STAR-Diff: Surgical Trocar-Adaptive, RCM-aware Diffusion Policy

Anonymous Authors

***Abstract*— Automating laparoscopic surgery holds significant promise for improving surgical outcomes. While some recent works have addressed varying trocar positions, they rely on wrist-mounted cameras to implicitly infer trocar configurations and use relative pose actions without explicitly constraining the action space to the RCM manifold—potentially causing issues when low-level controllers must project infeasible commands onto valid motions. Moreover, most prior works depend on proprietary systems such as the da Vinci Surgical System with articulated EndoWrist mechanisms, which alleviate RCM constraint challenges. In contrast, we present STAR-Diff (Surgical Trocar-Adaptive, RCM-aware Diffusion Policy), a novel imitation learning framework designed for standard 6-DoF manipulators without specialized surgical wrists, using only a fixed external camera. Our key contributions are: (1) RCM-aware action space design—we propose four parameterizations of the 4-dimensional RCM constraint manifold, categorized as fully-relative or mixed (relative-absolute), each guaranteeing constraint satisfaction by construction; and (2) trocar-centric observation space design—we introduce trocar-centric spatial encoding that augments RGB images with dense coordinate maps expressed in the trocar frame. For mixed parameterizations containing trocar-invariant components, we propose a hierarchical architecture where a diffusion model predicts the trocar-invariant component and a deterministic network completes the trocar-dependent component. STAR-Diff democratizes surgical automation beyond proprietary platforms, enabling generalization to unseen trocar configurations while guaranteeing zero RCM constraint violations by construction.**

I. Introduction

Minimally invasive surgery (MIS), particularly laparoscopic procedures, has revolutionized surgical practice by reducing patient trauma, hospital stays, and recovery times [**?**]. However, the inherent challenges of operating through small incisions—limited field of view, reduced haptic feedback, and constrained instrument motion—place significant cognitive and physical demands on surgeons. Robot-assisted surgical systems, exemplified by the da Vinci platform [**?**], have emerged as transformative tools that enhance precision and ergonomics. Yet, the potential for autonomous execution of surgical subtasks remains largely unrealized in clinical settings.

Recent advances in robot learning, particularly imitation learning and diffusion-based policies [**?**], [**?**], have demonstrated remarkable success in contact-rich manipulation tasks. Several works have applied these techniques to surgical automation [**?**], [**?**], achieving impressive results on benchmark tasks such as peg transfer, tissue manipulation, and suturing. However, critical limitations in existing approaches fundamentally constrain their practical applicability.

A. Limitations of Existing Approaches

(1) Implicit RCM Handling without Constraint Guarantees. While some recent works have addressed varying trocar positions [**?**], [**?**], [**?**], they do so without explicitly modeling the RCM constraint manifold. These approaches typically:

- Rely on **wrist-mounted cameras** to implicitly infer trocar configurations from visual observations, rather than explicitly conditioning on trocar frame information.
- Use **relative pose actions** with respect to the current end-effector pose, which does not inherently respect RCM constraints.
- Do not constrain the action space to the **RCM-feasible manifold**, meaning the policy can output actions that violate the RCM constraint.

This design leads to a critical issue: when the low-level controller receives an infeasible action command that would violate the RCM constraint, it must project or clamp the command onto a feasible motion. This projection can introduce **tracking errors**, cause **discontinuous motions**, and potentially lead to **task failures**—especially in precision-critical surgical manipulation.

(2) Reliance on Specialized Hardware. The majority of surgical automation research uses the **da Vinci Research Kit (dVRK)** [**?**], which features 7+ DoF instruments with articulated "wrist" joints (EndoWrist). This redundancy significantly relaxes the RCM constraint's impact on end-effector reachability—the wrist can compensate for approach angles that would be impossible with a rigid instrument. Consequently, prior works have not needed to carefully design RCM-aware action spaces for standard manipulators.

(3) Trocar Variation in Clinical Practice. Even for approaches that handle varying trocars implicitly, clinical practice demands robust adaptation to diverse trocar configurations based on:

- **Patient anatomy**: Body habitus, organ size, and pathology location necessitate patient-specific port configurations [**?**].
- **Procedure type**: Different procedures require distinct trocar arrangements.
- **Surgeon preference**: Individual surgeons develop personalized port placement preferences [**?**].

B. Our Setting: Democratizing Surgical Automation

In this work, we target a challenging and practical setting:

(1) Standard 6-DoF Manipulator without EndoWrist. We assume a general-purpose industrial manipulator (e.g., Fairino FR5) equipped with a rigid laparoscopic instrument.

