ROBOTIC OPTICAL COHERENCE TOMOGRAPHY-GUIDED GEOMETRIC MODELING OF SOFT TISSUE: 3D REPRESENTATIONS FOR DATA-DRIVEN CLOSED-LOOP SHAPE CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

In medical robotics, data-driven soft-tissue shape control requires deforming tissue toward desired morphologies through sequential physical interactions. This challenge is particularly important in microscale manipulation, where localized deformation is difficult to observe, represent, and control, while physics-based models can be computationally expensive and difficult to adapt to realistic environments. Optical coherence tomography (OCT) provides high-resolution volumetric perception of tissue deformation, but target-conditioned control requires effective representations of both observed and desired tissue states. In this work, we introduce an OCT-guided data-driven framework for tissue shape control, learned from user-guided demonstrations and executed with closed-loop OCT feedback. We study full-volume, tissue-masked sparse-volume, and voxelized point-cloud representations, with or without residual conditioning, using deterministic (Explicit) and generative diffusion (Diffusion) policies. Across 12 configurations and 576 phantom rollouts on 3-DoF depth and 6-DoF directional tasks, OCT-derived 3D representations support both Explicit and Diffusion policies, with performance depending on representation and residual conditioning. Residual conditioning reduces directional Shape RMSE from 0.748 to 0.555 mm for full-volume Diffusion, while voxelized point-cloud Explicit with residuals achieves 0.534 mm for depth and 0.443 mm for directional control, with comparable accuracy at intermediate unseen targets and consistent performance across deployments. We further evaluate the framework in 60 *ex vivo* porcine rollouts, extending to biological tissue with final geometric errors comparable to human demonstrations. These results establish volumetric OCT as current and goal representations for learned closed-loop tissue shape control across deterministic and generative action models.

1 INTRODUCTION

In medical robotics, microscale tissue manipulation requires precise control of deformable tissue through physical interaction, where tissue configuration can directly affect task execution (Thach et al., 2024; Wang et al., 2023). For example, robotic palpation relies on controlled tool-tissue interaction to induce and regulate local deformation toward a desired morphological state. The control objective is therefore not simply a tool pose, but a target tissue morphology achieved through sequential interactions and successive observation of tissue response. Goal conditioned deformable manipulation has demonstrated this principle using image contours and point clouds (Navarro-Alarcon & Liu, 2018; Thach et al., 2022; 2024). Extending this paradigm to volumetric medical imaging is particularly relevant to microscale manipulation, where 3D image volumes preserve spatial and structural information (Zhou et al., 2026) and allow desired tissue morphologies to be represented directly in the imaging space using both intensity based and geometric representations.

At the microscale, morphology-based control requires high-resolution volumetric observations that capture both surface geometry and depth-dependent tissue structure. Optical coherence tomography (OCT) provides such observations, resolving local surface morphology and subsurface structure within a 3D imaging volume (Huang et al., 1991; Ma et al., 2025). This volumetric perception

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

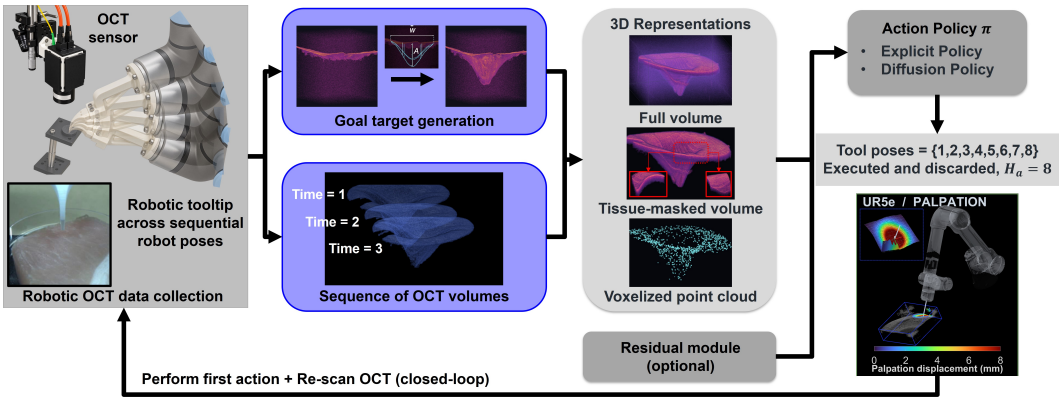


Figure 1: **Closed-loop OCT-guided tissue shape control.** The OCT captures volumetric tissue deformation induced by the robotic tooltip, while parameterized target generation constructs desired morphologies from predeformed OCT states. Current and target OCT volumes are encoded as full volumes, tissue-masked volumes, or voxelized point clouds, with optional current–target residual conditioning. These representations condition an Explicit or Diffusion policy to predict tooltip poses, with the first action executed before an OCT scan updates the state for closed-loop prediction.

has enabled robotic OCT systems for learned interaction, insertion control, and deformation-aware instrument navigation (Keller et al., 2020; Edwards et al., 2022; Arikan et al., 2025). Based on these capabilities, we develop ***OCT-guided soft-tissue shape control***, a learning-based framework (Fig. 1) that uses OCT feedback to guide tool motion toward a target tissue morphology represented in the same imaging space. This closed-loop formulation enables the robot to adapt its actions to observed deformation, providing a foundation for image-guided tissue shaping in microscale interventions.

Realizing closed-loop OCT-guided shape control requires jointly addressing target construction, OCT representation, and action prediction. Prior work has used geometric and learned goal shapes for deformable-object manipulation (Thach et al., 2022; 2026), while OCT-derived point clouds have been explored for registration and reconstruction (Liu et al., 2024; Fang et al., 2025). Building upon these directions, we introduce an OCT-guided robotic framework that constructs parameterized target morphologies from measured tissue states and learns to deform tissue toward these targets from user demonstrations. We investigate full OCT volumes, tissue-masked intensity volumes, and voxelized point clouds for representing current and target states, and compare deterministic Explicit and Diffusion action models (Chi et al., 2023; Ze et al., 2024), with and without representation-specific current–target residual conditioning. These policies predict incremental 6-DoF tooltip motions, while successive OCT scans provide closed-loop feedback toward the desired tissue morphology.

In this work, we evaluate 12 policy configurations on physical robotic shape control tasks spanning depth and directional deformations, including intermediate targets unseen during training, and further demonstrate the framework on *ex vivo* porcine tissue with varying surface geometries. Our results demonstrate the feasibility of using volumetric OCT observations for closed loop shape control with both Explicit and Diffusion policies, while voxelized point cloud representations provide accurate and repeatable control through compact geometric encoding. In addition, current to target residual conditioning further improves performance across most configurations and strengthens target dependent responses, particularly for full volume representations. Overall, this work establishes a systematic framework for studying how volumetric and geometric OCT representations support learned closed loop tissue shape control, connecting high resolution medical imaging with deterministic Explicit and generative Diffusion action policies for deformable tissue manipulation.

2 Related Work

OCT-guided robotic intervention: OCT has increasingly been integrated into robotic intervention through image-based visual servoing (Duflot et al., 2018), learning-based control from demonstrations (Keller et al., 2020), and data-driven modeling of tool–tissue interactions (Edwards et al., 2022; Ma et al., 2025). For example, in retinal interventions, OCT-guided systems combine instrument

localization, motion planning, and feedback control for autonomous navigation and targeted injection (Zhang et al., 2024; Bian et al., 2026; Liao et al., 2026), while deformation-aware approaches incorporate anatomical changes to adapt instrument positioning during tissue interaction (Arikan et al., 2025; Draelos et al., 2019). These methods primarily use OCT to localize anatomy, guide instruments, or adapt trajectories relative to anatomical targets. In contrast, we formulate the observed tissue morphology as the controlled state and a desired 3D tissue configuration as the goal, extending OCT-guided robotic intervention toward target-conditioned volumetric tissue shape control.

OCT representations for geometry and interaction: Learning from OCT has explored both volumetric appearance and explicit tissue geometry. Three-dimensional convolutional networks have been used to estimate interaction forces directly from OCT volumes (Gessert et al., 2018), while spatiotemporal models capture tissue deformation across successive scans (Gessert et al., 2020). Beyond volumetric intensity, geometric learning on segmented OCT point clouds has been compared with volumetric convolution for glaucoma diagnosis (Thiéry et al., 2023). Point-cloud representations also support anatomical reconstruction: C2P-Net performs non-rigid registration of anatomical templates to partial OCT point clouds (Liu et al., 2024), while point cloud-based volumetric montaging integrates robotic OCT acquisitions into an expanded field of view (Fang et al., 2025). Though these studies demonstrate the utility of different OCT representations for perception and reconstruction, we investigate their role in closed-loop action learning by comparing volumetric intensity, tissue-masked intensity, and point geometry for target-conditioned physical tissue deformation.

Goal conditioning and action prediction for deformable manipulation: Robotic shape control requires both specifying desired configurations and translating current–goal relationships into effective manipulation actions. Prior goal-conditioned shape servoing represents targets using image contours (Navarro-Alarcon & Liu, 2018) or partial-view point clouds (Thach et al., 2022). Beyond explicitly specified goals, DefGoalNet learns context-dependent target shapes from demonstrations (Thach et al., 2024), while DiffDef models multimodal goal distributions via diffusion and passes generated goals to DeformerNet for action prediction (Thach et al., 2026). Model-based approaches such as RoboCraft and DexDeform instead leverage learned dynamics or differentiable physics for deformable-object planning (Shi et al., 2022; Li et al., 2023). In this work, we construct volumetric OCT targets by applying parameterized deformations to measured tissue states, enabling direct control over local target depth, width, and direction while preserving specimen-specific OCT structure. These targets support both intensity-based and geometric representations, enabling systematic evaluation of how 3D goal representation affects closed-loop tissue shape control.

Given current and desired states, action prediction translates their relationship into goal-directed manipulation. Deterministic policies directly regress actions, whereas Diffusion Policy generates visuomotor action sequences through conditional denoising (Chi et al., 2023; Wolf et al., 2025). Reuss et al. incorporate explicit goals into score-based action generation (Reuss et al., 2023), while DP3 combines compact point-cloud representations with diffusion-based action prediction (Ze et al., 2024). Movement Primitive Diffusion further integrates diffusion-based imitation learning with structured motion generation for deformable-object manipulation (Scheikl et al., 2024). Goal conditioning can explicitly encode current–goal discrepancies, as in DeformerNet's use of feature differences (Thach et al., 2022). Motivated by these approaches, we compare Explicit and Diffusion policies with and without representation-specific current–target residual conditioning, examining how action prediction and goal conditioning interact with the information retained by different OCT representations.

3 OCT-Guided Tissue Shape Control Framework

We formulate soft-tissue shape control as a goal-conditioned robot learning problem with closed-loop volumetric OCT feedback. A policy learned from user-guided demonstrations predicts tooltip motions conditioned on current OCT observations and a target tissue morphology. The framework integrates parameterized target generation (Appendix A.2), 3D OCT representations, goal-conditioning schemes, and Explicit and Diffusion policies within a unified closed-loop formulation.

3.1 Problem formulation and closed-loop control

Let V_{pre} denote a predeformation OCT scan and $G = \mathcal{T}(V_{\mathrm{pre}}; \eta)$ as a target volume constructed with deformation parameters η. The target remains fixed within a rollout. At step t, the policy receives

the target, an observation history $V_{t-H_o+1:t}$, and the corresponding tool poses $\mathbf{p}_{t-H_o+1:t}$ to predict

$$\widehat{\mathbf{A}}_t = \pi_\theta(V_{t-H_o+1:t}, G, \mathbf{p}_{t-H_o+1:t}) \in \mathbb{R}^{H_a \times 6}. \tag{1}$$

Here H_o is the observation horizon and H_a is the action prediction horizon. Each row of $\widehat{\mathbf{A}}_t$ contains normalized translation and rotation increments of the tooltip pose (policy in Fig. 1). These intermediate poses are defined relative to the initial tooltip pose, while action increments are executed in the current tooltip frame. The predicted action sequence is clipped component-wise to $[-1, 1]$ in normalized action space. The first action is then denormalized to translational increments in mm for $(\Delta x, \Delta y, \Delta z)$ and rotational increments in degrees for $(\Delta\phi, \Delta\theta, \Delta\psi)$, and executed. A new OCT scan then updates the observation history for the next prediction. This receding-horizon procedure closes the loop between observed deformation and tooltip motion (Chi et al., 2023).

3.2 OCT REPRESENTATIONS

We represent current OCT observations and target G as full intensity volumes, tissue-masked sparse volumes, or voxelized point clouds. Each policy uses the same representation, spatial calibration, and preprocessing for current and target states. These representations retain different combinations of OCT appearance and tissue geometry, enabling us to study how representation affects target-conditioned policy learning.

Full Volume: This representation retains the dense OCT intensity volume as an $H \times W \times D$ voxel grid, preserving tissue structure and surrounding background signals. A 3D convolutional encoder processes the complete volume to learn spatial and intensity features from the OCT observation.

Tissue-masked Sparse Volume: This representation retains the tissue voxels as an $M \times 4$ sparse representation, where each voxel is described by its 3D coordinate and OCT intensity, (x, y, z, I). A sparse 3D convolutional encoder processes the occupied voxels and their local neighborhoods to capture tissue-specific intensity and structural information while excluding background signals.

Tissue-masked Voxelized Point Cloud: This representation retains the tissue voxel coordinates within the 3D OCT image space as an $M \times 3$ point set, (x, y, z), including the tissue interior while discarding OCT intensity information. A PointNet-style encoder (Qi et al., 2017) applies shared pointwise transformations followed by symmetric pooling to obtain a geometry-only embedding.

3.3 REPRESENTATION-SPECIFIC RESIDUAL CONDITIONING

Beyond the choice of OCT representation, we investigate how current and target tissue states are conditioned for action prediction. All policies receive features from both the current observation and target. Residual-conditioned variants additionally encode an explicit target–current discrepancy for each history observation, providing the policy with a direct representation of the remaining deformation. For volumetric representations, this residual is computed on the spatial voxel lattice, whereas for voxelized point clouds, it is computed in the learned feature space.

Full-volume Residual: For normalized current and target volumes $\widetilde{V}_t$ and $\widetilde{G}$, the residual is the voxelwise difference $R_t = \widetilde{G} - \widetilde{V}_t$, which is encoded as an additional volume.

Sparse-tissue Residual: Let M_t and M_g denote binary occupancy masks on the same 3D voxel grid, and let I_t and I_g denote intensities, set to zero outside their respective masks. On the coordinate union $\mathcal{U}_t = \{\mathbf{q} : M_t(\mathbf{q}) = 1 \text{ or } M_g(\mathbf{q}) = 1\}$, the residual has two channels,

$$\mathbf{r}_t(\mathbf{q}) = \begin{bmatrix} M_g(\mathbf{q})I_g(\mathbf{q}) - M_t(\mathbf{q})I_t(\mathbf{q}) \\ M_g(\mathbf{q}) - M_t(\mathbf{q}) \end{bmatrix}, \qquad \mathbf{q} \in \mathcal{U}_t. \tag{2}$$

The intensity channel captures local OCT appearance differences within the masked volumes, while the occupancy channel explicitly encodes geometric mismatch where tissue is present in only one state. Together, they distinguish changes in tissue appearance from changes in tissue occupancy, and a sparse encoder extracts features from this two-channel residual.

