TReF-6: Inferring Task-Relevant Frames from a Single Demonstration for One-Shot Skill Generalization

Yuxuan Ding **Shuangge Wang** **Tesca Fitzgerald**
Yale University
eason.ding@aya.yale.edu
{shuangge.wang, tesca.fitzgerald}@yale.edu

Abstract: Robots often struggle to generalize from a single demonstration due to the lack of a transferable and interpretable spatial representation. In this work, we introduce **TReF-6**, a method that infers a simplified, abstracted **6**DoF **T**ask-**Re**levant **F**rame from a single trajectory. Our approach identifies an influence point purely from the trajectory geometry to define the origin for a local frame, which serves as a reference for parameterizing a Dynamic Movement Primitive (DMP). This influence point captures the task's spatial structure, extending the standard DMP formulation beyond start-goal imitation. The inferred frame is semantically grounded via a vision-language model and localized in novel scenes by Grounded-SAM, enabling functionally consistent skill generalization. We validate TReF-6 in simulation and demonstrate robustness to trajectory noise. We further deploy an end-to-end pipeline on real-world manipulation tasks, showing that TReF-6 supports one-shot imitation learning that preserves task intent across diverse object configurations.

Keywords: Spatial Reference Frames, One-Shot Imitation Learning, Dynamic Movement Primitives

1 Introduction

Robots are increasingly expected to operate in dynamic, human-centered environments, whether assisting in homes [1], collaborating in warehouses [2], or supporting kitchen workflows [3]. These settings demand generalizable behavior: adapting motion to unseen objects, adjusting to new object placements and orientations, and aligning with different surface orientations. While humans perform such adaptations effortlessly, robots struggle to generalize skills from limited demonstrations.

Since training data can never fully prevent out-of-distribution (OOD) scenarios, researchers increasingly turn to extract structural representations, such as rigid-body poses [4] or keypoints [5, 6, 7], which provide a more stable foundation for generalization in manipulation tasks. Recent work has developed generalizable techniques for generalizing a task in terms of goal poses [8, 9, 10], subgoals [11], or contact interactions [12]. There has been less work, however, on generalizing the spatial constraints encoded in the trajectory itself, such as the curvature of opening the door indicating a hinge constraint besides just the handle position. In practice, the shape of a human demonstration reflects more than just start and end points. It encodes implicit constraints, such as obstacle avoidance [13], mechanical limitations (e.g. hinge constraints) [14], or ergonomic preferences [15]. For any two points in space, infinitely many paths exist, yet demonstrators tend to follow specific, repeatable curves. Prior works in [16] have shown that humans opt for these repeatable trajectories due to their similarities in more condensed, latent geometric structure. We believe that inferring these latent structures could be informative in generalizing trajectories to unseen objective configurations. Furthermore, we expect that they may correspond to semantic features that are detectable

9th Conference on Robot Learning (CoRL 2025), Seoul, Korea.