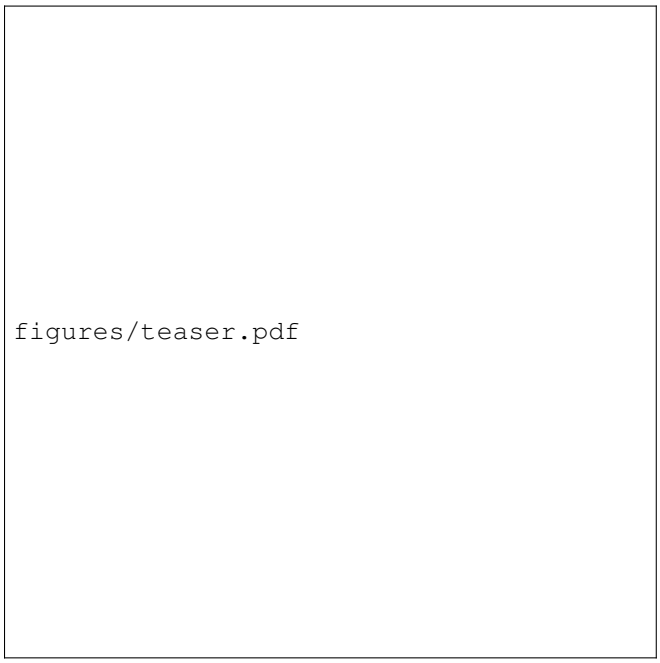

Fig. 1: **Problem Overview.** (a) Existing approaches either assume fixed trocars or handle varying trocars implicitly via wrist cameras without RCM constraint guarantees. (b) STAR-Diff explicitly parameterizes the RCM constraint manifold as the action space, guaranteeing feasibility by construction, and uses trocar-centric observation encoding with a fixed external camera.

Without additional wrist DoF, the RCM constraint directly couples tip position, approach angle, and axial rotation through the trocar point. This makes **explicit RCM-aware action space design** essential.

(2) Fixed External Camera without Wrist Sensing. We assume a single monocular camera mounted externally, observing the surgical workspace from a fixed viewpoint. Unlike wrist-mounted cameras that provide implicit trocar-frame observations, this configuration requires explicit handling of the observation-action frame mismatch.

We choose this setting to **democratize surgical automation research**: enabling experimentation with commodity hardware rather than requiring proprietary platforms or specialized sensing.

C. Our Approach: STAR-Diff

STAR-Diff addresses the above limitations through:

(1) RCM-aware Action Space Design. We directly parameterize the action space using the 4-dimensional RCM constraint manifold $\mathcal{M}_{\text{RCM}}$. We identify four candidate parameterizations:

- **Fully-relative**: Both components expressed in the trocar frame.
- **Mixed (relative-absolute)**: One trocar-invariant absolute component and one trocar-dependent relative component.

All parameterizations guarantee RCM compliance by construction—the low-level controller never receives infeasible commands.

(2) Trocar-Centric Observation Space Design. We introduce **trocar-centric spatial encoding** that augments RGB observations with per-pixel 3D coordinates expressed in the trocar frame, explicitly bridging the observation-action frame gap without relying on wrist cameras.

(3) Architecture Design. For fully-relative parameterizations, we use a single diffusion model. For mixed parameterizations, we propose a **hierarchical architecture** where a diffusion model predicts the trocar-invariant component and a deterministic MLP completes the trocar-dependent component.

We summarize our contributions:

1) We formalize **trocar-adaptive surgical manipulation** for standard 6-DoF manipulators with fixed external cameras.
2) We propose four **RCM-aware action space parameterizations** guaranteeing constraint satisfaction by construction.
3) We introduce **trocar-centric spatial encoding** for observation space design.
4) We present **single and hierarchical architectures** tailored to each action space category.

II. Related Work

A. Learning-based Surgical Automation

Machine learning applications to surgical robotics span perception, planning, and control. Early work focused on learning from demonstrations for surgical subtasks [**?**], [**?**]. Recent approaches leverage deep learning for end-to-end visuomotor policies. SurRoL [**?**] provides a simulation platform for surgical RL, while ORBIT-Surgical [**?**] extends this to GPU-accelerated simulation.

Imitation learning has emerged as the dominant paradigm. Diffusion Policy [**?**] advanced the state-of-the-art by capturing multimodal action distributions through denoising diffusion models.

