

Figure 5: Example VAD-Net Execution

Our VAD-Net method (an example execution of which is shown in Fig. 5) is an ensemble consisting of five independently initialized instances of the DeformerNet architecture. The original DeformerNet architecture is composed of two PointConv-based feature extractors that independently process the current and goal geometries, P_{cm} and P_g , both represented as partial-view point clouds of the deformable object. Each point cloud initially has the shape 1024×3 , corresponding to 1024 3D points. To encode task-relevant context, the current geometry P_c is augmented with the manipulation point m to produce $P_{cm} \in \mathbb{R}^{1024 \times 4}$, where the first three channels encode the spatial coordinates of each point, and the fourth channel is a binary indicator marking the 50 points nearest to m . This augmented input enables the model to focus on regions relevant to the planned manipulation.

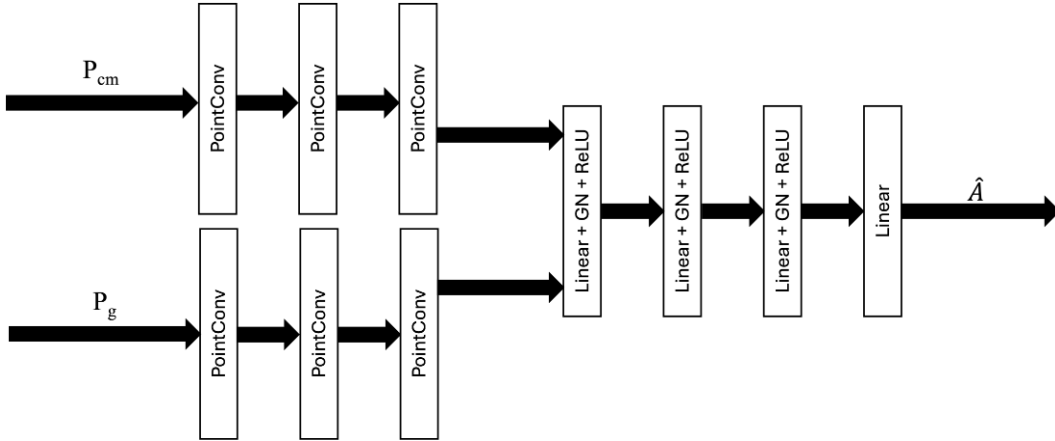


Figure 6: DeformerNet Architecture

In our VAD-Net ensemble, each of the five independently initialized instances of the DeformerNet architecture outputs a predicted 4×4 transformation matrix \hat{A} composed of a rotation matrix $\hat{R} \in \text{SO}(3)$, and a translation vector $\hat{t} \in \mathbb{R}^3$.

Fig. 6 shows a visualization of the full DeformerNet architecture. The feature extractors contain three sequential PointConv layers with increasing feature dimensions of 64, 128, and 256. The encodings of P_{cm} and P_g are then concatenated into a vector of length 512. This vector is then passed through a sequence of four feedforward fully connected layers separated by a GroupNorm

490 layer and a ReLU activation function to produce a 9-dimensional output vector containing $\hat{\mathbf{t}}$ and a
 491 6D vector which gets mapped to a rotation matrix $\hat{\mathbf{R}}$ using the method described in Zhou et al. [61].
 492 Training data is collected entirely in Isaac Gym [57] on a deformable box object. Given a current
 493 geometry of the object P_c and a random manipulation point \mathbf{m} , a random action A is applied to the
 494 robot’s end-effector. The final geometry of the object is recorded as P_g . The model is then trained
 495 on tuples of the form $(P_c, P_g, \mathbf{m}, A)$ where A contains the ground-truth translation \mathbf{t} and rotation
 496 R . The model loss is a linear combination of the mean squared error between the $\hat{\mathbf{t}}$ and \mathbf{t} and the
 497 geodesic distance between $\hat{\mathbf{R}}$ and R .
 498 We train the model using an Adam optimizer with a learning rate of 0.001 over 200 epochs using a
 499 dataset containing 11,566 training and 1,285 test examples. After 100 epochs, the learning rate is
 500 reduced to 0.0001.

Appendix B Additional Handoff Trial Results

To further analyze the performance of our uncertainty-aware handoff framework, we include additional experimental results that provide deeper insight into the relationship between predictive uncertainty and task success, as well as the efficacy of the handoff policies.