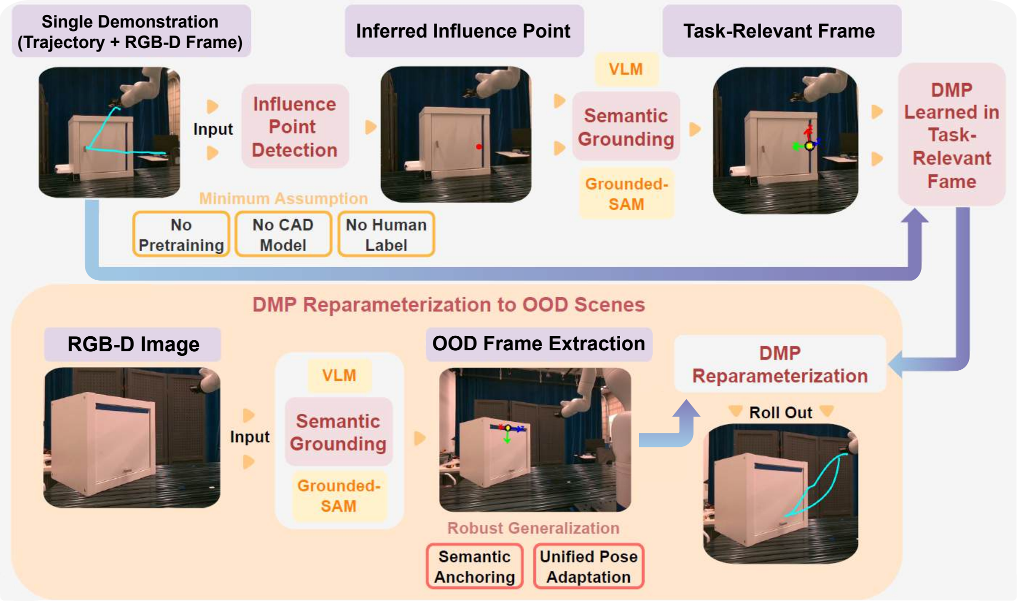


Figure 1: **Overview of TReF-6.** Given a single demonstration, TReF-6 infers an implicit influence point, semantically grounded by a vision-language model (VLM), and extracts a 6-DoF reference frame from the segmentation provided by Grounded-SAM. With minimum assumptions, the inferred frame enables robust OOD generalization.

by vision-language models (VLMs) [17]. This motivates our central research question: *Can we use a single demonstration to infer a task-relevant spatial reference frame that is* **(1)** *semantically identifiable (anchors to scene features)* and **(2)** *functionally meaningful (preserves constraints), enabling generalization across object poses and configurations?*

In this work, we propose **TReF-6** (Figure 1), a framework that infers a task-relevant, 6DoF reference frame from a single demonstration that enables simple dynamical controllers such as DMPs to generalize their motion robustly. Our contributions include:

1. Formalizing the problem of inferring a task-relevant frame as an optimization over its geometric consistency to the trajectory dynamics;
2. An efficient optimization algorithm that is robust to trajectory noise and does not rely on object priors, human labels, or dense annotations; and
3. Simulated and physical robot evaluations that demonstrate our method's ability to generalize to new spatial variations of demonstrated tasks.

2 Related Work

Dynamic Movement Primitives (DMPs) encode a demonstrated trajectory as a stable dynamical system [18, 19] that converges toward a goal. DMPs can generalize the shape of a demonstrated motion to target a new goal pose, yet are functionally agnostic to the surrounding environment unless manually augmented with additional force-like terms that account for external influences [20] or goal and sub-goal related task frames [21]. Without access to an appropriate task-specific reference, DMPs are trained over just a single trajectory, so by definition they "overfit" to the environmental configuration and constraints that were present during the initial demonstration.

Data-driven approaches such as TP-GMMs [22] and their extensions [23] involve conditioning a demonstrated trajectory on external task parameters. These approaches improve adaptation across different object poses, placements, and orientations, but require extensive training data (including synthetic data) to generalize effectively [22, 24], or assume that relevant task-relevant frames, such as those aligned with object or interaction axes, are provided or can be easily extracted [25].

Affordance learning focuses on the actionable possibilities in an environment, such as regions on an object that can be grasped, pushed, or opened [26], capturing the physical interactions that enable task completion. Recent efforts in this field have focused on visually identifying object surfaces that support affordances based on robot-collected data [27] or human demonstration videos [28, 29]. Others use depth data to address occlusion and collision constraints in articulated objects [30]. While these approaches have shown promising results, they involve training visual backbones with target object types or multi-view RGB inputs. In real-world scenarios where a robot must interact with unseen novel objects, especially under varying camera viewpoints with crowded objects and random occlusions, such reliance becomes a bottleneck in terms of data requirements and system robustness.

One-shot imitation learning targets this challenge of adapting to novel task constraints or objects. This may involve grounding the demonstration using external context—language descriptions or videos of related tasks to make sense of the action in a broader setting [31]. Other methods focus on low-level structure, aligning object parts or motion trajectories across different scenes to find transferable patterns [32, 8, 33] or with rich contact information [12]. Recent work also looks for structure that is consistent across different environments, for example, regions in 3D space that consistently guide behavior across tasks [34], or prior demonstrations retrieved based on graph-based similarity metrics [35]. These approaches enable a robot to quickly generalize its task knowledge to novel constraints and objects, and reinforce the importance of a reasoning module that correctly and efficiently identifies task-relevant features for one-shot learning [36].

Although efficient at generalization, one-shot methods still require a large amount of data at training to acquire the ability to learn-to-learn and generalize from a single new example. Compared to end-to-end one-shot policy learning approaches, DMPs offer advantages in stability, data efficiency, and ease of adaptation once a transferable frame is available. This motivates our focus on building a minimal-assumption method to automatically extract a task-relevant spatial frame, enabling classical DMPs to generalize motion trajectories without the need for large-scale task distributions, external labels, or object CAD models.