Handling Trocar Variation. Recent works on generative policies for surgery—SRT [**?**], SRT-H [**?**], and SutureBot [**?**]—have addressed varying trocar configurations. However, these approaches:

- Use **wrist-mounted cameras** to implicitly infer trocar frame information from visual observations, rather than explicitly conditioning on trocar parameters.
- Employ **relative pose actions** without constraining outputs to the RCM-feasible manifold.
- Rely on low-level controllers to **project infeasible commands** onto valid motions, which can introduce tracking errors and discontinuities.

In contrast, **STAR-Diff explicitly parameterizes the RCM constraint manifold**, guaranteeing that every policy output is feasible by construction. We also demonstrate that explicit trocar conditioning with a fixed external camera can achieve robust generalization without wrist-mounted sensing.

B. Remote Center of Motion Constraints

The RCM constraint ensures instrument motion does not traumatize tissue at the insertion point. **Hardware-based** solutions use specialized kinematic designs [**?**], [**?**]. **Software-based** approaches employ constrained optimization [**?**] or control barrier functions [**?**]—but these add computational overhead and may introduce tracking errors when projecting infeasible commands. **Learning-based** methods incorporate RCM as penalty terms [**?**], but penalties do not guarantee constraint satisfaction.

Our approach differs fundamentally: by parameterizing the RCM manifold directly as the action space, **constraint satisfaction is guaranteed by construction**, eliminating the need for runtime projection.

C. Spatial Encoding for Visuomotor Learning

Bridging observation and action coordinate frames is a key challenge. Point cloud observations [**?**] provide 3D information. Neural descriptor fields [**?**] learn SE(3)-equivariant representations. Coordinate convolutions [**?**] append pixel coordinates as input channels. **We construct dense trocar-frame coordinate maps that embed geometric priors about the RCM constraint.**

D. Diffusion Models for Robot Control

Diffusion models [**?**] have entered robotics through Diffusion Policy [**?**]. Extensions include 3D Diffusion Policy [**?**] and Consistency Policy [**?**]. Our work adapts diffusion to surgical constraints through RCM-aware action spaces and trocar-centric observations.

III. Problem Formulation

A. Coordinate Frames

We establish the following coordinate frames (Fig. **??**):

- $\{W\}$: World frame, fixed.
- $\{C\}$: Camera frame, fixed relative to $\{W\}$ with known transform ${}^{W}T_{C} \in \mathrm{SE}(3)$.
- $\{T\}$: Trocar frame, centered at the trocar point $\mathbf{p}_T \in \mathbb{R}^3$. The transform ${}^{W}T_{T}(\lambda_{\text{trocar}}) \in \mathrm{SE}(3)$ depends on trocar parameters.
- $\{E\}$: End-effector frame, at the jaw center of the laparoscopic grasper.

B. Trocar Parameterization

We parameterize trocar position using spherical coordinates centered at a reference point $\mathbf{o}_{\text{ref}}$:

$$\lambda_{\text{trocar}} = (r, \theta, \phi) \in \mathbb{R}^{+} \times [0, \pi] \times [0, 2\pi), \tag{1}$$

where r is radial distance, θ is polar angle, and ϕ is azimuthal angle. The trocar position in world coordinates is:

$$\mathbf{p}_T(\lambda_{\text{trocar}}) = \mathbf{o}_{\text{ref}} + r \begin{bmatrix} \sin\theta\cos\phi \\ \sin\theta\sin\phi \\ \cos\theta \end{bmatrix}. \tag{2}$$

C. The RCM Constraint Manifold

[Remote Center of Motion] A manipulator satisfies the RCM constraint at point $\mathbf{p}_T$ if the instrument's longitudinal axis always passes through $\mathbf{p}_T$:

$$\mathbf{p}_E = \mathbf{p}_T + d \cdot R_E \mathbf{e}_z, \quad d \in \mathbb{R}, \tag{3}$$

where $R_E \in \mathrm{SO}(3)$ and $\mathbf{p}_E \in \mathbb{R}^3$ are end-effector orientation and position, $\mathbf{e}_z = [0, 0, 1]^{\top}$ is the instrument's longitudinal axis, and d is insertion depth.

[RCM Manifold Structure] The set of poses satisfying the RCM constraint forms a 4-dimensional submanifold $\mathcal{M}_{\text{RCM}}(\lambda_{\text{trocar}}) \subset \mathrm{SE}(3)$, diffeomorphic to $S^2 \times S^1 \times \mathbb{R}$, corresponding to instrument direction (2 DoF), axial rotation (1 DoF), and insertion depth (1 DoF).