Voxelized Point Cloud Residual: For voxelized point clouds P_t and P_g, a shared encoder E produces pooled features $\mathbf{z}_t = E(P_t)$ and $\mathbf{z}_g = E(P_g)$. We define the residual as $\mathbf{r}_t = \mathbf{z}_g - \mathbf{z}_t$, providing target–current discrepancy conditioning in feature space without requiring pointwise correspondence (Thach et al., 2022).

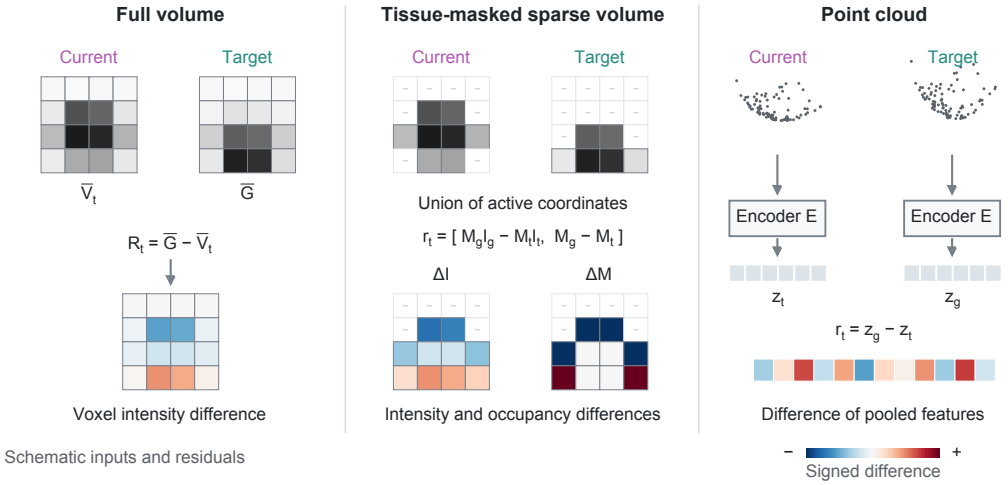


Figure 2: **Representation-specific residual conditioning.** Full volumes use voxelwise intensity differences, sparse tissue volumes use intensity and occupancy differences over occupied coordinates, and point clouds use pooled feature differences. Values are schematic.

3.4 Action sequence prediction

Observation and target features, relative tool poses, and optional residual features are combined into the conditioning input c_t. Conditioned on c_t, each policy predicts a sequence of H_a normalized 6-DoF tool-motion increments. We compare deterministic behavioral cloning (Explicit) with diffusion-based policies that model conditional distributions over action sequences.

Explicit Policy: We adopt a deterministic explicit policy (Chi et al., 2023) that maps the conditioning features c_t directly to an action sequence. A set of H_a learned action queries attends to the conditioning features through a multi-head cross-attention decoder. The decoder jointly predicts the full action sequence in a single forward pass, $\widehat{\mathbf{A}}_t = f_\theta^{\text{explicit}}(c_t)$.

Diffusion Policy: We train a conditional diffusion model (Chi et al., 2023) to recover expert actions from noisy sequences. During training, a normalized expert action sequence $\mathbf{A}_t$ is corrupted with Gaussian noise,

$$\mathbf{A}_t^{(k)} = \sqrt{\bar{\alpha}_k}\,\mathbf{A}_t + \sqrt{1-\bar{\alpha}_k}\,\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, I), \tag{3}$$

where k denotes the diffusion step and $\bar{\alpha}_k$ controls the mixture of the expert action sequence and Gaussian noise. Conditioned on the noisy action sequence, k, and c_t, a Transformer predicts the clean action sequence using self-attention over action tokens and cross-attention to the conditioning features. At inference, the policy starts from Gaussian noise and iteratively denoises the action sequence through successive clean-action predictions.

Training and model selection. Both policy types are trained to reconstruct normalized expert action sequences. Let $\mathcal{J}_t$ contain the available future-action positions in a demonstration sequence. The training loss is

$$\ell_t = \frac{1}{6|\mathcal{J}_t|} \sum_{j \in \mathcal{J}_t} \left\| \widehat{\mathbf{A}}_t[j] - \mathbf{A}_t[j] \right\|_2^2. \tag{4}$$

The loss averages squared prediction error over valid future steps and all six action components. For diffusion policies, it is additionally averaged over sampled diffusion steps and noise realizations. Unavailable future actions are zero-padded and do not directly contribute to the loss.

All configurations share the demonstration split and closed-loop execution protocol in Section 3.1. For each model, we select the checkpoint with the lowest validation action MSE in normalized space, with equal weighting across trajectories (implementation and training details in Appendix A).

4 REAL-ROBOT EVALUATION

We evaluated the proposed framework on physical tissue phantoms and *ex vivo* porcine tissue to determine whether learned policies (i) achieve desired tissue morphologies, (ii) adapt actions to different targets, and (iii) generalize to intermediate targets unseen during training.

4.1 EXPERIMENTAL SETUP

The platform combines a 6-DoF UR5e robot, a printed palpation tooltip, a 1064-nm swept-source OCT system operating at 400 kHz, and an RGB camera (Fig. 1). Experiments use a soft silicone 40A phantom with an internal empty layer, together with *ex vivo* porcine tissue. Two operators collected 40 depth and 51 directional phantom demonstrations and 30 demonstrations per porcine task, totaling 151 trajectories and 1,714 OCT observations. All configurations use an approximately 80/20 trajectory-level training/validation split (data acquisition and preprocessing details in Appendix A).

The phantom depth task evaluates 2–5 mm targets in 0.5-mm increments, with demonstrations at integer depths. The directional task evaluates $0°$–$180°$ in $22.5°$ increments, with demonstrations every $45°$, a depth of 3.5 mm, and tilt strength of 0.7. Both use a target width of 6 mm. Twelve policy configurations combine three OCT representations (full volume, tissue-masked sparse volume, and voxelized point cloud), with or without residual conditioning, and Explicit or Diffusion prediction. Each checkpoint is evaluated over three target sweeps, yielding 252 depth and 324 directional trials. Between sweeps, the phantom is repositioned and rescanned, with deployment seeds varied where applicable. Each rollout executes 12 actions with OCT feedback toward a fixed target generated from the initial scan. Repeated sweeps quantify deployment variability for each checkpoint.

4.2 EVALUATION METRICS

Shape accuracy: Let P_t and Q denote occupied voxel locations in the observed and target tissue masks, in calibrated millimeters. With $d(\mathbf{p}, Q) = \min_{\mathbf{q}\in Q} \|\mathbf{p} - \mathbf{q}\|_2$, symmetric Shape RMSE is

$$D_t = \sqrt{\frac{1}{2|P_t|}\sum_{\mathbf{p}\in P_t} d(\mathbf{p}, Q)^2 + \frac{1}{2|Q|}\sum_{\mathbf{q}\in Q} d(\mathbf{q}, P_t)^2}. \tag{5}$$

The observation-to-target component ($P_t \to Q$) measures how far observed tissue lies outside the target support, whereas the target-to-observation component ($Q \to P_t$) measures how far uncovered target regions lie from the observed tissue. Evaluation uses the native OCT grid without registration or policy-input resampling. We report final error D_{12}, reduction $\rho = 100(1 - D_{12}/D_0)$, and the number of rollouts satisfying $D_{12} < D_0$. Means and SDs summarize three target-averaged sweeps (HD95 and seen/unseen results in Appendix B.1).

Target-dependent action response and demonstration agreement: Distance correlation r_d measures association between target parameters and composed endpoint commands; response spread W measures their across-target variation. Larger r_d indicates stronger association; W describes the magnitude of across-target variation. Separately, nearest-demonstration distance compares component-wise sums of command increments with the closest of ten demonstrations at the same target, independently for translation and rotation. It uses demonstrated targets only: four depths and five directions. Lower distance indicates closer command agreement (details in Appendix B.2.3).

4.3 CROSS-TASK SHAPE ACCURACY AND TARGET INTERPOLATION

We first evaluate closed-loop shape accuracy, target generalization, and deployment repeatability on phantoms. Closed-loop control reduces Shape RMSE in 512/576 phantom trials: 219/252 depth and 293/324 directional rollouts (Fig. 3). All four voxelized point-cloud configurations improve in at least 19/21 depth trials and all 27 directional trials, totaling 187/192, indicating improvement across action models and conditioning choices. The lowest-error configuration differs by task: full-volume Explicit with residuals achieves the lowest depth error (0.508 ± 0.008 mm), while voxelized point-cloud Explicit with residuals achieves the lowest directional error (0.443 ± 0.013 mm) and a depth error of 0.534 ± 0.044 mm. The latter provides strong cross-task performance, with mean paired error reductions of 32.4% and 47.6% for depth and direction, respectively. Voxelized point-cloud

Explicit with residuals maintains accuracy at intermediate targets, with unseen versus seen errors of 0.537 versus 0.531 mm for depth and 0.427 versus 0.457 mm for direction, supporting interpolation within the tested target families. Across three deployment sweeps, its target-averaged final RMSE ranges 0.508–0.584 mm for depth and 0.435–0.458 mm for direction, supporting repeatable closed-loop control across phantom repositioning and newly acquired predeformation scans.

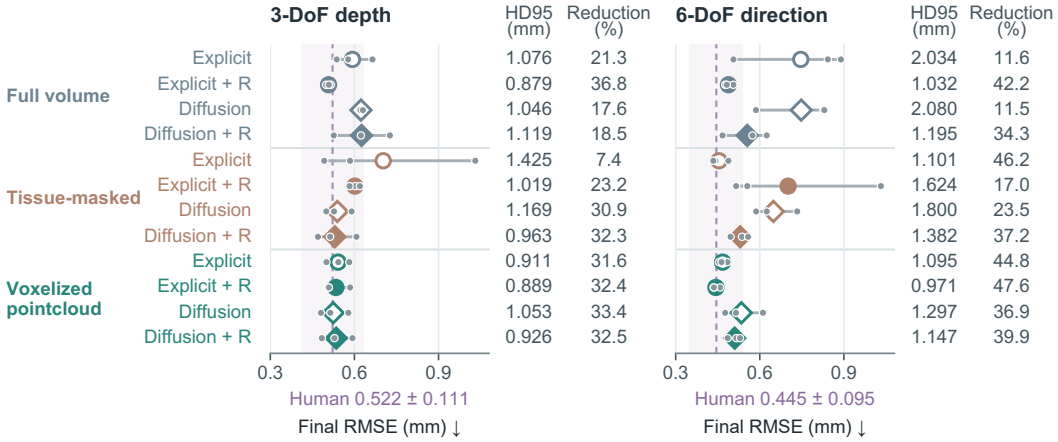


Figure 3: **Phantom shape accuracy and reliability.** Large symbols show mean final RMSE; three gray dots and their range show target-averaged deployment sweeps. Adjacent columns report mean final HD95 (details in Appendix B) and per-rollout RMSE reduction. Purple dashed lines and bands show human endpoint means and reported ± 1 sample SD of final Shape RMSE across individual human demonstrations. Circles: Explicit; diamonds: Diffusion; filled: residual conditioning (R).

4.4 Effect of residual conditioning

Residual conditioning lowers final RMSE in 9/12 matched comparisons, with the largest gains in full-volume directional control and recovery of target-dependent responses. Full-volume Explicit improves from 0.746 to 0.488 mm and Diffusion from 0.748 to 0.555 mm, with improving-rollout counts increasing from 14/27 to 27/27 and 19/27 to 26/27, respectively. Voxelized point-cloud Explicit shows smaller gains, from 0.541 to 0.534 mm for depth and 0.467 to 0.443 mm for direction.

Figure 4 separates target responsiveness from agreement with demonstrated commands. For Tissue-masked Diffusion in directional control, adding residuals increases translation spread from 0.063 to 6.403 mm and target association from 0.603 to 0.874, alongside lower shape error (0.648 to 0.530 mm). These increases indicate that endpoint translation commands vary more across targets and are more strongly associated with the requested direction. In depth control, nearest-demonstration distance decreases in all six matched pairs.

Stronger target responses do not necessarily improve shape accuracy. Full-volume Diffusion's depth association increases from 0.456 to 0.955 and spread from 0.340 to 0.920 mm, while RMSE remains nearly unchanged (0.623 versus 0.625 mm). For tissue-masked Explicit, directional translation association rises from 0.810 to 0.919, but nearest-demonstration distance increases from 2.337 to 3.859 mm and shape error from 0.455 to 0.701 mm. Thus, residuals can strengthen target-dependent commands, but their geometric benefit depends on the resulting tissue deformation.

4.5 Explicit policy versus diffusion-based action prediction

Diffusion has lower depth RMSE in three of six matched comparisons, while Explicit has lower directional RMSE in five of six. With point-cloud residual inputs, the depth means of the two action models are nearly equal (0.534 versus 0.535 mm), but Explicit has lower directional error (0.443 versus 0.510 mm) and smaller nearest-demonstration distances in all three command measures. For full-volume residual inputs, Diffusion instead has closer directional command agreement, while Explicit achieves lower shape error (0.488 versus 0.555 mm). Action-model choice therefore interacts with the representation and evaluation criterion. For voxelized point-cloud inputs with residuals, Ex-

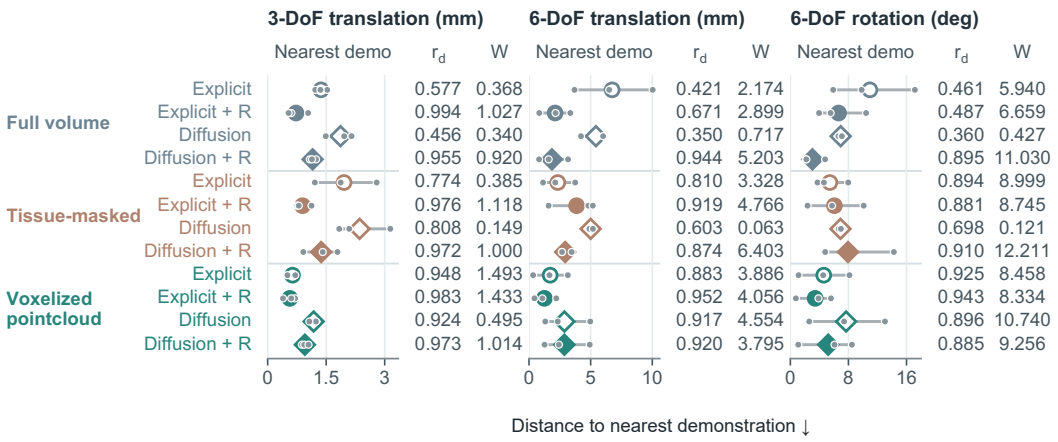


Figure 4: **Target response and agreement with human demonstrations.** Nearest-demonstration distances ("Nearest Demo") show the mean across three deployment sweeps in color, with sweep means and ranges in gray. Columns report pooled r_d and W across all targets, while demonstration distances use seen targets only. Units apply to distances and W; symbols follow Fig. 3.

plicit policy combines comparable depth accuracy, lower directional error, and faster prepared-input inference: 2.42 ms versus 94.26 ms for Diffusion on the timed depth checkpoints (Appendix A.6).