3 Methodology

We introduce **TReF-6**[1], a trajectory-based framework for inferring a task-relevant local frame from a single demonstration for skill generalization. Unlike prior work that relies on large-scale training data or assumes external knowledge about object geometry, our method leverages a novel optimization-based formulation to infer latent influence point from motion dynamics, which is general-purpose and operates on a single trajectory. The end-to-end pipeline consists of three stages:

1. **Influence Point Inference:** Optimize a directional consistency score to identify a 3D spatial point that best captures the trajectory's dynamics.
2. **Semantic Grounding:** Refine the point by aligning it with a semantically relevant and spatially aligned visual feature identified by a VLM, and extract a local frame based on surface normals and interaction direction.
3. **DMP Reparameterization:** Transform the trajectory into the inferred local frame, fit DMPs, and reuse them in novel scenes by extracting a new frame.

3.1 Influence Point Inference: Defining a Directional Consistency Score

We formulate the frame inference problem as identifying a coordinate transformation informed by the trajectory. We hypothesize that the demonstrated motion is shaped by a dominant spatial constraint like rotation around a hinge (axis constraint), movement along a shelf (plane constraint), or alignment toward a socket (point constraint). Rather than designing separate models for each type,

[1] GitHub repository: https://github.com/iqr-lab/tref-6

we infer a single 6-DoF frame centered at a latent point $p \in \mathbb{R}^3$ that reflects the underlying structure influencing the motion, serving as a general-purpose reference for motion generalization.

Given a single trajectory $\{x_t\}_{t=1}^{T} \in \mathbb{R}^3$, we hypothesize that the motion can be captured by a position-only, latent influence pointing from x_t to a fixed point $p \in \mathbb{R}^3$. We optimize p for both *temporal consistency* and *directional agreement*, whether the observed acceleration aligns consistently toward a candidate point, inspired by prior work in shared-control robotics to disambiguate user intent [37]. Specifically, we define the **directional consistency score**, $\mathcal{S} : \mathbb{R}^3 \to \mathbb{R}$, which compares the predicted direction from x_t to p with the acceleration, $\ddot{x}_t$. We define $\mathcal{S}(p)$ as:

$$\mathcal{S}(p) = -\frac{1}{T}\sum_{t=1}^{T} \left|\left| \frac{p - x_t}{||p - x_t|| + \epsilon} - \ddot{x}_t \right|\right| \tag{1}$$

where $\epsilon > 0$ is a small positive constant. A higher $\mathcal{S}(p)$ value indicates stronger consistency between the trajectory's underlying dynamics and the candidate point p. A 2D example of the score landscape is shown in Appendix Figure 9, where regions around the ground-truth candidate point of influence have higher $\mathcal{S}(p)$ values. By relying on the second derivative, the score inherently reflects the temporal dynamics of the trajectory. Since direction difference is normalized, $\mathcal{S}$ provides an objective that is robust to variations in force magnitude.

3.2 Influence Point Inference: Optimizing for Directional Consistency

We can then estimate the latent intent point by solving for $p^* = \arg\max_{p \in \mathbb{R}^3} \mathcal{S}(p)$. Due to vector normalization and temporal error aggregation, we find that the score landscape exhibits many local optima and extensive flat regions, especially under noise in acceleration direction, as illustrated in Appendix Figure 5, making the result highly sensitive to initialization. Empirically, we find that choosing an initialization point with (1) *sharp gradients* and (2) *proximity to the ground truth*, is essential for successful optimization - an insight aligned with prior works in optimization [38, 39]. Formally, if we denote the small angular deviation between $\frac{p-x_t}{||p-x_t||}$ as θ, we can apply the law of cosines to approximate the squared residual and its partial derivative as follows:

$$\left|\left| \frac{p - x_t}{||p - x_t|| + \epsilon} - \ddot{x}_t \right|\right|^2 \approx ||\ddot{x}_t||^2 - 2||\ddot{x}_t|| \cos\theta + 1 \tag{2}$$

$$\implies \frac{\partial}{\partial\theta} \left|\left| \frac{p - x_t}{||p - x_t|| + \epsilon} - \ddot{x}_t \right|\right|^2 \approx \frac{\partial}{\partial\theta}\left(||\ddot{x}_t||^2 - 2||\ddot{x}_t|| \cos\theta + 1\right) = 2||\ddot{x}_t|| \sin\theta \tag{3}$$

where the change in residual magnitude due to the deviations in θ is proportional to $||\ddot{x}_t||$. We initialize the optimization near these regions by computing $p_0 \sim \frac{1}{k}\sum_{\tau \in \mathcal{I}} x_j + \mathcal{N}(0, \sigma^2 \mathbf{I})$, where $\mathcal{I}$ is the top$-k$ timesteps with the largest $||\ddot{x}_t||$ magnitudes. This initialization places the optimization near points of sharper gradients. We empirically show in Section 4.1 that this approach significantly improves convergence and solution quality. Due to the non-smoothness of normalization near $p \approx x_t$, we approximate gradients using central finite differences and use Adam [40] to smooth fluctuations in the gradient and accelerate convergence. Algorithm 1 details the entire optimization algorithm.