D. Problem Statement

Given:

- A dataset $\mathcal{D} = \{(\tau_i, \lambda_i)\}_{i=1}^{N}$ of expert demonstrations under various trocar configurations.
- A distribution $p(\lambda_{\text{trocar}})$ over trocar configurations.

Find: A policy $\pi(a \mid o, \lambda_{\text{trocar}})$ that:

1) Maximizes task success across $p(\lambda_{\text{trocar}})$, including unseen configurations.
2) Guarantees RCM constraint satisfaction by construction.
3) Enables efficient real-time inference.

IV. Methodology

We present STAR-Diff, a diffusion-based framework for trocar-adaptive surgical manipulation. Our methodology centers on two design pillars: **RCM-aware action space** and **trocar-centric observation space**, with architecture choices tailored to each action space category.

A. Overview

STAR-Diff consists of three components (Fig. **??**):

1) **RCM-aware Action Space**: Four parameterizations of the RCM constraint manifold, categorized as fully-relative or mixed.
2) **Trocar-Centric Observation Space**: RGB images augmented with dense trocar-frame coordinate maps.
3) **Policy Architecture**: Single model for fully-relative; hierarchical model for mixed action spaces.

B. RCM-aware Action Space Design

We directly parameterize the action space using the RCM constraint manifold. We identify four candidate parameterizations:

Fully-Relative:	
$x_{\text{RCM}}^{(1)} = (d_{\text{rel}}, R_{\text{rel}})$	[Depth, Rotation]
$x_{\text{RCM}}^{(2)} = (\gamma_{\text{rel}}, \mathbf{p}_{\text{rel}})$	[Roll, Position]
Mixed (Relative, Absolute):	
$x_{\text{RCM}}^{(3)} = (d_{\text{rel}}, R_{\text{abs}})$	[Depth, Rotation]
$x_{\text{RCM}}^{(4)} = (\gamma_{\text{rel}}, \mathbf{p}_{\text{abs}})$	[Roll, Position]

(4)

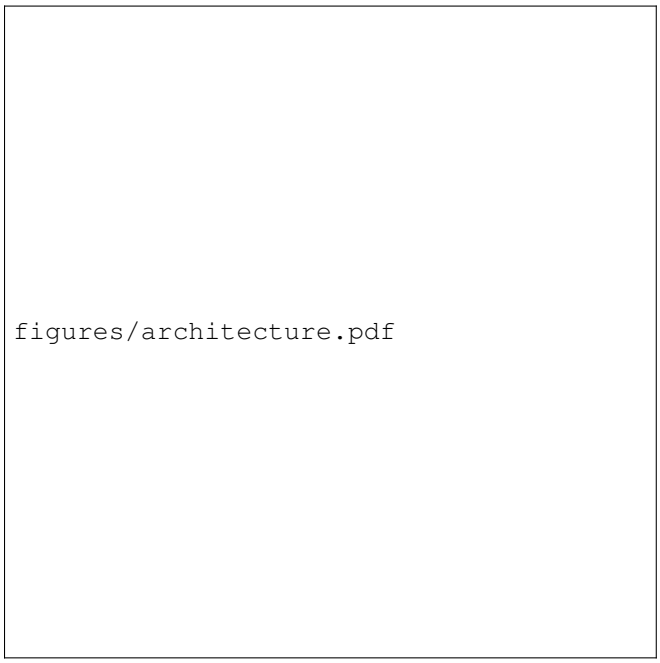

Fig. 2: **STAR-Diff Architecture.** (Left) Single architecture for fully-relative action spaces. (Right) Hierarchical architecture for mixed action spaces.

TABLE I: Properties of RCM Action Space Parameterizations

	$x^{(1)}_{\text{RCM}}$	$x^{(2)}_{\text{RCM}}$	$x^{(3)}_{\text{RCM}}$	$x^{(4)}_{\text{RCM}}$
Category	Fully-Relative		Mixed	
Trocar-invariant			R_{abs}	$\mathbf{p}_{\text{abs}}$
Config. consistency	–	–	Poor	Good
Architecture	Single	Single	Hier.	Hier.

"Relative" quantities are in the trocar frame $\{T\}$; "absolute" quantities are in the world frame $\{W\}$.

1) Fully-Relative Parameterizations: $x^{(1)}_{\mathbf{RCM}} = (d_{\mathbf{rel}}, R_{\mathbf{rel}})$: Insertion depth and rotation in the trocar frame.