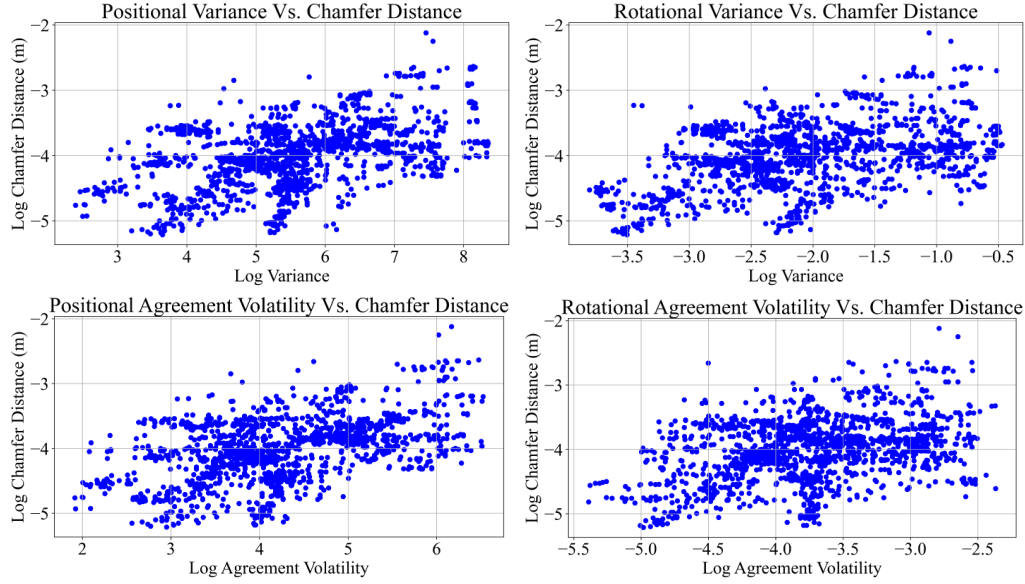


Figure 7: Uncertainty Metric Correlations

Correlation Between Uncertainty Metrics and Task Error Fig. 7 shows log plots of the correlation between each of our uncertainty metrics (ensemble variance and agreement volatility) and the Chamfer distance between the current and goal tissue geometries after taking the predicted action. Across both positional and rotational components, we observe a positive linear relationship. This supports our hypothesis that agreement volatility captures higher-order information about predictive stability and is indicative of downstream task success. These results reinforce the decision to use agreement volatility as an uncertainty feature in the learned handoff policy.

Task Efficiency Across Conditions Fig. 8 compares the relative task efficiency between the Variance Only and VAD-Net policies. For each trial in our 30-trial test set, we record which condition (Variance Only or VAD-Net) completed the task in less total time. We report the percentage of trials in which each method was faster. VAD-Net completed the task faster in approximately 80% of cases (with an average reduction in time of 32.07%), demonstrating its ability to operate more efficiently than the Variance Only baseline. This result highlights the practical advantage of incorporating agreement volatility into the handoff-policy, enabling the system to operate more efficiently while still maintaining high success rates.

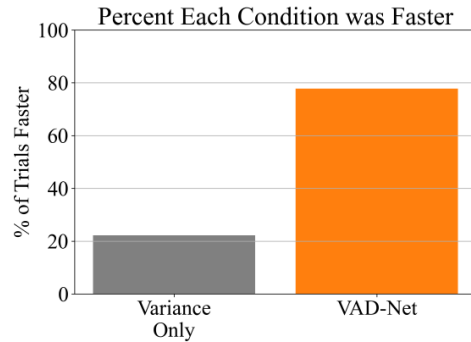


Figure 8: Relative Condition Speed

Table 1 presents a summary of key performance metrics across the three experimental conditions. As expected, the Fully Autonomous baseline

Table 1: Comparison of key performance metrics across control conditions.

Metric	Fully Autonomous	Variance Only	VAD-Net
Success Rate (%)	50	90	100
Avg. Teleoperation Time (s)	–	12.197	4.571
Avg. Autonomous Time (s)	–	16.249	16.566
% Teleoperation	–	36.292	27.170
% Autonomous	–	63.708	72.83
Avg. # of Handoffs	–	2	2.1

530 achieves the lowest task success rate, reinforcing the need for uncertainty-aware handoffs. The Vari-
 531 ance Only policy improves the success rate to 90%, but still requires substantial human intervention
 532 with an average of 12.2 seconds spent in teleoperation mode and 2.0 handoffs per trial. In contrast,
 533 VAD-Net achieves a 100% success rate while reducing average teleoperation time to just 4.6 sec-
 534 onds, indicating greater trust in the policy’s predictions. Despite a similar number of handoffs, the
 535 more efficient decision-making enabled by agreement volatility reduces overall reliance on human
 536 intervention. These results demonstrate that VAD-Net not only enhances reliability in challenging
 537 manipulation tasks, but also enables more autonomous operation without compromising safety.

538 Appendix C Uncertainty Attribution Examples

539 In this section, we present additional insight into VAD-Net’s uncertainty attribution mechanism dur-
 540 ing real-world execution on the dVRK. VAD-Net generates spatial uncertainty maps that highlight
 541 regions in the model input that have the greatest influence on the model’s uncertainty, enabling both
 542 interpretability and collaborative handoffs.