3.3 Semantic Grounding

We then use the optimized influence point p^* as the origin of the task-relevant 6DoF Frame. To ensure the frame is semantically meaningful and transferable across scenes, we refine this point by aligning it with visual features identified by GPT-4o [41]. Specifically, we implement a two-phase querying process. In the first phase, we prompt the model to generate a high-level task label based on the initial state RGB image overlaid with the demonstration trajectory. In the second phase, using the predicted task label, we query GPT-4o again with the same RGB image now overlaid with the inferred influence point to identify the visual features associated with both the influence point and the interaction point (where the robot first begins its interact with the environment). We provide the full prompts in Appendix 9.8. The model returns a natural language description of the features, which we use to guide segmentation via Grounded-SAM and refine the point. Once the refined point

Algorithm 1 Trajectory-based Influence Point Identification

```
Require: Trajectory positions $\{x_t, \ddot{x}_t\}_{t=1}^{T}$, learning rate $\eta > 0$, total steps $N$, initialization count $k$
$\mathcal{I} = \arg\max_{\substack{\mathbb{J}\subseteq\{1,\dots,n\} \\ |J|=k}} \sum_{j\in\mathbb{J}} ||\ddot{x}_t||$        ▷ Select top-$k$ time steps with highest norms
Sample $p \sim \frac{1}{k}\sum_{\tau\in\mathcal{I}} x_j + \mathcal{N}(0, \sigma^2\mathbf{I})$        ▷ Sample Initial State
for $i = 1$ to $N$ do
    $\nabla_p \mathcal{S}(p) \approx \left[\frac{\mathcal{S}(p+\epsilon e_i) - \mathcal{S}(p-\epsilon e_i)}{2\epsilon}\right]_{i=1}^{3}$        ▷ Estimate Gradient
    $p \leftarrow p - \eta \nabla_p \mathcal{S}(p)$
end for
return $p$
```

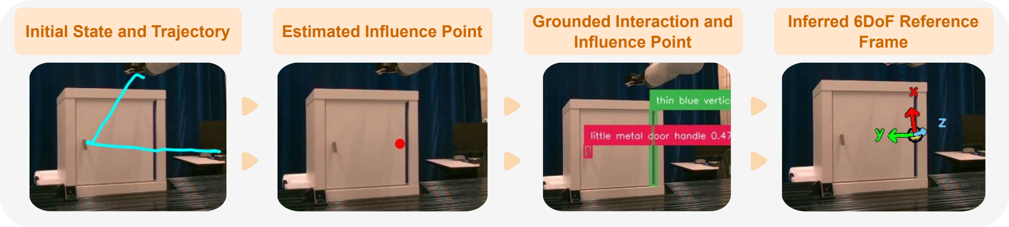

Initial State and Trajectory
Estimated Influence Point
Grounded Interaction and Influence Point
Inferred 6DoF Reference Frame
thin blue vertic
little metal door handle 0.47
x
y
z

Figure 2: 6DoF Frame Extraction for Door Opening Demonstration.

is established, we compute the surface normal at that location to define the z-axis of the frame. We then define the yz-plane using the z-axis and the vector from the refined point to the interaction point. The x-axis is then computed as the unit vector orthogonal to this plane. Finally, the y-axis is obtained as the cross product of the z- and x-axes. This captures both local geometry (via the surface normal) and task-relevant directionality (via the interaction point), as shown in Figure 2.

3.4 DMP Reparameterization

To enable environment-adaptive generalization, we apply the standard DMP formulation in a task-relevant frame centered at the refined influence point p^*. Rather than modifying the DMP dynamics, we reparameterize the demonstrated trajectory by transforming it from the robot base frame, referred to as the world frame, into the task-relevant frame. Given a demonstration of length T, the Cartesian position x_t and quaternion orientation q_t at time t of the trajectory are transformed from the world frame into the inferred task-relevant frame. We then compute the relative motion from the starting pose in this local frame:

$$\Delta x_t = x_t^{\text{local}} - x_0^{\text{local}} \qquad\qquad \Delta q_t = q_t^{\text{local}} \otimes {q_0^{\text{local}}}^{-1} \tag{4}$$