$x^{(2)}_{\mathbf{RCM}} = (\gamma_{\mathbf{rel}}, \mathbf{p}_{\mathbf{rel}})$: Axial roll and position in the trocar frame.

Both components are trocar-dependent: the same task state corresponds to different action values for different trocar placements.

2) Mixed Parameterizations: $x^{(3)}_{\mathbf{RCM}} = (d_{\mathbf{rel}}, R_{\mathbf{abs}})$: Relative depth with absolute rotation.

$x^{(4)}_{\mathbf{RCM}} = (\gamma_{\mathbf{rel}}, \mathbf{p}_{\mathbf{abs}})$: Relative roll with absolute position.

Mixed parameterizations contain one **trocar-invariant** component (absolute) and one **trocar-dependent** component (relative).

3) Trocar-Invariance Analysis: [Trocar-Invariance of Absolute Position] For visual servoing tasks where the target is defined by visual features in a fixed camera frame, the desired absolute position $\mathbf{p}^*_{\text{abs}}$ is independent of trocar parameters λ_{trocar}.

$x^{(3)}_{\text{RCM}}$ exhibits **configuration inconsistency**: the same absolute rotation corresponds to different physical configurations depending on λ_{trocar}. $x^{(4)}_{\text{RCM}}$ exhibits **configuration consistency**: the same $\mathbf{p}_{\text{abs}}$ corresponds to similar tip positions regardless of trocar placement.

4) RCM Constraint Guarantee: [RCM Compliance by Construction] Any action produced using $x^{(1)}_{\text{RCM}}$–$x^{(4)}_{\text{RCM}}$, when converted to an end-effector pose via the transformations in Appendix **??**, satisfies the RCM constraint (**??**).

This guarantee is *unconditional*: regardless of neural network outputs, the resulting pose always satisfies RCM. No runtime projection is needed.

C. Trocar-Centric Observation Space Design

We address the observation-action frame mismatch through trocar-centric spatial encoding.

1) Trocar Parameter Conditioning: The trocar configuration $\lambda_{\text{trocar}} = (r, \theta, \phi)$ is explicitly provided:

$$\mathbf{e}_\lambda = \text{MLP}_{\text{embed}}(\lambda_{\text{trocar}}) \in \mathbb{R}^{d_\lambda}. \tag{5}$$

2) Dense Trocar-Frame Coordinate Maps: We augment RGB observations with per-pixel 3D coordinates in the trocar frame.

Step 1: 2D-to-3D Lifting. Project each pixel to camera-frame coordinates:

$$\mathbf{P}_{\text{cam}}(u, v) = Z(u, v) \cdot K^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}. \tag{6}$$

Step 2: Trocar-Centric Transformation.

$$\mathbf{P}_{\text{tro}}(u, v) = R_{T \leftarrow C} \cdot \mathbf{P}_{\text{cam}}(u, v) + t_{T \leftarrow C}. \tag{7}$$

Step 3: Augmented Observation.

$$\mathbf{O}_{\text{aug}} = \text{Concat}(\mathbf{I}_{\text{RGB}}, \mathbf{P}^{\text{norm}}_{\text{tro}}) \in \mathbb{R}^{H \times W \times 6}. \tag{8}$$

D. Policy Architecture

1) Single Architecture for Fully-Relative Action Spaces: For $x^{(1)}_{\text{RCM}}$ and $x^{(2)}_{\text{RCM}}$, a single diffusion model predicts the complete action:

$$\pi_{\text{single}} : (\mathbf{O}_{\text{aug}}, \mathbf{x}_{\text{history}}, \lambda_{\text{trocar}}) \mapsto x^{(k)}_{\text{RCM}}, \quad k \in \{1, 2\} \tag{9}$$

2) Hierarchical Architecture for Mixed Action Spaces: For $x^{(3)}_{\text{RCM}}$ and $x^{(4)}_{\text{RCM}}$, we use a two-stage hierarchical architecture:

Stage 1: Absolute Component (Diffusion Model)

$$\pi_{\text{abs}} : (\mathbf{O}_{\text{aug}}, \mathbf{x}^{\text{history}}_{\text{abs}}) \mapsto \mathbf{x}^{t+1:t+H}_{\text{abs}} \tag{10}$$

Trocar parameters are **not conditioned** since the absolute component is trocar-invariant.