4.6 Ex vivo porcine demonstration study

To extend the phantom evaluation to biological tissue, we evaluate the framework on *ex vivo* porcine tissue. Separately trained Full-volume Diffusion and Point-cloud Explicit policies, both with residuals, are tested at five depths (2–4 mm) and five directions (0°–90°) in three deployment sweep of 60 rollouts. Target width is 3 mm; directional targets use depth 5 mm and tilt strength 0.7. Intermediate depths (2.5, 3.5 mm) and directions (22.5°, 67.5°) are absent from demonstrations.

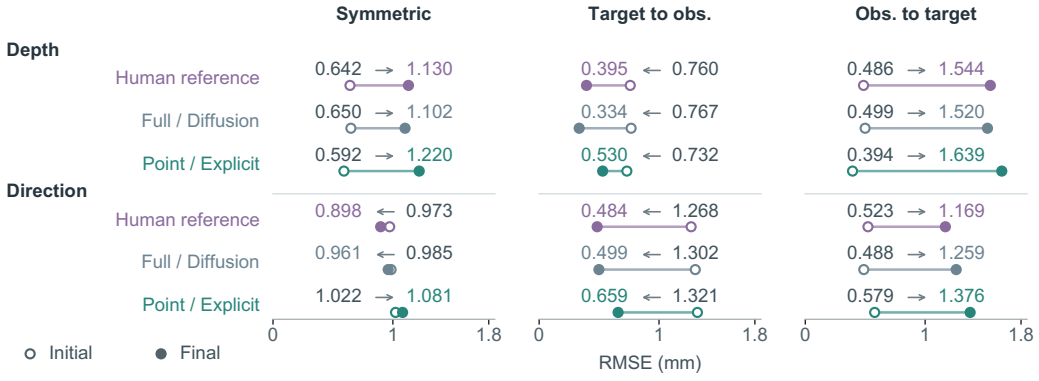


Figure 5: **Porcine tissue shape error before and after control.** Open and filled markers show initial and final mean RMSE, respectively; arrows indicate the direction of error change after control. Symmetric denotes Shape RMSE combining both distance directions; obs. denotes observed tissue geometry. Human references include 30 demonstrations per task, showing policy and human errors on a similar scale and with similar trends across conditions, as indicated by arrow orientation.

Figure 5 compares policy outcomes with human demonstration references. For depth control, final Shape RMSEs are 1.102 and 1.220 mm for the two policies versus 1.130 mm for human demonstrations. For directional shaping, the corresponding errors are 0.961 and 1.081 mm versus 0.898 mm for human demonstrations. Despite the complex *ex vivo* tissue geometry, policy and human demonstration errors show similar trends across most conditions and remain on a similar scale. Both policies also show strong target-dependent directional command responses (Appendix Table 15).

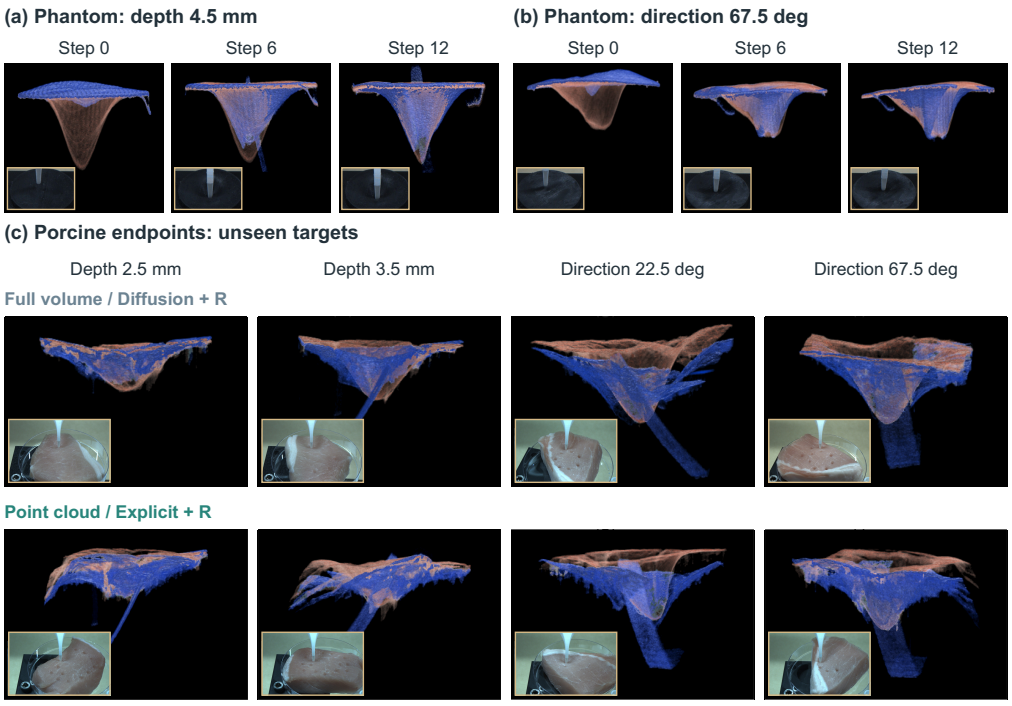


Figure 6: **Phantom deformation states and porcine endpoint geometry.** (a,b) OCT-derived geometry at initial, intermediate, and final phantom states. (c) Porcine endpoints at unseen targets. Blue: observed tissue; brown: fixed target; insets: RGB observations. The porcine overlays show remaining differences in depression width and surrounding geometry.

Target-to-observation RMSE decreases in 58/60 policy rollouts and all 60 human demonstrations, whereas observation-to-target RMSE increases in every trajectory in both cohorts (Table 14). This paired pattern reflects coupled tissue deformation: approaching the target depression can simultaneously alter its width, sidewalls, and surrounding geometry, as illustrated in Fig. 6c.

5 Discussion and Conclusion

In this work, we develop an OCT-guided, data-driven framework for closed-loop tissue shape control, which uses volumetric OCT to represent current and target tissue states and learns incremental tool motions from user-guided demonstrations. Across full-volume, tissue-masked sparse-volume, and voxelized point-cloud representations, OCT supports both deterministic (Explicit) and generative diffusion-based (Diffusion) action prediction. Voxelized point-cloud Explicit with residual conditioning provides accurate shape matching, performance at intermediate unseen targets, and repeatable deployment, while residual conditioning substantially improves full-volume policies with both action models. Porcine experiments further extend the framework to biological tissue, where learned policies produce target-dependent deformation with shape errors on a similar scale to human demonstrations.

However, several limitations remain. Model inference requires approximately 182–403 ms for full-volume, 35–615 ms for tissue-masked sparse-volume, and 2–94 ms for voxelized point-cloud representations across Explicit and Diffusion configurations (Appendix of Table 6). Combined with OCT acquisition and robot execution latency, this cost limits the closed-loop update rate. The current experiments also focus on short-horizon, localized tissue shaping and do not address longer interventions requiring temporal context and coordinated subtasks. Future work will investigate more efficient OCT representations and inference, long-horizon control, and anatomical and task-specific constraints for target generation. In summary, these results provide a foundation for integrating volumetric medical imaging, goal-conditioned learning, and generative action policies for more complex image-guided tissue manipulation.

AI USE STATEMENT

AI tools were used to assist with language editing and identifying relevant literature. The scientific content and initial manuscript drafts were developed by the authors, and all suggested references were independently reviewed and verified for relevance and accuracy. The authors reviewed all AI-assisted content and take full responsibility for the final text, claims, citations, and results.

ETHICS STATEMENT

This work evaluates robotic tissue manipulation using physical tissue phantoms and *ex vivo* porcine tissue. The study focuses on engineering development of OCT-guided robotic manipulation with data collected from user demonstrations. All experimental procedures and data collection were conducted in accordance with applicable institutional research and safety requirements.

REPRODUCIBILITY STATEMENT

We provide implementation details to support reproducibility. The main text describes the OCT representations, goal-conditioning strategies, policy formulations, and closed-loop control procedure. Appendix A.2 details parameterized OCT target generation, while Appendix A provides OCT acquisition, preprocessing, training, and implementation details. Appendix B.1 specifies the phantom evaluation protocol, target configurations, deployment procedure, and evaluation metrics. Additional experimental results and analyses are provided in the appendix and supplementary materials.

REFERENCES

Demir Arikan, Peiyao Zhang, Michael Sommersperger, Shervin Dehghani, Mojtaba Esfandiari, Russell H. Taylor, M. Ali Nasseri, Peter Gehlbach, Nassir Navab, and Iulian Iordachita. Real-time deformation-aware control for autonomous robotic subretinal injection under iOCT guidance. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10531–10537, 2025. doi: 10.1109/ICRA55743.2025.11128595. URL https://arxiv.org/abs/2411.06557.

Gui-Bin Bian, Yawen Deng, Zhen Li, Qiang Ye, Yupeng Zhai, Yong Huang, Yingxiong Xie, Weihong Yu, Zhangwanyu Wei, and Zhangguo Yu. Autonomous robotic intraocular surgery for targeted retinal injections. *Science Robotics*, 11(110):eadx7359, 2026. doi: 10.1126/scirobotics.adx7359. URL https://pubmed.ncbi.nlm.nih.gov/41533804/.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin CM Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems*, 2023. doi: 10.15607/RSS.2023.XIX.026. URL https://www.roboticsproceedings.org/rss19/p026.html.

Mark Draelos, Gao Tang, Brenton Keller, Anthony Kuo, Kris Hauser, and Joseph A Izatt. Optical coherence tomography guided robotic needle insertion for deep anterior lamellar keratoplasty. *IEEE transactions on bio-medical engineering*, 67(7):2073, 2019.

Lesley-Ann Duflot, Brahim Tamadazte, Nicolas Andreff, and Alexandre Krupa. Wavelet-based visual servoing using OCT images. In *2018 7th IEEE International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, pp. 621–626, 2018. doi: 10.1109/BIOROB.2018.8487223. URL https://inria.hal.science/hal-01826328.

William Edwards, Gao Tang, Yuan Tian, Mark Draelos, Joseph Izatt, Anthony Kuo, and Kris Hauser. Data-driven modelling and control for robot needle insertion in deep anterior lamellar keratoplasty. *IEEE Robotics and Automation Letters*, 7(2):1526–1533, 2022. doi: 10.1109/LRA.2022.3140458. URL https://ieeexplore.ieee.org/document/9670688.

Raymond Fang, Pengpeng Zhang, Tingwei Zhang, Zihang Yan, Daniel Kim, Edison Sun, Roman Kuranov, Junghun Kweon, Alex S. Huang, and Hao F. Zhang. Robotic optical coherence tomography with expanded three-dimensional field-of-view using point-cloud-based volumetric montaging. *IEEE Transactions on Medical Imaging*, 44(11):4639–4651, 2025. doi: 10.1109/TMI.2025.3580383. URL https://pubmed.ncbi.nlm.nih.gov/40526559/.

Nils Gessert, Jens Beringhoff, Christoph Otte, and Alexander Schlaefer. Force estimation from OCT volumes using 3D CNNs. *International Journal of Computer Assisted Radiology and Surgery*, 13 (7):1073–1082, 2018. doi: 10.1007/s11548-018-1777-8. URL https://arxiv.org/abs/1804.10002.

Nils Gessert, Marcel Bengs, Matthias Schlüter, and Alexander Schlaefer. Deep learning with 4D spatio-temporal data representations for OCT-based force estimation. *Medical Image Analysis*, 64:101730, 2020. doi: 10.1016/j.media.2020.101730. URL https://arxiv.org/abs/2005.10033.

David Huang, Eric A Swanson, Charles P Lin, Joel S Schuman, William G Stinson, Warren Chang, Michael R Hee, Thomas Flotte, Kenton Gregory, Carmen A Puliafito, and James G Fujimoto. Optical coherence tomography. *Science*, 254(5035):1178–1181, 1991. doi: 10.1126/science.1957169.

Brenton Keller, Mark Draelos, Kevin Zhou, Ruobing Qian, Anthony N. Kuo, George Konidaris, Kris Hauser, and Joseph A. Izatt. Optical coherence tomography-guided robotic ophthalmic microsurgery via reinforcement learning from demonstration. *IEEE Transactions on Robotics*, 36(4):1207–1218, 2020. doi: 10.1109/TRO.2020.2980158. URL https://doi.org/10.1109/TRO.2020.2980158.

Sizhe Li, Zhiao Huang, Tao Chen, Tao Du, Hao Su, Joshua B. Tenenbaum, and Chuang Gan. DexDeform: Dexterous deformable object manipulation with human demonstrations and differentiable physics. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=LIV7-_7pYPl.

Guiqiu Liao, Sujit Kumar, Phillippe Zanne, Benoit Rosa, Florent Nageotte, Michel De Mathelin, Diego Dall'Alba, and Michalina J Gora. Automatic oct scanning of moving soft tissue using continuum endoscopic robots. *IEEE Transactions on Medical Robotics and Bionics*, 2026.

Peng Liu, Jonas Golde, Joseph Morgenstern, Sebastian Bodenstedt, Chenpan Li, Yujia Hu, Zhaoyu Chen, Edmund Koch, Marcus Neudert, and Stefanie Speidel. Non-rigid point cloud registration for middle ear diagnostics with endoscopic optical coherence tomography. *International Journal of Computer Assisted Radiology and Surgery*, 19(1):139–145, 2024. doi: 10.1007/s11548-023-02960-9. URL https://link.springer.com/article/10.1007/s11548-023-02960-9.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Guangshen Ma, Morgan McCloud, Yuan Tian, Amit Narawane, Harvey Shi, Robert Trout, Ryan P McNabb, Anthony N Kuo, and Mark Draelos. Robotics and optical coherence tomography: current works and future perspectives. *Biomedical Optics Express*, 16(2):578–602, 2025.

David Navarro-Alarcon and Yun-Hui Liu. Fourier-based shape servoing: A new feedback method to actively deform soft objects into desired 2-D image contours. *IEEE Transactions on Robotics*, 34 (1):272–279, 2018. doi: 10.1109/TRO.2017.2765333. URL https://ieeexplore.ieee.org/document/8106734/.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171. PMLR, 2021. URL https://proceedings.mlr.press/v139/nichol21a.html.

Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. URL https://arxiv.org/abs/1612.00593.

Moritz Reuss, Maximilian Li, Xiaogang Jia, and Rudolf Lioutikov. Goal-conditioned imitation learning using score-based diffusion policies. In *Proceedings of Robotics: Science and Systems*, 2023. doi: 10.15607/RSS.2023.XIX.028. URL https://www.roboticsproceedings.org/rss19/p028.html.