where x_0, q_0 denote the initial end-effector pose relative to the influence point, and $\otimes$ is quaternion multiplication. This initial pose corresponds to the location of the tool at the beginning of task execution (e.g., where a brush first makes contact with the robot). We then fit Cartesian and quaternion DMPs over Δx_t and Δq_t, respectively. At deployment, a new influence point p^* and associated task-relevant frame are inferred, and the new starting pose $(x_0^{\text{new}}, q_0^{\text{new}})$ is computed accordingly. The DMPs then roll out relative motions with respect to $(x_0^{\text{new}}, q_0^{\text{new}})$. The definition of the roll out functions $\text{DMP}_{\text{pos}}(x_0^{\text{new}})$ and $\text{DMP}_{\text{quat}}(q_0^{\text{new}})$ is deferred to Appendix 9.6. The generated trajectory is finally mapped back into the world frame using the inverse transformation.

4 Experiment

To evaluate whether our proposed method could infer a task-relevant spatial reference frame that is (1) semantically identifiable and (2) functionally meaningful from a single demonstration, we designed controlled 3D simulations with known influence point to assess inference precision under varying levels of directional and magnitude noise in acceleration, using Mean Euclidean Distance

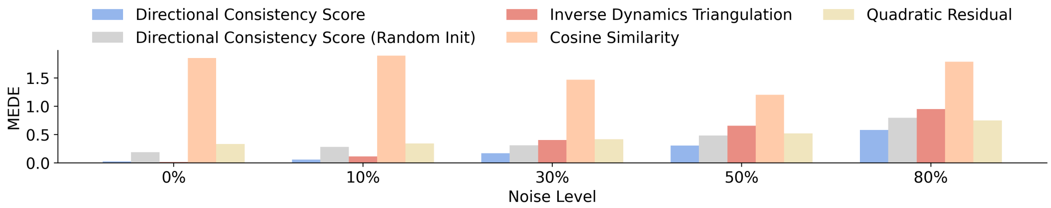


Figure 3: Mean Euclidean Distance Error (MEDE) comparison of spatial influence inference methods under varying levels of noise (0% to 80%). Complete results are summarized in Appendix 9.4

Error (MEDE) between the inferred and ground-truth influence point as the metric. In real-world tasks, we assess whether the inferred frame can be semantically grounded and enable one-shot skill generation. While most sophisticated baselines such as affordance-based imitation or goal-conditioned policies exist, they typically rely on object CAD models [8], extensive training [34], or rich contact information [12], which are unavailable in our setting. We thus benchmark against privileged DMPs, which have access to additional information such as object positions or operate in environments that mimic the demonstration setup, defined in Section 4.2, as the strongest feasible baseline. Improvement over the baseline, measured by task success under an OOD environment, indicates that the inferred frame captures structure critical for generalizing the demonstrated behavior.

4.1 Simulated Environment

In real-world settings, demonstrations are inherently noisy, and our method relies on the second-order trajectory, which amplifies noise. Therefore, to test the robustness of our method with noisy demonstrations, we design a 3D simulation where the ground-truth influence point and motion dynamics are fully controllable. A point mass, simulating a robot end-effector, starts at the origin with a random velocity sampled from $[-0.5, 0.5]^3$, and is influenced by a randomly placed, ground-truth influence point $\hat{p}$ within a bounded region $[-5, 5]^3$. A directionally noisy force points from the particle to the influence, scaled by a coefficient $\alpha_t \sim \mathcal{U}(0, ||\hat{p} - x_t||)$, mimicking diminishing attraction. We add a constant Gaussian noise to both direction and magnitude of control. The state evolves throughout a locally linearized dynamics for 100 steps.

We evaluated our method with and without random initialization along with three other scoring objectives and one inverse dynamics baseline for inferring the spatial influence point p. All methods were tested under the same simulated conditions. We reported the MEDE between the predicted and ground-truth influence points over 50 randomized seeds. These baselines were selected to represent diverse but plausible strategies for trajectory-based influence point inference without requiring access to ground-truth labels, including a physically grounded method (inverse dynamics triangulation), optimization over residual scores (quadratic residual score), and directional alignment (cosine similarity score).