Stage 2: Relative Component (Deterministic MLP)

$$f_{\text{rel}} : (\mathbf{O}_{\text{aug}}, \mathbf{x}^{t+1:t+H}_{\text{abs}}, \mathbf{x}^{\text{history}}_{\text{rel}}, \lambda_{\text{trocar}}) \mapsto \mathbf{x}^{t+1:t+H}_{\text{rel}} \tag{11}$$

E. Diffusion Policy Details

We adopt DDPM [**?**]. The noise prediction network ϵ_θ is trained with:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{a}_0, k, \boldsymbol{\epsilon}} \left[\|\boldsymbol{\epsilon} - \epsilon_\theta(\mathbf{a}_k, \mathbf{c}, k)\|^2 \right]. \tag{12}$$

For hierarchical architectures, the MLP is trained with:

$$\mathcal{L}_{\text{rel}} = \mathbb{E} \left[\|\mathbf{x}^{t+1:t+H}_{\text{rel}} - f_{\text{rel}}(\cdot)\|^2 \right]. \tag{13}$$

Total loss: $\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda \mathcal{L}_{\text{rel}}$ with $\lambda = 0.1$.

TABLE II: Compared Methods

Group	Method	Action Space	Obs.	Arch.
Baselines	DP-SE3	$\mathrm{SE}(3)$	RGB	Single
	DP-SE3-Proj	$\mathrm{SE}(3)$ + Proj	RGB	Single
Fully-Rel.	STAR-1	$x_{\text{RCM}}^{(1)}$	RGB+TC	Single
	STAR-2	$x_{\text{RCM}}^{(2)}$	RGB+TC	Single
Mixed	STAR-3-S	$x_{\text{RCM}}^{(3)}$	RGB+TC	Single
	STAR-3-H	$x_{\text{RCM}}^{(3)}$	RGB+TC	Hier.
	STAR-4-S	$x_{\text{RCM}}^{(4)}$	RGB+TC	Single
	STAR-4-H	$x_{\text{RCM}}^{(4)}$	RGB+TC	Hier.

TC: Trocar-centric coordinates. S: Single. H: Hierarchical.

F. Action Execution via Low-Level Control

1) Action-to-Pose Transformation: All four parameterizations are converted to $\mathrm{SE}(3)$ poses via $\mathcal{T} : x_{\text{RCM}}^{(k)} \times \lambda_{\text{trocar}} \to \mathrm{SE}(3)$ (Appendix **??**).

2) Low-Level Control: Given desired pose $(R_E^*, \mathbf{p}_E^*)$:

1) **Inverse Kinematics**: $\mathbf{q}^* = \mathrm{IK}(R_E^*, \mathbf{p}_E^*; \mathbf{q}_{\text{current}})$
2) **Joint Space Controller**: $\boldsymbol{\tau} = K_p(\mathbf{q}^* - \mathbf{q}) + K_d(\dot{\mathbf{q}}^* - \dot{\mathbf{q}})$

Since all policy outputs lie on the RCM manifold by construction, no projection or clamping is required.

V. Experiments

We design experiments to evaluate:

- **RQ1**: Generalization to unseen trocar positions
- **RQ2**: Comparison of RCM action space parameterizations
- **RQ3**: Effect of trocar-centric spatial encoding
- **RQ4**: Single vs. hierarchical architecture for mixed action spaces

A. Experimental Setup

1) Task: We evaluate on **Peg Transfer** from the FLS benchmark [**?**], implemented in SurRoL [**?**] simulation with a 6-DoF Fairino FR5 and laparoscopic grasper, observed by a fixed RGB-D camera.

2) Trocar Configuration: 12 trocar configurations total:

- **Training**: 8 configurations (67%)
- **Test (Seen)**: Same 8 as training
- **Test (Unseen)**: 4 held-out configurations (33%)

3) Metrics: **Task Success Rate (SR)**: Percentage of successful transfers within 60 seconds (30 trials per configuration).

Generalization Gap: $\Delta_{\text{gen}} = \text{SR}_{\text{seen}} - \text{SR}_{\text{unseen}}$.

B. Methods

Baselines: DP-SE3 (Diffusion Policy in $\mathrm{SE}(3)$) and DP-SE3-Proj (with post-hoc RCM projection).

Fully-Relative: STAR-1 ($x_{\text{RCM}}^{(1)}$) and STAR-2 ($x_{\text{RCM}}^{(2)}$) with single architecture.

Mixed: STAR-3-S/H ($x_{\text{RCM}}^{(3)}$) and STAR-4-S/H ($x_{\text{RCM}}^{(4)}$) with single and hierarchical architectures.