Paul Maria Scheikl, Nicolas Schreiber, Christoph Haas, Niklas Freymuth, Gerhard Neumann, Rudolf Lioutikov, and Franziska Mathis-Ullrich. Movement primitive diffusion: Learning gentle robotic manipulation of deformable objects. *IEEE Robotics and Automation Letters*, 9(6): 5338–5345, 2024. doi: 10.1109/LRA.2024.3382529.

Haochen Shi, Huazhe Xu, Zhiao Huang, Yunzhu Li, and Jiajun Wu. RoboCraft: Learning to see, simulate, and shape elasto-plastic objects with graph networks. In *Proceedings of Robotics: Science and Systems*, 2022. URL https://arxiv.org/abs/2205.02909.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Yankui Sun, Tian Zhang, Yue Zhao, and Yufan He. 3D automatic segmentation method for retinal optical coherence tomography volume data using boundary surface enhancement. *Journal of Innovative Optical Health Sciences*, 9(2):1650008, 2016. doi: 10.1142/S1793545816500085. URL https://arxiv.org/abs/1508.00966.

Bao Thach, Brian Y. Cho, Alan Kuntz, and Tucker Hermans. Learning visual shape control of novel 3D deformable objects from partial-view point clouds. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2022. URL https://arxiv.org/abs/2110.04685.

Bao Thach, Tanner Watts, Shing-Hei Ho, Tucker Hermans, and Alan Kuntz. DefGoalNet: Contextual goal learning from demonstrations for deformable object manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3145–3152, 2024. URL https://arxiv.org/abs/2309.14463.

Bao Thach, Tanner Watts, Siyeon Kim, Britton Jordan, Mohanraj Devendran Shanthi, Shing-Hei Ho, James M. Ferguson, Tucker Hermans, and Alan Kuntz. DiffDef: A diffusion model for generating multimodal goal shapes from demonstrations for deformable object manipulation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026. URL https://arxiv.org/abs/2506.18779v2.

Alexandre H. Thiéry, Fabian Braeu, Tin A. Tun, Tin Aung, and Michaël J. A. Girard. Medical application of geometric deep learning for the diagnosis of glaucoma. *Translational Vision Science & Technology*, 12(2):23, 2023. doi: 10.1167/tvst.12.2.23. URL https://pubmed.ncbi.nlm.nih.gov/36790820/.

Shengyin Wang, Rafael Papallas, Matteo Leonetti, and Mehmet Dogar. Goal-conditioned action space reduction for deformable object manipulation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

Rosa Wolf, Yitian Shi, Sheng Liu, and Rania Rayyes. Diffusion models for robotic manipulation: A survey. *Frontiers in Robotics and AI*, 12:1606247, 2025.

Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3D diffusion policy: Generalizable visuomotor policy learning via simple 3D representations. In *Proceedings of Robotics: Science and Systems*, 2024. URL https://3d-diffusion-policy.github.io/.

Peiyao Zhang, Ji Woong Kim, Peter Gehlbach, Iulian Iordachita, and Marin Kobilarov. Autonomous needle navigation in subretinal injections via iOCT. *IEEE Robotics and Automation Letters*, 9(5): 4154–4161, 2024. doi: 10.1109/LRA.2024.3375710. URL https://pubmed.ncbi.nlm.nih.gov/38550718/.

Tianxing Zhou, Feiyang Xue, Zhangchen Ye, Tianyuan Yuan, Hang Zhao, and Tao Jiang. Volumedp: Modeling volumetric representation for manipulation policy learning. *arXiv preprint arXiv:2603.17720*, 2026.

A Implementation and Reproducibility

A.1 Demonstrations and dataset partitions

In this study, we constructed the training and validation datasets from user-guided manipulation demonstrations. Each demonstration is one complete target-conditioned manipulation trajectory,

$$\mathcal{D}_i = \left(G_i,\, \{V_{i,t}\}_{t=0}^{L_i},\, \{\mathbf{p}_{i,t}\}_{t=0}^{L_i},\, \{\mathbf{a}_{i,t}\}_{t=0}^{L_i-1} \right), \tag{6}$$

where G_i is the fixed target volume, $V_{i,0}$ is the predeformation OCT scan, and $V_{i,1:L_i}$ are the post-action OCT observations. The six-dimensional vector $\mathbf{p}_{i,t}$ describes the tooltip pose relative to its initial pose, using translation in millimetres and a rotation vector in degrees. The action $\mathbf{a}_{i,t} \in \mathbb{R}^6$ is the local tooltip command applied at state t, comprising three translational increments in millimetres and three extrinsic Euler-xyz rotational increments in degrees. And each trajectory with L_i actions contains $L_i + 1$ observations and poses. The recorded trajectories may contain 9–15 actions.

Table 1 summarizes 151 demonstrations, 1,563 actions, and 1,714 OCT observations, including the initial scans and excluding the target volumes. Complete trajectories are split approximately 80/20 within each target condition using seed 42. Evaluation uses separately collected closed-loop rollouts, including intermediate target conditions absent from both training and validation.

Table 1: Complete demonstration inventory and trajectory-level partitions.

Dataset	Demonstrations	Train/validation	Actions	Observations
Phantom depth	40	32/8	405	445
Phantom directional	51	41/10	535	586
Porcine depth	30	24/6	302	332
Porcine directional	30	24/6	321	351
Total	151	121/30	1,563	1,714

A.2 OCT Parameterized target generation

We construct target OCT volumes from predeformation scans by extracting visible tissue using adaptive A-scan boundary detection and spatial continuity constraints (Sun et al., 2016). Connectivity filtering and smoothing refine the tissue surface, which is then deformed by a parameterized displacement field. Measured OCT intensities are resampled into the target geometry, preserving specimen-specific structure while enabling controlled target shapes.

Specifically, let $\mathbf{x}$ denote lateral position in millimetres, $\mathbf{c}_b$ the center of the OCT field of view, A the nominal peak deformation depth, and w the target base width. A depth-only target is defined by the Gaussian displacement field:

$$d_{\text{depth}}(\mathbf{x}) = A \exp\left(-\frac{\|\mathbf{x} - \mathbf{c}_b\|_2^2}{2\sigma_w^2} \right), \qquad \sigma_w = \frac{w}{2\sqrt{-2\log q_0}}, \qquad q_0 = 0.05. \tag{7}$$

Here, A is expressed in millimetres, and w is the diameter of the contour at which the displacement equals $0.05A$. The Gaussian tail is not truncated outside this contour.

Directional targets use two additional parameters: the direction ϕ and the tilt strength $\tau \in [0, 1]$. The unit direction vector and displaced peak center are

$$\mathbf{e}_\phi = \begin{bmatrix} \sin\phi \\ -\cos\phi \end{bmatrix}, \qquad \mathbf{c}_p = \mathbf{c}_b + 0.30\,\tau w\,\mathbf{e}_\phi. \tag{8}$$

Thus, $\tau = 0$ leaves the peak at the scan center and recovers the depth-only target, whereas increasing τ moves the peak toward direction ϕ.

To preserve a centered and smooth target base, only the inner Gaussian contours are shifted. For a normalized displacement level $q \in [q_0, 1]$, the contour center and radius are

$$\mathbf{c}(q) = \mathbf{c}_b + \frac{q - q_0}{1 - q_0}\,(\mathbf{c}_p - \mathbf{c}_b), \qquad r(q) = \sigma_w \sqrt{-2\log q}. \tag{9}$$

Consequently, the outer 5% contour remains centered at $\mathbf{c}_b$, while progressively higher contours move toward $\mathbf{c}_p$. Each lateral position is assigned the largest displacement level whose contour contains that position. This produces a smooth single-peaked directional deformation rather than translating the entire Gaussian footprint.

Table 2: Target parameters used for user demonstration collection. The dash denotes a depth-only target. Direction angles follow the native OCT row–column convention.

Dataset	A (mm)	w (mm)	ϕ	τ
Phantom depth	$2, 3, 4, 5$	6	–	–
Phantom directional	3.5	6	$0, 45, 90, 135, 180^\circ$	0.7
Porcine depth	$2, 3, 4$	3	–	–
Porcine directional	5.0	3	$0, 45, 90^\circ$	0.7

Tissue displacement exposes regions without corresponding source-tissue samples. These regions are filled with measured air samples from the predeformation scan rather than zeros or synthetic Gaussian noise, preserving the measured OCT background distribution and avoiding artificial empty bands around the displaced tissue. The generated target retains the native OCT grid and intensity type and remains fixed throughout the demonstration and closed-loop rollout.

Target Volume Construction: We construct the target by resampling the measured OCT intensities according to the deformation field (Fig. 7),

$$G(\mathbf{q}) = V_{\mathrm{pre}}(\mathcal{W}_\eta(\mathbf{q}))\,, \tag{10}$$

where $\mathcal{W}_\eta$ maps target locations to sampling locations in the predeformation volume. For depth deformation, we define $\mathcal{W}_{A,w}(\mathbf{x}, z) = (\mathbf{x}, z - d_{\mathrm{depth}}(\mathbf{x}))$, so each target location samples the original volume at its axial position before displacement. Directional targets use the corresponding asymmetric displacement field. Resampling is applied within the deformed tissue region, and measured air samples fill newly exposed and out-of-bounds regions.

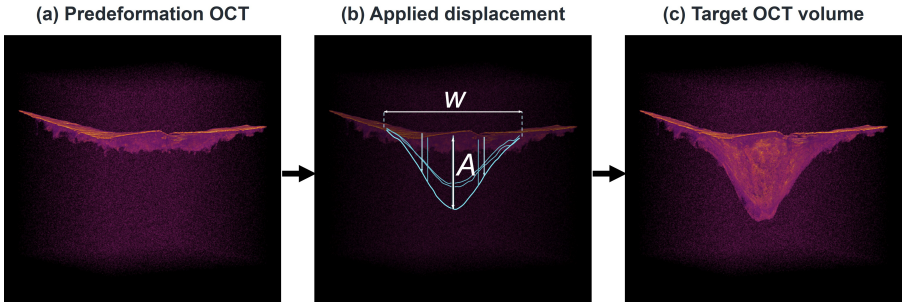


Figure 7: **Parameterized OCT target generation.** A measured predeformation OCT volume is transformed into a target tissue morphology using a parameterized displacement field with deformation depth A and width w defined in Equation 7. The resulting target volume preserves the measured OCT structure within the deformed geometry.

A.3 OCT PREPROCESSING AND INPUT REPRESENTATIONS

We used a 1064 nm, 400 kHz swept-source OCT system (Thorlabs) with a fiber-based Mach–Zehnder interferometer for volumetric acquisition. The interference signal was detected by a balanced photoreceiver and digitized at 1 GS s^{-1} using an ATS-9364 digitizer (AlazarTech).

For the OCT data processing, the raw observations and target volumes are stored as signed `int8` arrays of shape $[1, 256, 256, 768, 1]$. After removing singleton dimensions, the array order is row–column–depth (HWD). The lateral voxel spacing is 0.12857 mm for phantom scans and 0.05298 mm for *ex vivo* scans, while the axial voxel spacing is 0.01325 mm and 0.01329 mm, respectively.

Artifact removal. A trajectory-specific artifact mask is estimated once from the predeformation scan and reused for every observation and the fixed target. We first form a coarse occupancy volume by thresholding local bright-voxel fractions and identify three-dimensional connected components. The scanner-top artifact is identified as the shallow component that touches the top boundary and spans most of the lateral field of view, and the resulting mask covers the leading axial band. For phantom data, the lower support plate is identified separately as a deeper component below the tissue that spans the lateral boundary and has a predominantly peripheral footprint. This classification is not applied to *ex vivo* scans because the *ex vivo* setup contains no lower support plate. Masked voxels are replaced with samples from measured predeformation air above the tissue surface, excluding both the scanner-top band and a guard interval immediately above the detected tissue surface.

Surface detection and visible-tissue support extraction. The same adaptive surface detector is used for both target generation and representation preprocessing. After artifact removal, the detector estimates the upper tissue surface in each A-scan from local background statistics by requiring a persistent axial intensity rise and retaining the largest laterally depth-consistent surface sheet. The reliable lateral footprint is then hole-filled and morphologically closed to suppress small gaps.

For target generation, the predeformation surface and its reliable footprint define the region over which the parameterized displacement field is applied. The measured OCT volume is inverse-warped according to this deformation, while newly exposed regions are filled with measured air sampled from the predeformation scan.

For the Tissue-masked Sparse Volume and Voxelized Point Cloud representations, we additionally construct a solid visible-tissue support mask. Within the reliable lateral footprint, OCT intensity is evaluated from the detected upper surface to approximately 300 axial indices below. At each lateral location, an adaptive threshold combines a fixed tissue-intensity criterion with measured-air statistics from the predeformation scan, and the deepest satisfying voxel defines the lower tissue boundary. All voxels between the upper and lower boundaries are labeled as tissue, forming a solid visible-tissue support region. This mask represents the OCT-supported visible tissue volume rather than only bright surface voxels or a complete anatomical segmentation. The procedure is applied independently to each current observation and target volume.

Normalization and coordinate convention: OCT intensities are normalized using the fixed global constants,

$$\widetilde{I} = \mathrm{clip}\left(\frac{I - \mu_I}{\sigma_I}, -5, 5\right), \qquad \mu_I = 10.377689, \quad \sigma_I = 6.639139. \tag{11}$$

The same fixed transformation is used during training, validation, and closed-loop inference. Volumes are converted from HWD to depth–height–width (DHW) order and flipped along the established scanner X axis. Observations and targets use the same cleanup, normalization, orientation, and representation-specific resampling procedure.

Full Volume: The Full Volume representation retains the cleaned tissue and measured air background as a dense OCT intensity volume. The normalized volume is trilinearly resampled to a fixed $256 \times 256 \times 256$ voxel grid. No tissue mask is provided as an additional input channel.

Tissue-masked Sparse Volume: The Tissue-masked Sparse Volume retains OCT intensities only within the visible-tissue mask. After resampling to a fixed $72 \times 128 \times 128$ DHW grid, each occupied voxel is represented by its integer grid coordinate, $\mathbf{C} \in \mathbb{Z}^{N\times 3}$, and normalized OCT intensity, $\mathbf{F} \in \mathbb{R}^{N\times 1}$, forming an $N \times 4$ sparse representation. The number of occupied voxels N varies with tissue occupancy. A sparse 3D convolutional encoder processes the occupied voxels and their local neighborhoods and pools them into a fixed 128-dimensional feature.

Voxelized Point Cloud: The Voxelized Point Cloud is derived from the same visible-tissue mask and retains geometry only. We sample 1,024 occupied voxel locations from the $72 \times 128 \times 128$ grid, reorder their DHW indices into XYZ coordinates, and normalize them to $[-1, 1]^3$ using the fixed grid bounds. The resulting 1024×3 representation contains only normalized tissue voxel coordinates, without OCT intensity.

Table 3: Policy inputs after preprocessing. B denotes batch size; N_t and N_g denote the variable numbers of occupied sparse voxels in an observation and target, respectively. Sparse entries are shown per frame within one policy sample.