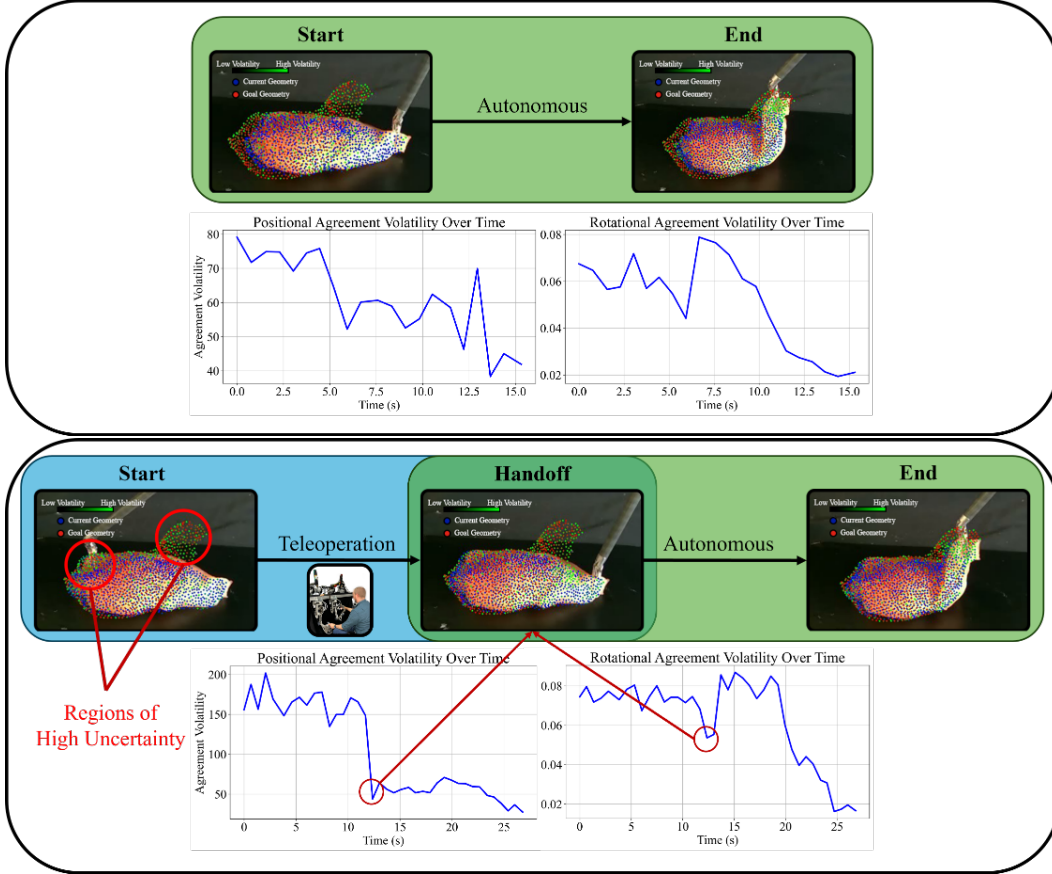


Figure 9: VAD-Net Trial Examples

543 **Case 1: Fully Autonomous Execution** In the first example (top row of Fig. 9), VAD-Net suc-
 544 cessfully operates fully autonomously without the need for human intervention. Success without
 545 the need for human intervention is currently atypical in out-of-distribution scenarios, as mentioned
 546 above and motivating our approach, however when the system is capable, this demonstrative exam-
 547 ple shows our method identifying that no human intervention is needed and autonomous manipula-
 548 tion proceeds successfully. While in this case the spatial uncertainty maps still identify regions with
 549 the highest influence over uncertainty, the agreement volatility plots show a relatively low overall
 550 agreement volatility. This indicates high ensemble confidence, allowing the system to complete the
 551 soft-tissue manipulation task without requesting a human intervention.

552 **Case 2: Intervention and Recovery** In contrast, the second example (bottom row of Fig. 9)
 553 demonstrates an out-of-distribution scenario where the manipulation point is poorly selected, leading
 554 to a high chance of task failure. In this case, the agreement volatility plots show a relatively large
 555 overall initial agreement volatility particularly for the positional component. The spatial uncertainty
 556 map also immediately identifies the high agreement volatility near the manipulation point as well
 557 as the non-overlapping geometry of the goal point cloud. As a result, VAD-Net triggers a human
 558 intervention. After the human corrects the manipulation point, we see a steep decline in both the

559 positional and rotational agreement volatilities, indicating an increase in model confidence. VAD-
560 Net subsequently reclaims control and successfully completes the task autonomously. This example
561 highlights VAD-Net’s ability not only to detect uncertainty in real-time, but also to attribute that
562 uncertainty to regions of the input and reclaim control after human intervention.

563 Together, these cases demonstrate the value of agreement volatility not only as a signal of model
564 uncertainty but also a tool for interpretability, enabling users to understand the source of uncertainty,
565 which is critical for high-stakes applications like surgical robotics.