Directional Consistency Score (ours) achieved the best overall performance across all noise levels, maintaining low error and variance even as the force signal became increasingly corrupted, consistently outperforming random initialization baseline across all noise settings, with an average of 55.0% reduction in error and an average of 86.8% reduction in variance. Appendix Figure 7 illustrates a representative worst-case outcome under 50% noise using random initialization. In contrast, across all baselines, we observed distinct failure modes under extreme noisy conditions. **Inverse Dynamics Triangulation** was highly sensitive to perturbations in the direction of the force vector, as it relied on accurate ray intersections from $\ddot{x}_t$ orientations. **Cosine Similarity Score** suffered in 3D settings due to vanishing gradients, leading to poor convergence and high variance. **Quadratic Residual Score** broke down when the force magnitudes were irregular or noisy—as was the case in our simulations where magnitudes were modulated stochastically. In contrast, our proposed **Directional Consistency Score** remained robust by comparing normalized force predictions directly with observed accelerations, achieving the best performance across varying noise levels.

Figure 3 summarizes these results. Mathematical formulations of Cosine Similarity Score, Quadratic Residual Score, and Inverse Dynamics Triangulation are summarized in Appendix 9.2. More detailed analysis and discussion of each approach as well as a comprehensive simulation experiment results are summarized in Appendix 9.4.

4.2 Real-World Experiment

Our real-world experiments aim to validate that **TReF-6** enables functionally meaningful one-shot skill generalization, even when paired with simple downstream controllers such as DMPs. We evaluate its effectiveness across three real-world manipulation tasks on a 7-DoF Kinova Gen3 robot where trajectory shape and alignment are critical: (1) *peg-in-hole dropping*, (2) *cabinet door opening*, and (3) *surface wiping*. Each task includes one demonstration and multiple OOD variants for evaluation. The peg-in-hole task tests for semantic transferability and precision, with variations in object shape and color, as well as rod color and height. The door-opening task examines generalization across spatial positioning and rotations, varying cabinet position, hinge placement, and cabinet orientation. The wiping task evaluates adaptability to surface tilt and the ability to maintain continuous surface contact without excessive force, with changes in stain appearance and the flatness of the board.

To isolate the contribution of our method, we adapt DMPs as the shared motion controller and compare executions with and without our inferred local reference frame. Since vanilla DMPs lack semantic grounding, we provide a privileged setup for baseline DMPs: objects and rods in the peg-in-hole task are placed in the original demonstration locations; for door opening, the handle position is explicitly specified; and for wiping, the whiteboard brush and tilted surface are arranged to match the demonstration.

(1) Peg-in-hole Dropping: In this task, the robot is provided with a single demonstration, illustrated in Appendix Figure 6, and both methods successfully reproduce the demonstrated behavior in the original setting. We intentionally designed the task to be challenging and tightly constrained: as illustrated in Appendix Figure 11, slight misalignments between the object and the rod lead to failure. Our method maintains a high success rate under variations, as shown in Figure 4. While the baseline achieves a similar success rate in the pickup phase due to its privileged setup, its performance in completing the full task, assuming the object has been picked up, is significantly lower than ours.

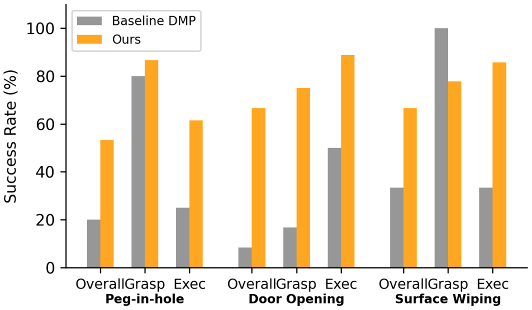


Figure 4: Real world experiment results. Each task includes: Overall success, Grasp success, and Execution success given successful grasp. Complete results were summarized in Appendix 9.5

(2) Cabinet Door Opening: The robot is provided with a single demonstration, illustrated in Appendix Figure 6, where it opens a cabinet door directly in front of it, left to right. Successful task execution requires rotating the handle around a hinge axis - a simple linear pulling would result in failure, as it would drag the entire cabinet without opening the door. Both the baseline and our method reproduce the demonstrated behavior in the original setting. However, DMP fails to adapt its motion to the new hinge geometry or cabinet orientation. In contrast, our method generalizes the motion successfully across multiple variants with an overall success rate of 66.7% as shown in Figure 4.

(3) Surface Wiping: In this task, the robot is provided with a single demonstration, illustrated in Appendix Figure 6. A successful execution requires wiping out the ink on the whiteboard while maintaining contact with the surface, without deforming it to the point of touching the threshold placed behind. Both the baseline and our method reproduce the demonstrated behavior under the original setup. As shown in Appendix Table 6, the baseline performs reliably in this setting and under stain color changes. However, its motion does not adapt to different surface orientations.