Input	Observation history	Fixed target
Full volume	$[B, 3, 256, 256, 256]$	$[B, 256, 256, 256]$
Tissue-only	Three $\{[N_t, 3]_{\text{coord}}, [N_t, 1]_{\text{intensity}}\}$ sets	One $\{[N_g, 3]_{\text{coord}}, [N_g, 1]_{\text{intensity}}\}$ set
Voxelized Point cloud	$[B, 3, 1024, 3]$	$[B, 1024, 3]$
Relative tool pose	$[B, 3, 6]$	–

Tooltip pose input and action units. At time t, the pose input is derived from the tooltip transform $T_{i,0}^{-1}T_{i,t}$ and represented by a three-dimensional translation in millimetres and a three-dimensional rotation vector in degrees. All six pose components are divided by 20 before entering the model. Action labels contain local translational increments in millimetres and extrinsic Euler-xyz rotational increments in degree and all six action components are divided by 2.

A.4 Representation encoders and action models

All condition tokens have width 128. Table 4 summarizes the representation encoders and token counts. For Full Volume inputs, the three observations share one dense encoder, while the fixed target uses a separate encoder; residual conditioning adds a third encoder for residual volumes. For the Tissue-masked Sparse Volume, observations and target share a sparse encoder, with an additional sparse encoder for target–observation residuals. For the Voxelized Point Cloud, observations and target share a PointNet encoder, while residuals are computed by subtracting their embeddings.

A $6 \to 128 \to 128$ pose MLP with SiLU produces one token per history state. Temporal and token-type embeddings distinguish observation, target, pose, and residual tokens. Full Volume tokens additionally receive fixed 3D Fourier spatial features over the pooled 4^3 token grid. Tissue-masked sparse-volume inputs use integer grid coordinates to locate occupied voxels and normalized OCT intensity as the voxel feature; voxelized point-cloud inputs use scanner-fixed normalized XYZ coordinates.

Table 4: Representation encoders and total condition-token counts without/with residual conditioning. Counts include three observations, one fixed target, and three relative tooltip poses.

Input	Encoder	Tokens
Full volume	Four dense 3-D convolutional residual stages with channels $16/32/64/128$, GroupNorm, SiLU, and adaptive average pooling to 4^3 tokens per volume	259 / 451
Tissue-only	Sparse 3-D encoder with 27-neighbor blocks of channels $32/64/128/128$ and strides $1/2/2/2$, followed by global mean pooling over active voxels	7 / 10
Voxelized point cloud	Pointwise MLP $3 \to 64 \to 128 \to 128$ with ReLU, max pooling over 1,024 points, and pooled MLP $128 \to 128 \to 128$	7 / 10

Each Full Volume input contributes $4^3 = 64$ tokens. Three observations, one target, and three pose tokens yield $3(64) + 64 + 3 = 259$ tokens; three residual volumes increase the total to 451. Tissue-masked Sparse Volume and Voxelized Point Cloud inputs each produce one token per observation, target, or residual, yielding seven tokens without residual conditioning and ten with it.

Explicit Policy. The Explicit Policy is a deterministic direct action-regression decoder. Eight learned action queries, corresponding to the eight prediction-horizon steps, attend to the condition tokens through one four-head cross-attention layer. The attended features are added to the learned queries and passed through LayerNorm and an MLP ($128 \to 512 \to 6$, with SiLU), producing a normalized 8×6 action chunk in one forward pass.

Diffusion Policy. The Diffusion policy predicts the clean normalized 8×6 action chunk from a noisy sequence using a 100-step cosine diffusion schedule (Nichol & Dhariwal, 2021). Noisy

actions are projected to width 128 and combined with learned action-position embeddings. A diffusion-timestep embedding conditions four Transformer blocks, each with action self-attention, cross-attention to the condition tokens, four attention heads, and feed-forward width 512. Inference starts from Gaussian action noise and uses 20 deterministic DDIM-style updates (Song et al., 2021).

Both policies use the same representation-specific condition tokens. Their normalized outputs are bounded to $[-1, 1]$ and converted back to physical action units defined in Appendix A.3. Fig. 8 illustrates the shared conditioning pipeline, two action decoders, and closed-loop execution procedures.

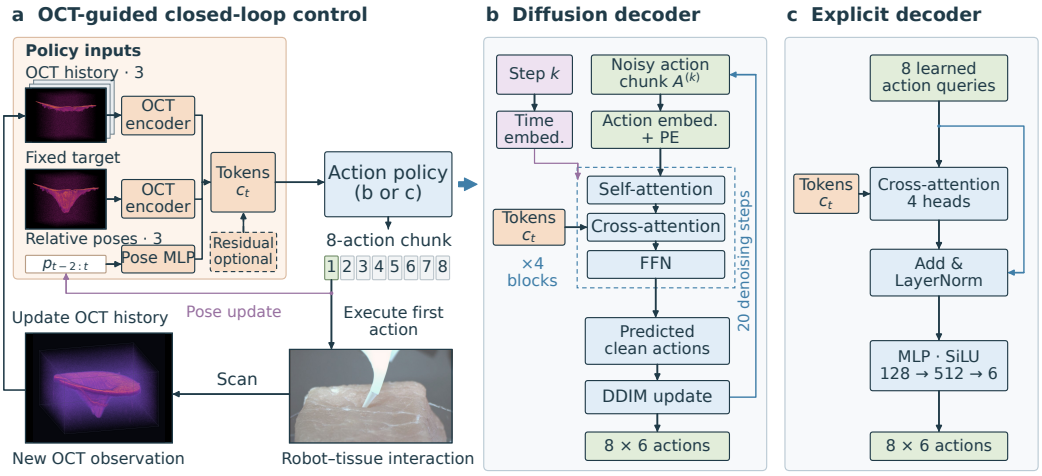


Figure 8: **OCT-guided control and action decoders.** (a) OCT history, fixed targets, relative commanded poses, and optional residuals form condition tokens. The stack denotes three observations. Only the first of eight predicted actions is executed before rescanning; pose history is updated by command composition. (b) Four time-conditioned Transformer blocks predict clean actions for 20 DDIM-style updates. PE denotes learned action-position embeddings. (c) Explicit uses eight learned queries and one four-head cross-attention operation. Query skip addition, LayerNorm, and a SiLU MLP produce the action chunk. Representation encoders and token counts are specified in Table 4.

A.5 Training and Checkpoint Selection

Training samples and objective. We optimize all policies with AdamW (Loshchilov & Hutter, 2019) using the settings in Table 5. Each action defines a sample with history indices $[\max(0, t-2), \max(0, t-1), t]$ and up to eight future actions from t. The oldest observation and pose are repeated to initialize the history. Missing future actions are zero-padded and excluded from the masked six-dimensional MSE in Eq. (4); no terminal-action samples are added.

Table 5: Shared physical-policy training settings. The learning rate is constant.

Setting	Value	Setting	Value
Optimizer	AdamW	Learning rate	10^{-4}
Weight decay	10^{-4}	Batch size	1
Maximum/minimum epochs	100/10	Early-stop patience	10
Minimum improvement	10^{-4}	EMA decay	0.995
Gradient accumulation	None	Training seed	42

Checkpoint selection. Validation and deployment use EMA weights. For normalized expert commands $\bar{\mathbf{a}}_{i,t} = \mathbf{a}_{i,t}/2$, selection uses trajectory-macro first-action MSE:

$$E_{\text{val}} = \frac{1}{n_{\text{val}}} \sum_{i=1}^{n_{\text{val}}} \frac{1}{6L_i} \sum_{t=0}^{L_i-1} \|\widehat{\mathbf{a}}_{i,t,0} - \bar{\mathbf{a}}_{i,t}\|_2^2. \tag{12}$$

Normalized MSE is averaged within each validation trajectory and then equally across trajectories. Diffusion validation generates one chunk per sample from pure noise, resetting the sampling seed to 42 for each validation pass. A checkpoint replaces the current best only when this score improves by more than 10^{-4}. Early stopping uses the settings in Table 5. We train and select one task-specific checkpoint for each configuration: twelve configurations for each phantom task and two for each porcine task, giving 28 evaluated checkpoints.

A.6 Closed-loop execution and model-only runtime

The controller retains a fixed target and a three-state observation/pose history. It predicts eight actions, clips the normalized output to $[-1, 1]$, multiplies the first action by two, and executes it. Component bounds are therefore ± 2 mm and $\pm 2^\circ$. The local tooltip increment is composed as $T_{t+1} = T_t \Delta T_t$ and converted to an end-effector command for inverse kinematics. A new OCT scan and the updated commanded pose enter the next prediction. Pose history is propagated from the initial robot pose by command composition. Each quantitative rollout executes 12 actions and stores every step scan.

Table 6: Prepared-input model-only latency (mean $\pm$ sample SD, ms) for the twelve phantom depth checkpoints on an RTX 4500 Ada GPU: one fixed input per checkpoint, batch size one, FP16 autocast, five warm-ups, and 20 CUDA-synchronized calls.

Representation / conditioning	Explicit	Diffusion
Full volume	181.67 ± 0.08	272.85 ± 0.54
Full volume + residual	311.44 ± 0.17	402.91 ± 0.40
Tissue-only	35.06 ± 1.44	140.16 ± 14.73
Tissue-only + residual	517.86 ± 77.71	614.60 ± 39.28
Point cloud	2.16 ± 0.23	91.16 ± 0.57
Point cloud + residual	2.42 ± 0.11	94.26 ± 0.79

These values measure prepared-input model inference rather than the full control cycle. Timing includes encoders, residual operations, and the complete action head, including all 20 Diffusion updates and internal sparse bookkeeping. It excludes checkpoint loading, file I/O, preprocessing, representation extraction, OCT acquisition, inverse kinematics, and robot motion.

B Additional Real-Robot Results

This appendix reports the complete numerical results of the real-robot evaluation, including shape accuracy, target-dependent action responses, agreement with demonstrations, and closed-loop progression. Geometric results use the native-grid evaluation of 576 phantom and 60 porcine rollouts. Command analyses retain the recorded phantom and porcine trajectories.

B.1 Evaluation scope and reporting conventions

Evaluation data and target conditions. A target value is *seen* if it occurs in the corresponding demonstration training or validation data. *Unseen* values are intermediate target parameters absent from both partitions. All evaluation rollouts are newly executed closed-loop deployments and are separate from the demonstrations used for training and validation.

The evaluation contains 636 twelve-action rollouts, 7,632 executed actions, and 8,268 OCT states, including the initial state of each rollout. Phantom evaluation contains 252 depth and 324 directional rollouts. Porcine evaluation contains 30 depth and 30 directional rollouts across the two tested configurations.

Repeated deployments. Each phantom policy–target condition is evaluated in three independent deployment sweeps. Between sweeps, the phantom is repositioned to vary its placement height and orientation within the OCT field of view, a new predeformation scan is acquired, and the corresponding target is regenerated from that scan. Deployment seeds are also changed where applicable.

Table 7: Closed-loop evaluation inventory. Each phantom task evaluates all twelve policy configurations. Porcine experiments use task-specific checkpoints for the two configurations shown below. Each rollout contains twelve executed actions and thirteen OCT states, including the initial scan.

Task	Targets	Rollouts	Actions	OCT states
Phantom: all twelve configurations				
Depth	7	252	3,024	3,276
Directional	9	324	3,888	4,212
Porcine: Full-volume Diffusion + residual				
Depth	5	15	180	195
Directional	5	15	180	195
Porcine: Point-cloud Explicit + residual				
Depth	5	15	180	195
Directional	5	15	180	195
Total	–	**636**	**7,632**	**8,268**

For porcine evaluation, each configuration–target condition contains three independently executed rollouts. Because biological tissue cannot be reset to an identical initial state, these recordings are treated as repeated physical deployments rather than identical-state repeats.

Table 8: Phantom and porcine evaluation target conditions. Porcine experiments use task-specific checkpoints for Full-volume Diffusion + residual and Point-cloud Explicit + residual.

Material	Task	Target conditions	Fixed parameters
Phantom	Depth	**Seen:** 2.0, 3.0, 4.0, 5.0 mm **Unseen:** 2.5, 3.5, 4.5 mm	Width: 6 mm
Phantom	Directional	**Seen:** 0°, 45°, 90°, 135°, 180° **Unseen:** 22.5°, 67.5°, 112.5°, 157.5°	Depth: 3.5 mm Width: 6 mm Strength: 0.7
Porcine	Depth	**Seen:** 2.0, 3.0, 4.0 mm **Unseen:** 2.5, 3.5 mm	Width: 3 mm
Porcine	Directional	**Seen:** 0°, 45°, 90° **Unseen:** 22.5°, 67.5°	Depth: 5 mm Width: 3 mm Strength: 0.7

How means and variability are computed. For each policy configuration c, let $m_{c,k,r}$ denote a rollout-level statistic at target condition k in phantom deployment sweep r. For a reported all-target, seen-target, or unseen-target subset containing K target values, we compute

$$\bar{m}_{c,r} = \frac{1}{K}\sum_{k=1}^{K} m_{c,k,r}, \qquad \bar{m}_c = \frac{1}{3}\sum_{r=1}^{3} \bar{m}_{c,r}, \qquad s_{m,c} = \sqrt{\frac{1}{2}\sum_{r=1}^{3} (\bar{m}_{c,r} - \bar{m}_c)^2}. \tag{13}$$

Thus, statistics are reported as the mean $\pm$ sample SD of three target-averaged deployment sweeps for one fixed checkpoint.

Shape RMSE: geometric mismatch. We evaluate tissue–target alignment using the visible-tissue support masks defined in Section A.3. Let P_t and Q denote the sets of all occupied voxel locations within the observed and fixed-target support masks at step t. Each occupied voxel index is converted to physical coordinates using the material-specific lateral and axial voxel spacings reported in Section A.3. For a point $\mathbf{p}$ and point set Q, let $d(\mathbf{p}, Q) = \min_{\mathbf{q}\in Q} \|\mathbf{p} - \mathbf{q}\|_2$. We compute the two

directed nearest-neighbor RMSEs and their symmetric combination:

$$R_{P\to Q,t} = \sqrt{\frac{1}{|P_t|}\sum_{\mathbf{p}\in P_t} d(\mathbf{p}, Q)^2}, \qquad R_{Q\to P,t} = \sqrt{\frac{1}{|Q|}\sum_{\mathbf{q}\in Q} d(\mathbf{q}, P_t)^2},$$
$$D_t = \sqrt{\frac{R_{P\to Q,t}^2 + R_{Q\to P,t}^2}{2}}. \tag{14}$$

We report D_t as Shape RMSE in millimeters; lower values indicate closer geometric agreement with the target. Evaluation is performed directly on the native OCT grid, without registration or policy-input resampling.