TABLE III: Task Success Rate (%) on Peg Transfer

Group	Method	Seen↑	Unseen↑	Δ_{gen}↓
Baselines	DP-SE3	[XX.X]	[XX.X]	[XX.X]
	DP-SE3-Proj	[XX.X]	[XX.X]	[XX.X]
Fully-Rel.	STAR-1	[XX.X]	[XX.X]	[XX.X]
	STAR-2	[XX.X]	[XX.X]	[XX.X]
Mixed	STAR-3-S	[XX.X]	[XX.X]	[XX.X]
	STAR-3-H	[XX.X]	[XX.X]	[XX.X]
	STAR-4-S	[XX.X]	[XX.X]	[XX.X]
	STAR-4-H	[XX.X]	[XX.X]	[XX.X]

TABLE IV: Effect of Trocar-Centric Encoding (using STAR-4-H)

Observation	Seen	Unseen	Δ_{gen}
RGB only	[XX.X]	[XX.X]	[XX.X]
RGB + TC	[XX.X]	[XX.X]	[XX.X]

C. Main Results

D. Ablation: Observation Space

E. Ablation: Architecture (Single vs. Hierarchical)

Comparison of STAR-3-S vs. STAR-3-H and STAR-4-S vs. STAR-4-H in Table **??** evaluates the benefit of hierarchical decomposition for mixed action spaces.

VI. Discussion and Conclusion

A. Limitations

(1) Depth Requirement: Trocar-centric encoding requires depth; extending to monocular depth estimation would broaden applicability.

(2) 6-DoF Constraint: Extension to redundant manipulators (7+ DoF) is future work.

(3) Fixed Camera: Incorporating dynamic endoscope poses would enhance clinical applicability.

(4) Simulation-Only: Hardware validation is needed.

B. Conclusion

We presented STAR-Diff for trocar-adaptive surgical manipulation with the following contributions:

1) **RCM-aware Action Space**: Four parameterizations guaranteeing constraint satisfaction by construction—eliminating the need for runtime projection that plagues existing approaches.
2) **Trocar-Centric Observation Space**: Dense coordinate encoding bridging observation-action frame gaps without wrist-mounted cameras.
3) **Tailored Architectures**: Single model for fully-relative; hierarchical model for mixed action spaces.

Appendix

A. Preliminaries

Trocar position: $\mathbf{p}_T = \mathbf{o}_{\text{ref}} + r[\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta]^\top$

RotationAlign$(\mathbf{a}, \mathbf{b})$ computes rotation mapping $\mathbf{a}$ to $\mathbf{b}$ via Rodrigues' formula.

B. $x_{RCM}^{(1)} = (d_{rel}, R_{rel})$

$$R_E = {}^{W}R_T(\lambda_{\text{trocar}}) \cdot R_{\text{rel}} \tag{14}$$
$$\mathbf{p}_E = \mathbf{p}_T + d_{\text{rel}} \cdot R_E \mathbf{e}_z \tag{15}$$

C. $x_{RCM}^{(2)} = (\gamma_{rel}, \mathbf{p}_{rel})$

$$\mathbf{p}_E = {}^{W}R_T \cdot \mathbf{p}_{\text{rel}} + \mathbf{p}_T \tag{16}$$
$$\mathbf{v} = (\mathbf{p}_E - \mathbf{p}_T)/\|\mathbf{p}_E - \mathbf{p}_T\| \tag{17}$$
$$R_E = \text{RotationAlign}(\mathbf{e}_z, \mathbf{v}) \cdot R_z(\gamma_{\text{rel}}) \tag{18}$$

D. $x_{RCM}^{(3)} = (d_{rel}, R_{abs})$

$$R_E = R_{\text{abs}} \tag{19}$$
$$\mathbf{p}_E = \mathbf{p}_T + d_{\text{rel}} \cdot R_E \mathbf{e}_z \tag{20}$$

E. $x_{RCM}^{(4)} = (\gamma_{rel}, \mathbf{p}_{abs})$

$$\mathbf{p}_E = \mathbf{p}_{\text{abs}} \tag{21}$$
$$\mathbf{v} = (\mathbf{p}_{\text{abs}} - \mathbf{p}_T)/\|\mathbf{p}_{\text{abs}} - \mathbf{p}_T\| \tag{22}$$
$$R_E = \text{RotationAlign}(\mathbf{e}_z, \mathbf{v}) \cdot R_z(\gamma_{\text{rel}}) \tag{23}$$

Vision Encoder: ResNet-18 (6-channel) with spatial softmax, output $\in \mathbb{R}^{512}$.

Trocar Encoder: 2-layer MLP, output $\in \mathbb{R}^{64}$.