In contrast, our method achieves a much higher overall success rate of 66.7% across tilted surface variants. When evaluating only trails where the brush was successfully grasped, our method achieves 85.7% success, outperforming the baseline by 52.4%.

5 Discussion and Analysis

Overall, our real-world evaluations show that **TReF-6** enables functionally meaningful generalization across spatial variations with only a single demonstration by providing a semantically transferable task-relevant frame. We now highlight the key takeaways of our results.

Task-Specific Strength. Across all three tasks, **TReF-6** exhibits consistent improvements over baseline DMPs by leveraging task-specific spatial cues. In the *peg-in-hole dropping* task, it adapts to changes in rod height and color, successfully adjusting the hook trajectory, while baseline DMP fail to hook the rod securely before release. In the *door opening* task, **TReF-6** correctly infers the arc direction and hinge orientation, unlike baseline which follows a fixed left-to-right arc. In the *surface wiping* task, **TReF-6** realigns motion to maintain contact to different tilted surfaces, which the baseline cannot handle.

Partial Success Despite Failures. Even when task execution is not fully successful, **TReF-6** often produces structurally meaningful motions. Failures in peg-in-hole dropping tasks arise primarily from color misclassifications (e.g., identifying red or blue rods as a purple rod), yet the behavior remains robust enough to complete partial goals (e.g. reaching the wrong rod). Most failures in door opening task arise from unreachable grasp pose, particularly when the handle is located near the table surface, leading the robot to hit kinematic limits. When grasping fails, the resulting motion still preserves arc structure around the hinge, reflecting meaningful generalization without success in the full task.

Performance Correlates to Perception Accuracy. The quality of generalization is tightly coupled with the reliability of depth perception. One example frame extraction failure case of dropping on the green rod that caused by depth noise is visualized in Appendix Figure 12. Failure in the mirrored variant of door opening task is also attributed to poor depth perception when the door surface is nearly orthogonal to the camera, as shown in Appendix Figure 13. The lowest performance on surface wiping occurs at the 30° tilt, as shown in Appendix Figure 10. We hypothesize that shallow tilts, oblique angles, and small cross-section area of object lead to noisier or incomplete depth maps, leading to frame extraction failure. However, these perception challenges, such as segmentation errors or depth noise, may be mitigated in the future through advances in 3D perception and improvements to grounding models like Grounded-SAM, especially considering that we did not fine-tune the model for our tasks.

6 Conclusion

We presented a novel framework to augment DMPs' generalization capabilities: inferring a task-relevant spatial reference frame from a single demonstration that is both semantically identifiable and functionally meaningful. Unlike prior approaches that rely on object priors, multiple demonstrations, or predefined frames, our method extracts a full 6-DoF reference frame directly from the geometry of the demonstrated trajectory. By anchoring this inferred frame to semantic scene features, we enable skill generalization across diverse object poses and spatial configurations. Our physical robot experiments validate that our approach preserves the structural constraints of the task while adapting to challenging environment variations. These results highlight the promise of geometry-driven, semantically grounded reference frames as a foundation for scalable imitation learning. Moreover, our framework is agnostic to the type of downstream motion primitives; future works therefore could explore beyond DMPs such as probabilistic movement primitives and kernel-based models. Future works can also extend our approach with richer object representations to enable grasping strategies across diverse geometries.

7 Limitations

Since **TReF-6** relies on Grounded-SAM for localization, semantic mis-segmentation occasionally occurs. Since segmentation is not fine-tuned and operates out-of-the-box, addressing such external semantic ambiguities remains future work.

TReF-6 extracts a single influence point to define a task-relevant 6DoF frame. While more complex spatial constraints could theoretically require richer representations, our experiments demonstrate that a single frame abstraction is sufficient to generalize across diverse atomic tasks in practice.

TReF-6 focuses on motion generation after an object has been acquired, modeling the tool or object as a single representative point. It does not address grasp planning or complex contact interactions, and thus assumes that a feasible grasp has already been achieved prior to motion execution. Despite this simplification, experiments demonstrate that **TReF-6** achieved near-perfect success rates in task execution once the object is properly grasped. Future work will focus on extending the framework to integrate grasp planning, enabling the system to jointly reason about how to grasp and how to perform the task within the same spatial frame.

8 Acknowledgments

The authors would like to thank members of the Yale Inquisitive Robotics Lab and Qian Wang for their feedback and contributions.

References

[1] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, M. Y. Galliker, D. Ghosh, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, D. LeBlanc, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, A. Z. Ren, L. X. Shi, L. Smith, J. T. Springenberg, K. Stachowicz, J. Tanner, Q. Vuong, H. Walke, A. Walling, H. Wang, L. Yu, and U. Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization, 2025. URL https://arxiv.org/abs/2504.16054.