HD95: 95th-percentile geometric distance. Using all occupied voxel locations in the same visible-tissue support masks as Shape RMSE, we compute the 95th percentile of nearest-neighbor distances in each direction and report the larger value:

$$H_{95,t} = \max\{\mathrm{quantile}_{0.95}\{d(\mathbf{p}, Q) : \mathbf{p} \in P_t\},\ \mathrm{quantile}_{0.95}\{d(\mathbf{q}, P_t) : \mathbf{q} \in Q\}\}. \tag{15}$$

HD95 is reported in millimeters; lower values indicate closer point-set agreement at the 95th-percentile level. It complements Shape RMSE by describing the larger nearest-neighbor distances, while being less sensitive than the maximum distance to a small number of extreme values.

Reduction and improved counts. For rollout i, we define the percentage reduction in Shape RMSE and an improvement indicator as

$$\rho_i = 100\left(1 - \frac{D_{i,12}}{D_{i,0}}\right), \qquad I_i = \mathbf{1}[D_{i,12} < D_{i,0}]. \tag{16}$$

ρ_i measures the relative change from the initial to final Shape RMSE, and $\sum_i I_i$ reports the number of rollouts whose final Shape RMSE is lower than the initial value.

Endpoint results. Table 9 reports the final shape-matching results for all twelve configurations in the depth and directional tasks.

B.2 Operator shape reference at demonstrated goals

We evaluate the initial and final OCT states of all 151 operator demonstrations using the same material-specific visible-support extraction and geometric metrics as policy deployment. The 30 validation demonstrations serve as the primary operator reference; they are excluded from gradient updates and used only for checkpoint selection. The full 151-trajectory cohort summarizes all collected demonstrations. For matched target comparisons, the Point-cloud Explicit $+R$ policy includes only deployments at target parameters represented in the demonstrations. Each trajectory endpoint is its final recorded state, with goals weighted equally within each material and task. Operator and policy rollouts use distinct initial scans and may contain different numbers of actions.

B.2.1 Shape RMSE over execution

Figures 9 and 10 show Shape RMSE from the initial observation through all twelve executed actions. These curves complement the endpoint results by showing how tissue–target mismatch changes during the rollout.

B.2.2 Target-dependent action responses

Target association and response spread. We evaluate whether changing the requested target produces a corresponding change in the commanded motion. For each rollout, the twelve local commands are composed in the initial tool frame to obtain a final commanded translation vector and, for the 6-DoF task, a final commanded rotation vector.

For each deployment sweep, the seven depth targets or nine directional targets therefore produce a set of endpoint responses. We characterize this target-dependent response using two complementary

Table 9: Complete phantom endpoint results evaluation. D_{12} and H_{95} are in millimeters; ρ is the percentage reduction in Shape RMSE.

Configuration	R	D_{12} all	D_{12} seen	D_{12} unseen	H_{95}	ρ	Improved
3-DoF depth							
Full volume / Explicit	–	0.593 ± 0.065	0.602	0.581	1.076	21.3	17/21
Full volume / Explicit	✓	0.508 ± 0.008	0.499	0.520	0.879	36.8	21/21
Full volume / Diffusion	–	0.623 ± 0.006	0.637	0.606	1.046	17.6	15/21
Full volume / Diffusion	✓	0.625 ± 0.100	0.636	0.612	1.119	18.5	17/21
Tissue-only / Explicit	–	0.703 ± 0.288	0.710	0.693	1.425	7.4	16/21
Tissue-only / Explicit	✓	0.601 ± 0.018	0.590	0.616	1.019	23.2	18/21
Tissue-only / Diffusion	–	0.539 ± 0.046	0.557	0.515	1.169	30.9	18/21
Tissue-only / Diffusion	✓	0.530 ± 0.070	0.533	0.526	0.963	32.3	18/21
Point cloud / Explicit	–	0.541 ± 0.041	0.535	0.550	0.911	31.6	20/21
Point cloud / Explicit	✓	0.534 ± 0.044	0.531	0.537	0.889	32.4	20/21
Point cloud / Diffusion	–	0.524 ± 0.050	0.536	0.509	1.053	33.4	20/21
Point cloud / Diffusion	✓	0.535 ± 0.055	0.543	0.525	0.926	32.5	19/21
6-DoF directional							
Full volume / Explicit	–	0.746 ± 0.209	0.750	0.742	2.034	11.6	14/27
Full volume / Explicit	✓	0.488 ± 0.015	0.502	0.472	1.032	42.2	27/27
Full volume / Diffusion	–	0.748 ± 0.140	0.761	0.732	2.080	11.5	19/27
Full volume / Diffusion	✓	0.555 ± 0.080	0.570	0.537	1.195	34.3	26/27
Tissue-only / Explicit	–	0.455 ± 0.029	0.466	0.440	1.101	46.2	27/27
Tissue-only / Explicit	✓	0.701 ± 0.287	0.707	0.694	1.624	17.0	18/27
Tissue-only / Diffusion	–	0.648 ± 0.076	0.653	0.642	1.800	23.5	27/27
Tissue-only / Diffusion	✓	0.530 ± 0.032	0.538	0.521	1.382	37.2	27/27
Point cloud / Explicit	–	0.467 ± 0.015	0.477	0.455	1.095	44.8	27/27
Point cloud / Explicit	✓	0.443 ± 0.013	0.457	0.427	0.971	47.6	27/27
Point cloud / Diffusion	–	0.534 ± 0.069	0.546	0.519	1.297	36.9	27/27
Point cloud / Diffusion	✓	0.510 ± 0.023	0.513	0.507	1.147	39.9	27/27

Table 10: Shape outcomes at demonstrated goal parameters. D_0 and D_{end} are initial and endpoint Shape RMSE; $H_{95,\text{end}}$ is endpoint HD95 (all in mm). n counts trajectories and "Improved" means $D_{\text{end}} < D_0$. Operator trajectories end after 9–15 recorded actions; policy rollouts execute twelve. All and validation operator rows overlap; the latter are included in the former. Initial scans are distinct across cohorts.

Material / task	Cohort	n	$D_0 \to D_{\text{end}}$	$H_{95,\text{end}}$	Improved	Actions
Phantom / depth	Operators, all	40	$0.849 \to 0.522$	0.831	40/40	10–12
Phantom / depth	Operators, val.	8	$0.849 \to 0.552$	0.835	8/8	10–12
Phantom / depth	Point E+R	12	$0.838 \to 0.531$	0.896	11/12	12
Phantom / direction	Operators, all	51	$0.877 \to 0.445$	0.964	51/51	9–15
Phantom / direction	Operators, val.	10	$0.876 \to 0.421$	0.953	10/10	10–13
Phantom / direction	Point E+R	15	$0.852 \to 0.457$	1.004	15/15	12
Porcine / depth	Operators, all	30	$0.642 \to 1.130$	4.344	0/30	10–11
Porcine / depth	Operators, val.	6	$0.623 \to 1.213$	4.769	0/6	10
Porcine / depth	Point E+R	9	$0.599 \to 1.256$	4.812	0/9	12
Porcine / direction	Operators, all	30	$0.973 \to 0.898$	2.760	21/30	10–14
Porcine / direction	Operators, val.	6	$0.998 \to 0.909$	2.600	6/6	10–14
Porcine / direction	Point E+R	9	$1.029 \to 1.088$	3.075	2/9	12

quantities. First, distance correlation r_d measures whether targets that are more different also tend to produce more different endpoint commands. Larger $r_d \in [0, 1]$ indicates a stronger association between the requested target and the commanded response. Second, response spread W measures how much the endpoint commands actually vary across targets. Larger W indicates greater across-target variation, in millimeters for translation and degrees for rotation.

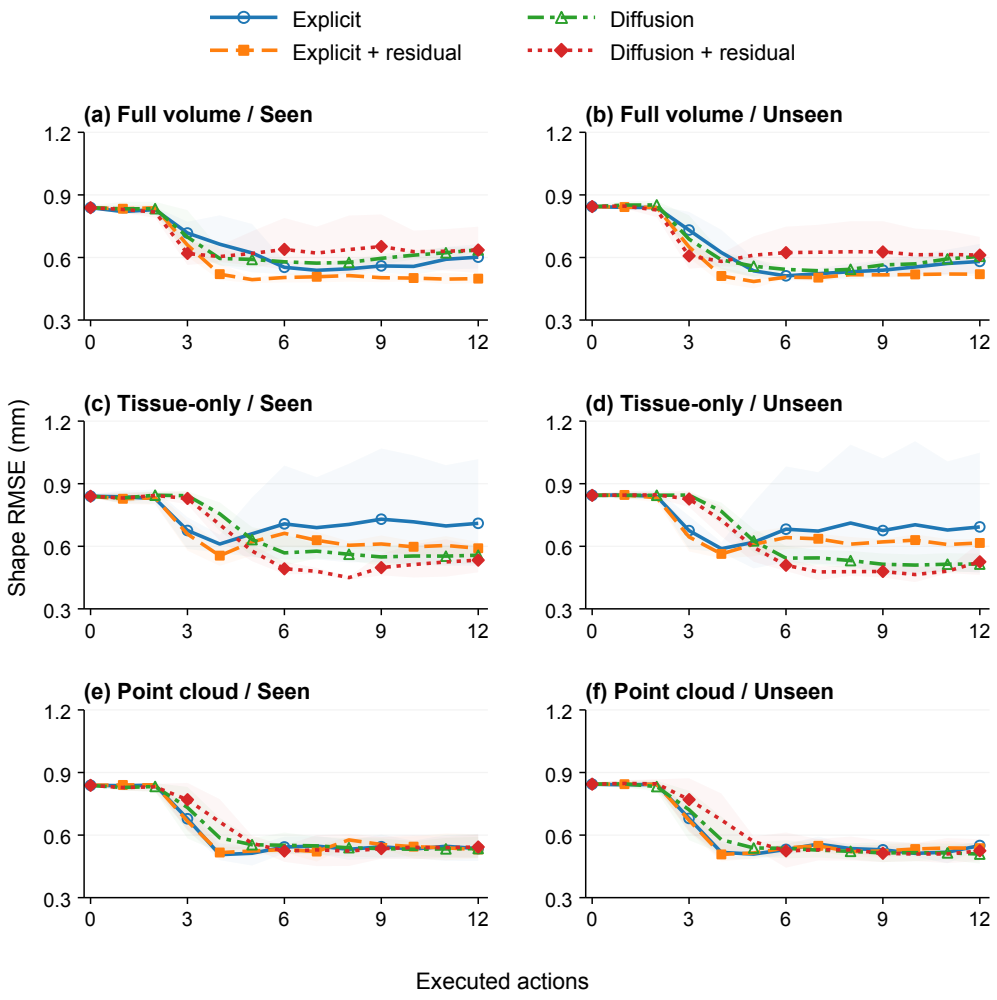


Figure 9: **Phantom depth error across observation steps.** Step 0 is the initial observation; steps 1–12 follow executed actions. Rows show representations and columns show seen/unseen targets. Curves average targets and three sweeps; shading spans sweep means.

Specifically, let $\mathbf{z}_i$ denote the endpoint translation or rotation vector of rollout i. Within each sweep, we subtract the average endpoint over all tested targets,

$$\mathbf{y}_i = \mathbf{z}_i - \frac{1}{K}\sum_{k=1}^{K}\mathbf{z}_k, \tag{17}$$

where $K = 7$ for depth and $K = 9$ for direction. Thus, $\mathbf{y}_i$ measures how the response to target i differs from the typical response in the same sweep.

For target association, depth is represented by its scalar value and direction by $(\cos\phi_i, \sin\phi_i)$. We compare the pairwise distances between target values with the pairwise distances between their corresponding endpoint responses. Thus, r_d is high when target pairs that are farther apart also tend to produce endpoint responses that are farther apart.

Let A and B denote the centered pairwise-distance matrices for the targets and endpoint responses, respectively, following the standard distance-correlation formulation. The target association and

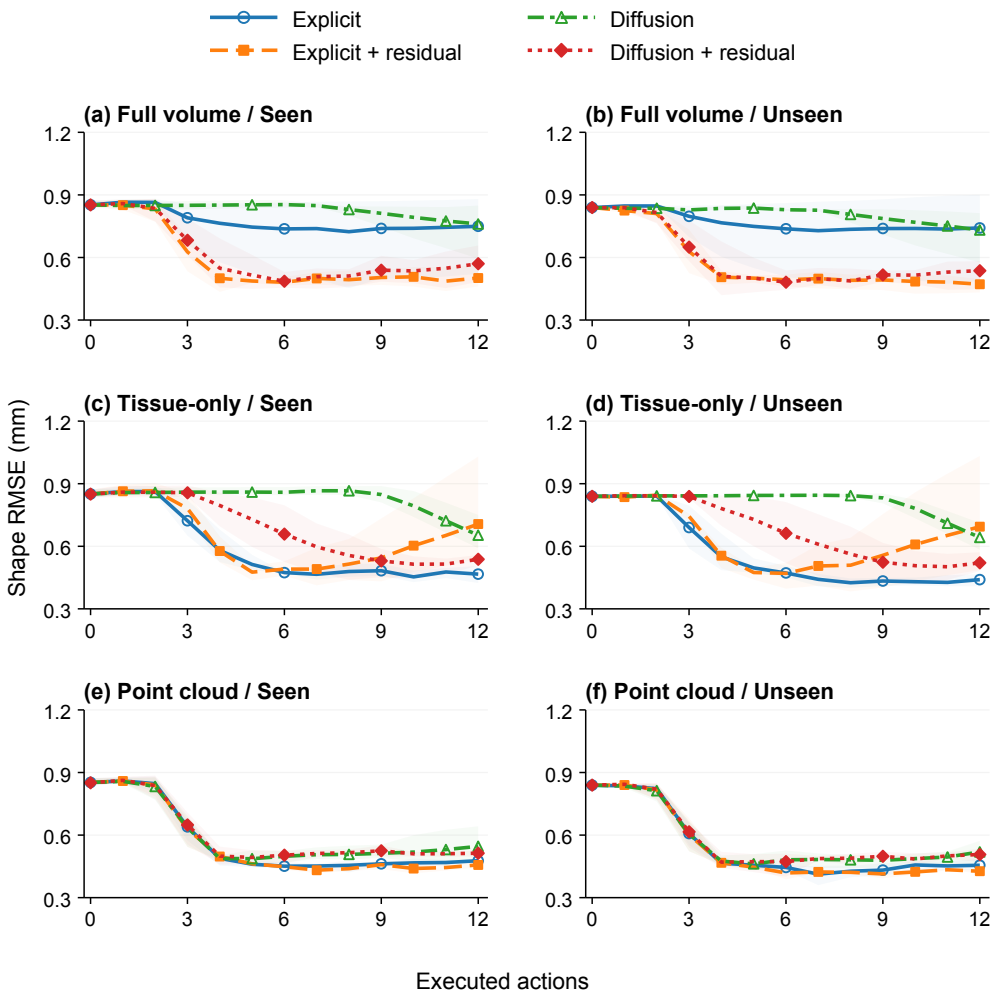


Figure 10: **Phantom directional error at every observation step.** Aggregation and shading follow Fig. 9. Curves show deployments across nine target directions, including four unseen directions.

response spread are

$$r_d = \sqrt{\max\left(0, \frac{\sum_{ij} A_{ij} B_{ij}}{\sqrt{\sum_{ij} A_{ij}^2 \sum_{ij} B_{ij}^2}}\right)}, \qquad W = \sqrt{\frac{1}{n} \sum_i \|\mathbf{y}_i\|_2^2}, \tag{18}$$

Table 11 includes all 21 depth and 27 directional rollouts per configuration, including unseen targets.