U-Net: 1D with 4 blocks, 256 base channels, FiLM conditioning.

Relative MLP: 3 residual blocks (256 dim, LayerNorm, GELU).

Data: 30 demos per trocar config at 30 Hz. Trocar range: $r \in [0, 5]$ cm, $\theta \in [60, 120]$, $\phi \in [0, 360]$.

Diffusion: $K = 100$ train / 10 DDIM inference, cosine schedule, $T_o = 2$, $H = 16$.

Optimization: AdamW, lr $= 10^{-4}$, batch 256, 100K iterations.

REFERENCES

[1] M. J. Mack, "Minimally invasive and robotic surgery," *JAMA*, vol. 285, no. 5, pp. 568–572, 2001.

[2] Intuitive Surgical, "da Vinci Surgical System," 2023.

[3] C. Chi *et al.*, "Diffusion policy: Visuomotor policy learning via action diffusion," in *Proc. RSS*, 2023.

[4] T. Z. Zhao *et al.*, "Learning fine-grained bimanual manipulation with low-cost hardware," in *Proc. RSS*, 2023.

[5] S. Scheikl *et al.*, "Movement primitive diffusion for dexterous surgical manipulation," in *Proc. ICRA*, 2024.

[6] J. Xu *et al.*, "SurRoL: An open-source reinforcement learning centered and dVRK compatible platform for surgical robot learning," in *Proc. IROS*, 2021.

[7] W. Reynolds, "The first laparoscopic cholecystectomy," *JSLS*, vol. 5, no. 1, pp. 89–94, 2001.

[8] D. Murphy *et al.*, "Surgeon preferences and peritoneal access," *Surg. Endosc.*, vol. 15, pp. 1236–1240, 2001.

[9] P. Kazanzides *et al.*, "An open-source research kit for the da Vinci surgical system," in *Proc. ICRA*, 2014.

[10] J. Kim *et al.*, "Surgical Robot Transformer (SRT): Imitation learning for surgical tasks," in *Proc. ICRA*, 2024.

[11] J. Kim *et al.*, "SRT-H: A hierarchical framework for autonomous surgery via language conditioned imitation learning," *arXiv preprint*, 2024.

[12] A. Chiu *et al.*, "SutureBot: A precision framework and benchmark for autonomous end-to-end suturing," *arXiv preprint*, 2024.

[13] R. H. Taylor *et al.*, "A steady-hand robotic system for microsurgical augmentation," *Int. J. Robot. Res.*, vol. 18, no. 12, pp. 1201–1210, 1999.

[14] J. Sandoval *et al.*, "Collaborative framework for robot-assisted minimally invasive surgery," in *Proc. ICRA*, 2017.

[15] J. Lu *et al.*, "Super: A surgical perception framework for endoscopic tissue manipulation," *IEEE RA-L*, vol. 6, no. 2, pp. 3977–3984, 2021.

[16] Y. Yu *et al.*, "ORBIT-Surgical: An open-simulation framework for learning surgical augmented dexterity," in *Proc. ICRA*, 2024.

[17] J. van den Berg *et al.*, "Superhuman performance of surgical tasks by robots using iterative learning," in *Proc. ICRA*, 2010.

[18] J. Schulman *et al.*, "A case study of trajectory transfer through non-rigid registration for a simplified suturing scenario," in *Proc. IROS*, 2013.

[19] G. S. Guthart and J. K. Salisbury, "The Intuitive telesurgery system: Overview and application," in *Proc. ICRA*, 2000.

[20] M. Selvaggio *et al.*, "Safe and efficient autonomous navigation for teleoperation in surgery," *IEEE RA-L*, vol. 3, no. 4, pp. 3373–3380, 2018.

[21] Y. Ze *et al.*, "3D diffusion policy," in *Proc. RSS*, 2024.

[22] A. Simeonov *et al.*, "Neural descriptor fields: SE(3)-equivariant object representations for manipulation," in *Proc. ICRA*, 2022.

[23] R. Liu *et al.*, "An intriguing failing of convolutional neural networks and the coordconv solution," in *Proc. NeurIPS*, 2018.

[24] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. NeurIPS*, 2020.

[25] V. Prasad *et al.*, "Consistency policy: Accelerated visuomotor policies via consistency distillation," in *Proc. RSS*, 2024.

[26] J. H. Peters *et al.*, "Development and validation of a comprehensive program of education and assessment of the basic fundamentals of laparoscopic surgery," *Surgery*, vol. 135, no. 1, pp. 21–27, 2004.