B.2.3 Agreement with demonstrations

We evaluate how closely each policy rollout matches the demonstrated commands at the same target. For policy configuration c, target k, and deployment sweep r, we sum the command increments over the twelve executed actions:

$$\mathbf{u}^{\mathrm{tr}}_{c,k,r} = \sum_{t=1}^{12} \Delta\mathbf{p}_{c,k,r,t}, \qquad \mathbf{u}^{\mathrm{rot}}_{c,k,r} = \sum_{t=1}^{12} \Delta\boldsymbol{\theta}_{c,k,r,t}. \tag{19}$$

These component-wise command sums describe accumulated commanded motion. The target-association analysis (Appendix B.2.2) uses endpoints obtained by composing the local motions.

Table 11: Target association and across-target response spread for all phantom configurations. r_d measures the association between target differences and endpoint-response differences; larger is stronger. W measures the RMS spread of endpoint responses across targets and is reported in millimeters for translation and degrees for rotation.

		3-DoF translation		6-DoF translation		6-DoF rotation	
Configuration	R	r_d	W	r_d	W	r_d	W
Full volume / Explicit	–	0.577	0.368	0.421	2.174	0.461	5.940
Full volume / Explicit	✓	0.994	1.027	0.671	2.899	0.487	6.659
Full volume / Diffusion	–	0.456	0.340	0.350	0.717	0.360	0.427
Full volume / Diffusion	✓	0.955	0.920	0.944	5.203	0.895	11.030
Tissue-only / Explicit	–	0.774	0.385	0.810	3.328	0.894	8.999
Tissue-only / Explicit	✓	0.976	1.118	0.919	4.766	0.881	8.745
Tissue-only / Diffusion	–	0.808	0.149	0.603	0.063	0.698	0.121
Tissue-only / Diffusion	✓	0.972	1.000	0.874	6.403	0.910	12.211
Point cloud / Explicit	–	0.948	1.493	0.883	3.886	0.925	8.458
Point cloud / Explicit	✓	0.983	1.433	0.952	4.056	0.943	8.334
Point cloud / Diffusion	–	0.924	0.495	0.917	4.554	0.896	10.740
Point cloud / Diffusion	✓	0.973	1.014	0.920	3.795	0.885	9.256

For demonstration d at target k, we similarly use its complete recorded sequence of length $L_{k,d}$:

$$\mathbf{u}_{k,d}^{\mathrm{tr,D}} = \sum_{t=1}^{L_{k,d}} \Delta\mathbf{p}_{k,d,t}, \qquad \mathbf{u}_{k,d}^{\mathrm{rot,D}} = \sum_{t=1}^{L_{k,d}} \Delta\boldsymbol{\theta}_{k,d,t}. \tag{20}$$

The command-reference analysis uses a balanced subset of ten demonstrations at each of the four phantom depth targets and five phantom directional targets, giving 40 depth and 50 directional reference trajectories. Their average cumulative command is

$$\bar{\mathbf{u}}_k^D = \frac{1}{10} \sum_{d=1}^{10} \mathbf{u}_{k,d}^D, \tag{21}$$

where translation and rotation are evaluated separately.

For each policy rollout, we measure its distance to (1) the average demonstrated command at the same target and (2) the closest individual demonstration:

$$m_{c,k,r}^{\mathrm{avg}} = \left\| \mathbf{u}_{c,k,r} - \bar{\mathbf{u}}_k^D \right\|_2, \qquad m_{c,k,r}^{\mathrm{near}} = \min_{1 \le d \le 10} \left\| \mathbf{u}_{c,k,r} - \mathbf{u}_{k,d}^D \right\|_2. \tag{22}$$

Lower values indicate closer agreement with the demonstrated commands. The average-reference distance measures agreement with the central demonstrated behavior at that target, while the nearest-demonstration distance measures agreement with the most similar recorded demonstration. Only targets with demonstrations are included: four depth targets and five directional targets. Table 12 reports the resulting configuration-level averages.

B.2.4 Step-by-step action trajectories

Figures 11–14 summarize the step-by-step action trajectories using one command component per figure. For the depth task, we report the y translation command because it is the primary deformation axis. For the directional task, we report the three rotation components, roll, pitch, and yaw, to show how rotational control changes across target directions.

Within each figure, columns enumerate all tested targets and rows enumerate the twelve policy configurations grouped by representation and action model. Each policy curve is the mean of the three deployment sweeps at that target. At demonstrated targets, the gray reference curve is the mean demonstration command at each action index.

Table 12: Cumulative-command agreement with demonstrations. "Demo avg." is the distance to the average cumulative command of the ten demonstrations at the same target; "Nearest demo" is the distance to the closest individual demonstration.

		3-DoF translation		6-DoF translation		6-DoF rotation	
Configuration	R	Demo avg.	Nearest demo	Demo avg.	Nearest demo	Demo avg.	Nearest demo
Full volume / Explicit	–	2.032	1.366	7.551	6.743	14.403	10.959
Full volume / Explicit	✓	0.745	0.726	3.200	2.114	9.174	6.643
Full volume / Diffusion	–	2.297	1.867	6.330	5.426	11.056	6.931
Full volume / Diffusion	✓	1.840	1.152	2.566	1.859	4.664	3.037
Tissue-only / Explicit	–	2.213	1.958	3.328	2.337	7.096	5.443
Tissue-only / Explicit	✓	1.346	0.895	4.672	3.859	8.010	6.070
Tissue-only / Diffusion	–	3.248	2.357	5.989	4.999	10.958	6.906
Tissue-only / Diffusion	✓	2.168	1.372	3.727	2.935	9.821	7.943
Point cloud / Explicit	–	0.840	0.637	2.255	1.714	5.850	4.599
Point cloud / Explicit	✓	0.732	0.567	1.604	1.247	4.535	3.413
Point cloud / Diffusion	–	1.881	1.179	3.566	2.872	9.619	7.677
Point cloud / Diffusion	✓	1.646	0.954	3.686	2.871	6.686	5.216

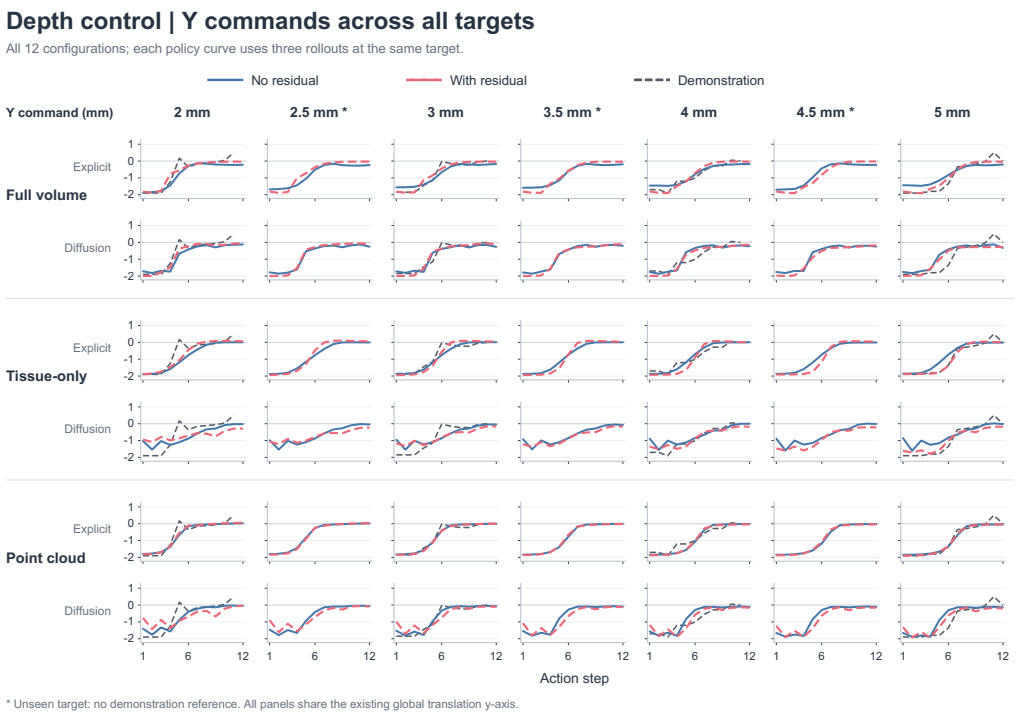


Figure 11: **Depth-task y translation commands across all targets.** Columns show the seven depth targets, with asterisks marking unseen intermediate targets. Rows show all twelve policy configurations grouped by representation and action model.

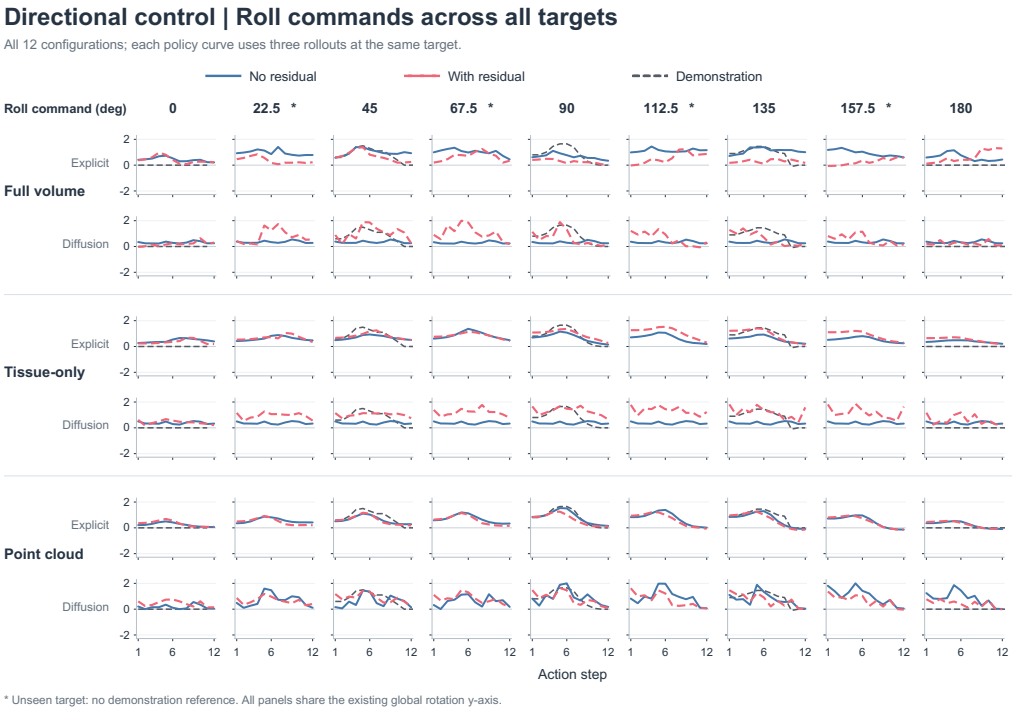


Figure 12: **Directional-task roll commands across all targets.** Columns show the nine tested target directions, with asterisks marking unseen intermediate targets. Rows: twelve policy configurations.

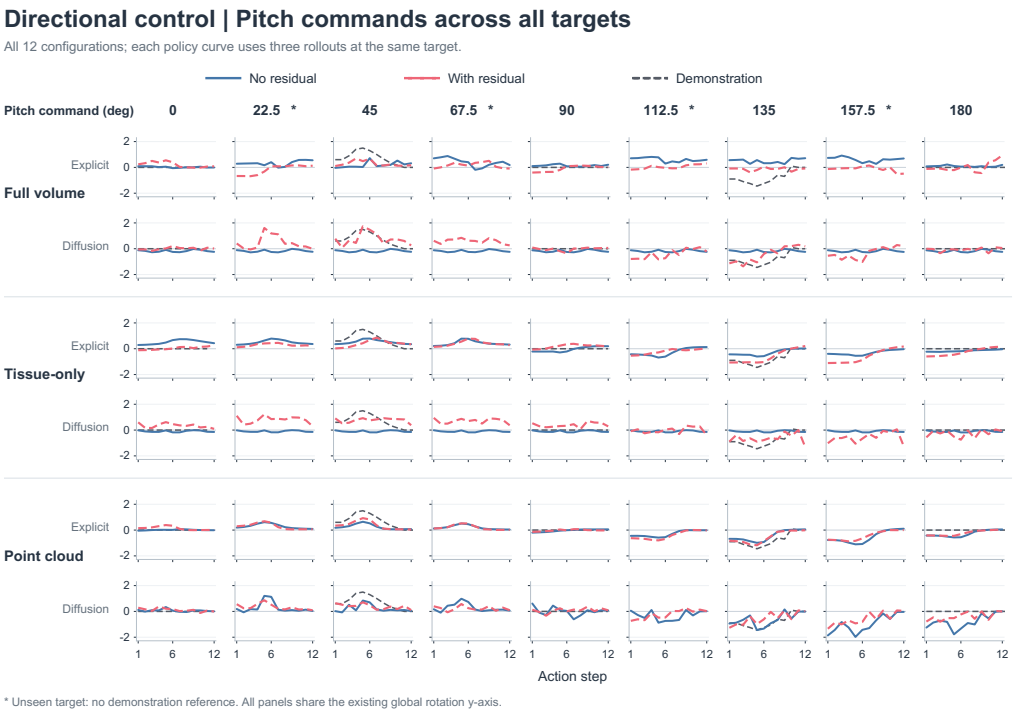


Figure 13: **Directional-task pitch commands across all targets.**

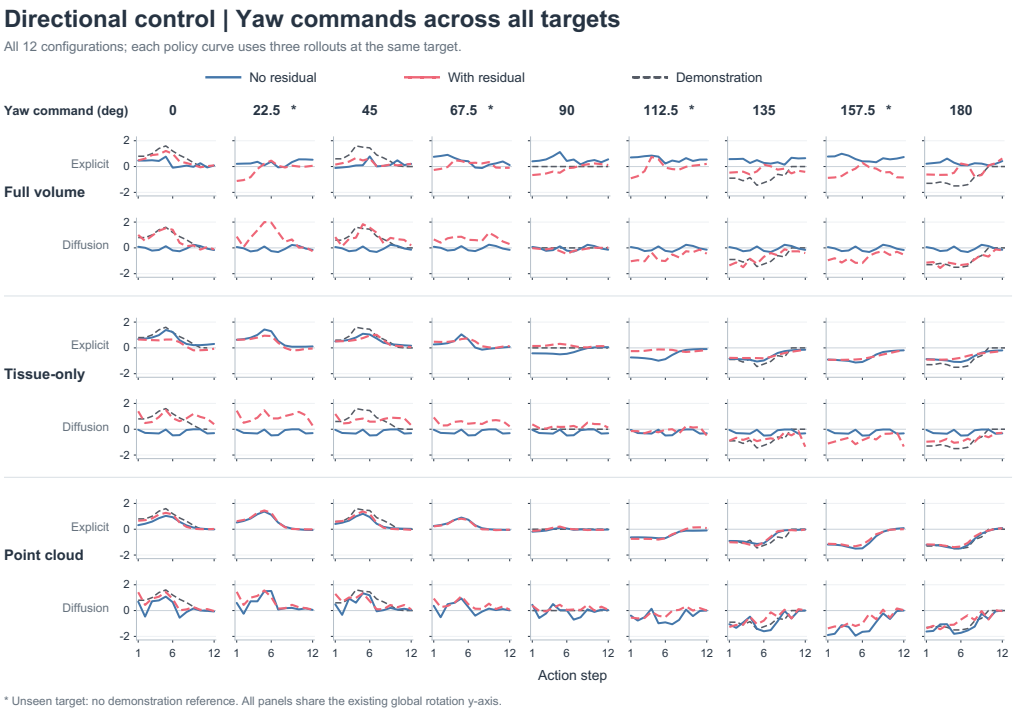


Figure 14: **Directional-task yaw commands across all targets.**

B.3 *Ex vivo* PORCINE RESULTS

We evaluate Full-volume Diffusion and Voxelized point-cloud Explicit, both with residual conditioning. Each configuration is tested at five depth and five directional targets with three deployments per target, yielding 60 twelve-action rollouts. Target conditions are listed in Table 8; geometric summaries follow Eq. (13) with $K = 5$ for each task.

B.3.1 SHAPE AGREEMENT

Endpoint results: Table 13 reports initial and final Shape RMSE and HD95 for both tasks, while Table 14 separates the two distance directions to relate endpoint errors to the observed deformation.

Table 13: Ex vivo endpoint geometry (mm) for 60 rollouts, with 15 per row. Both configurations use residual conditioning. D_0 and D_{12} are means $\pm$ sample SD across three target-averaged sweeps (Eq. (13)); HD95 columns report means.

Configuration	D_0	D_{12}	$H_{95,0}$	$H_{95,12}$	Improved
3-DoF depth					
Full volume / Diffusion	0.650 ± 0.044	1.102 ± 0.045	2.003	4.390	0/15
Point cloud / Explicit	0.592 ± 0.035	1.220 ± 0.088	1.906	4.690	0/15
6-DoF directional					
Full volume / Diffusion	0.985 ± 0.021	0.961 ± 0.024	3.376	2.892	8/15
Point cloud / Explicit	1.022 ± 0.028	1.081 ± 0.038	3.434	2.978	4/15

Two directions of geometric mismatch: We separate the target-to-observation and observation-to-target RMSE terms in Eq. (14), together with the directed q95 terms used in HD95 (Table 14). Target-to-observation RMSE decreases from the initial to final state in 58 of 60 rollouts; observation-to-target RMSE increases in all 60. The two exceptions are Point-cloud depth rollouts at the 2 mm target. For q95, the corresponding decrease and increase occur in all 60 rollouts.

Table 14: Directed components of Shape RMSE and HD95. Each entry shows the initial-to-final mean in millimeters. RMSE refers to the directed terms in Eq. (14); q95 refers to the two percentiles whose maximum defines HD95.

	Target → observation		Observation → target	
Configuration	RMSE	q95	RMSE	q95
3-DoF depth				
Human reference	0.760 → 0.395	1.988 → 0.759	0.486 → 1.544	1.216 → 4.344
Full volume / Diffusion	0.767 → 0.334	2.003 → 0.655	0.499 → 1.520	1.242 → 4.390
Point cloud / Explicit	0.732 → 0.530	1.906 → 0.951	0.394 → 1.639	0.972 → 4.690
6-DoF directional				
Human reference	1.268 → 0.484	3.308 → 0.936	0.523 → 1.169	1.248 → 2.760
Full volume / Diffusion	1.302 → 0.499	3.376 → 0.983	0.488 → 1.259	1.161 → 2.892
Point cloud / Explicit	1.321 → 0.659	3.434 → 1.192	0.579 → 1.376	1.398 → 2.978

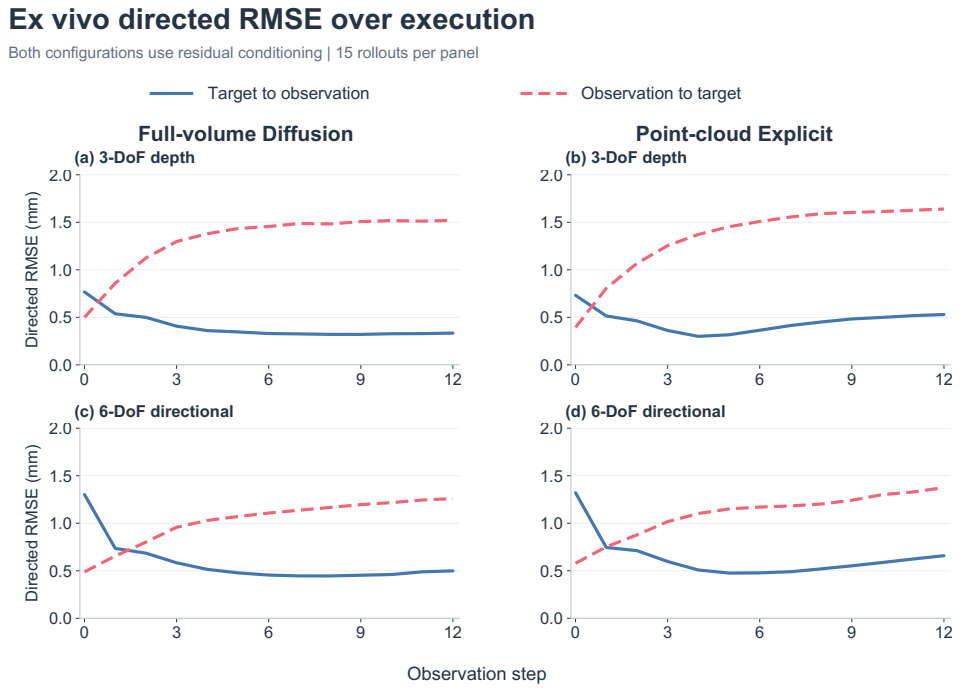


Figure 15: **Directed *ex vivo* RMSE over execution.** Rows show depth and directional tasks, and columns show the two policy configurations. Blue denotes target-to-observation RMSE and coral denotes observation-to-target RMSE.

Figure 15 shows the two RMSE terms over execution. Target-to-observation error falls during the early actions and rises slightly again later, particularly for Point-cloud Explicit. Observation-to-target error generally increases throughout execution.

Relation to tissue deformation. The initial decrease in target-to-observation RMSE is consistent with observed tissue approaching the desired depression. Its later rebound may reflect tissue moving away from some aligned target regions. For example, continued indentation could improve alignment at the center while widening the depression and increasing mismatch along its sidewalls.

Observation-to-target RMSE also includes observed tissue outside the target support. If these regions move farther from the target or occupy a larger fraction of the observed support, their contribution can outweigh the improvement at the center and increase this error. These trends are therefore

Table 15: Target-dependent ex vivo command responses. Statistics use the same definitions as in the phantom action-response analysis. r_d is the target-association score and W is the response spread.

Task	Configuration	Response	r_d	W
3-DoF depth				
Depth	Full volume / Diffusion	Translation	0.854	1.221
Depth	Point cloud / Explicit	Translation	0.660	0.692
6-DoF directional				
Directional	Full volume / Diffusion	Translation	0.926	0.984
Directional	Full volume / Diffusion	Rotation	0.974	2.051
Directional	Point cloud / Explicit	Translation	0.949	0.972
Directional	Point cloud / Explicit	Rotation	0.968	2.051

Table 16: Ex vivo cumulative-command agreement with demonstrations. 'Demo avg." and 'Nearest demo" use the same component-wise command sums and reference distances as Table 12.

	3-DoF translation		6-DoF translation		6-DoF rotation	
Configuration	Demo avg.	Nearest demo	Demo avg.	Nearest demo	Demo avg.	Nearest demo
Full volume / Diffusion	1.148	0.697	0.790	0.852	0.911	0.478
Point cloud / Explicit	0.717	0.586	1.315	0.914	1.526	1.069

consistent with partial alignment toward the intended indentation, while the width, sidewalls, and surrounding tissue remain mismatched.

B.3.2 Action responses

Target dependence: We next evaluate whether the executed commands vary systematically with the requested target, using the same target-association score r_d and response spread W as in the phantom study. Table 15 summarizes all retained ex vivo rollouts.

Agreement with demonstrations: Table 16 reports component-wise cumulative-command agreement using Eq. (22), evaluated at demonstrated depths (2, 3, 4 mm) and directions ($0°$, $45°$, $90°$).

B.3.3 Ex vivo OCT and Color Image examples

Figures 16 and 17 compare the final observed and target surfaces for both configurations in scanner coordinates. RGB insets show the corresponding tool–tissue interactions.

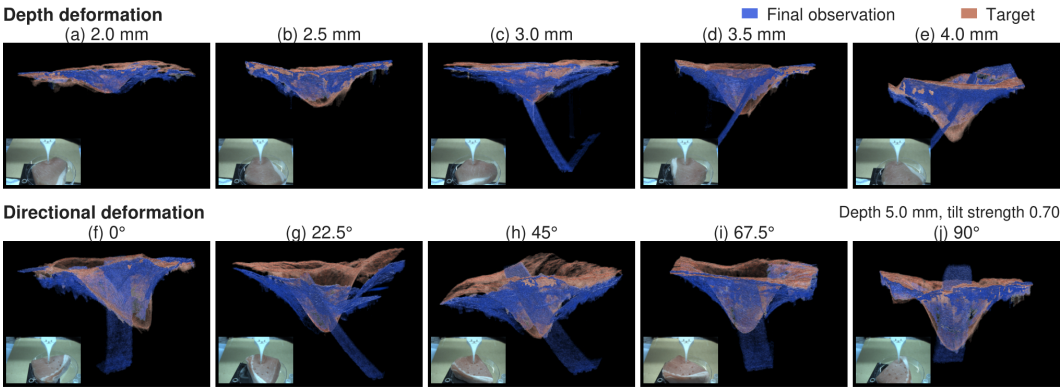


Figure 16: **Representative ex vivo endpoint comparisons for Full-volume Diffusion + residual.** Top row: depth targets 2.0–4.0 mm. Bottom row: directional targets $0°$–$90°$ at depth 5.0 mm and tilt strength 0.70. Blue shows the final observation and orange shows the target.

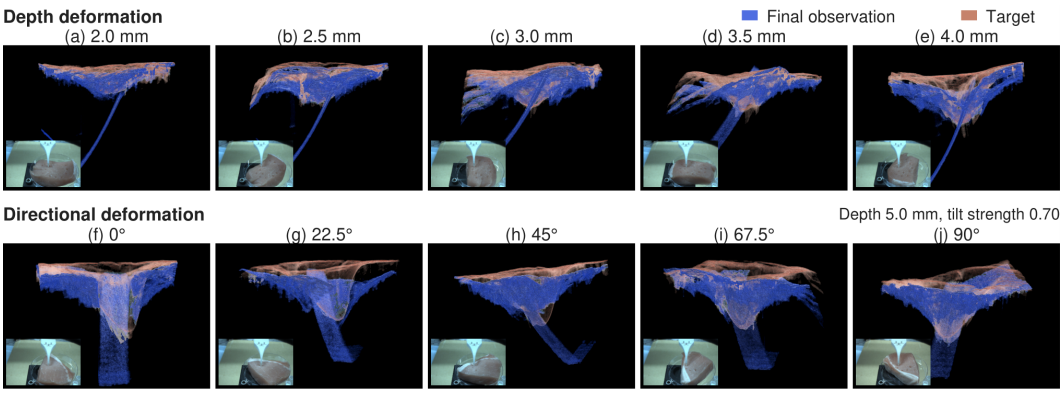


Figure 17: **Representative ex vivo endpoint comparisons for Voxelized point-cloud Explicit with residual conditioning.**

C Operators-Guided Demonstration Data Collection

Two operators independently collected demonstrations through an interactive interface, selecting incremental 6-DoF tooltip motions from the current tissue state and target morphology without prescribed trajectories. Each trajectory records an initial OCT scan and tooltip pose followed by 9–15 actions and the resulting scans and poses. Recorded trajectories are reviewed using tissue–target overlays, retaining demonstrations with feasible interactions and progressive target alignment for training.

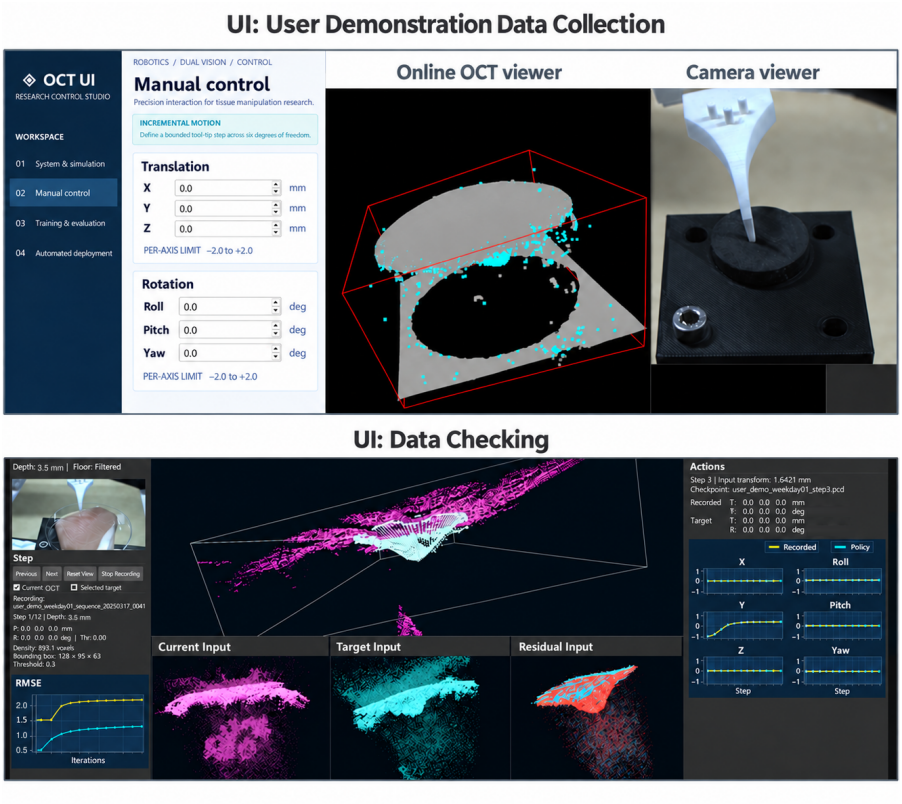


Figure 18: **Demonstration collection and verification.** Top: tooltip control with OCT and RGB feedback. Bottom: trajectory review with tissue–target overlays and geometric error.