[2] N. Kleer, M. Rekrut, J. Wolter, T. Schwartz, and M. Feld. A multimodal teach-in approach to the pick-and-place problem in human-robot collaboration. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '23, page 81–85, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450399708. doi:10.1145/3568294.3580047. URL https://doi.org/10.1145/3568294.3580047.

[3] H. Shi, H. Xu, S. Clarke, Y. Li, and J. Wu. Robocook: Long-horizon elasto-plastic object manipulation with diverse tools. *arXiv preprint arXiv:2306.14447*, 2023.

[4] B. Wen, W. Yang, J. Kautz, and S. Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects, 2024. URL https://arxiv.org/abs/2312.08344.

[5] L. Manuelli, W. Gao, P. Florence, and R. Tedrake. kpam: Keypoint affordances for category-level robotic manipulation, 2019. URL https://arxiv.org/abs/1903.06684.

[6] F. Liu, K. Fang, P. Abbeel, and S. Levine. Moka: Open-world robotic manipulation through mark-based visual prompting, 2024. URL https://arxiv.org/abs/2403.03174.

[7] W. Huang, C. Wang, Y. Li, R. Zhang, and L. Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation, 2024. URL https://arxiv.org/abs/2409.01652.

[8] T. Li, S. Sun, S. S. Aditya, and N. Figueroa. Elastic motion policy: An adaptive dynamical system for robust and efficient one-shot imitation learning, 2025. URL https://arxiv.org/abs/2503.08029.

[9] C. Pan, B. Okorn, H. Zhang, B. Eisner, and D. Held. TAX-pose: Task-specific cross-pose estimation for robot manipulation. In *6th Annual Conference on Robot Learning*, 2022. URL https://openreview.net/forum?id=YmJi0bTfeNX.

[10] A. Simeonov, Y. Du, A. Tagliasacchi, J. B. Tenenbaum, A. Rodriguez, P. Agrawal, and V. Sitzmann. Neural descriptor fields: Se(3)-equivariant object representations for manipulation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6394–6400, 2022. doi:10.1109/ICRA46639.2022.9812146.

[11] C. Ochoa, H. Oh, Y. Kwon, Y. Domae, and T. Matsubara. Ispil: Interactive sub-goal-planning imitation learning for long-horizon tasks with diverse goals. *IEEE Access*, 12:197616–197631, 2024. doi:10.1109/ACCESS.2024.3521302.

[12] H. Chang, A. Boularias, and S. Jain. Insert-one: One-shot robust visual-force servoing for novel object insertion with 6-dof tracking. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2935–2942, 2024. doi:10.1109/IROS58592.2024.10801884.

[13] A. Winn, X. Gao, S. Mishra, and A. A. Julius. Learning potential functions by demonstration for path planning. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 4654–4659, 2012. doi:10.1109/CDC.2012.6426153.

[14] G. Subramani, M. Zinn, and M. Gleicher. Inferring geometric constraints in human demonstrations. In A. Billard, A. Dragan, J. Peters, and J. Morimoto, editors, *Proceedings of The 2nd Conference on Robot Learning*, volume 87 of *Proceedings of Machine Learning Research*, pages 223–236. PMLR, 29–31 Oct 2018. URL `https://proceedings.mlr.press/v87/subramani18a.html`.

[15] A. Bestick, R. Pandya, R. Bajcsy, and A. D. Dragan. Learning human ergonomic preferences for handovers. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3257–3264, 2018. doi:10.1109/ICRA.2018.8461216.

[16] J. Jin, L. Petrich, Z. Zhang, M. Dehghan, and M. Jagersand. Visual geometric skill inference by watching human demonstration. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, page 8985–8991. IEEE, May 2020. doi:10.1109/icra40945.2020.9196570. URL `http://dx.doi.org/10.1109/ICRA40945.2020.9196570`.

[17] C. Zhang and G. H. Lee. Iaao: Interactive affordance learning for articulated objects in 3d environments, 2025. URL `https://arxiv.org/abs/2504.06827`.

[18] A. Ijspeert, J. Nakanishi, and S. Schaal. Movement imitation with nonlinear dynamical systems in humanoid robots. In *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No.02CH37292)*, volume 2, pages 1398–1403 vol.2, 2002. doi:10.1109/ROBOT.2002.1014739.

[19] A. J. Ijspeert, J. Nakanishi, H. Hoffmann, P. Pastor, and S. Schaal. Dynamical movement primitives: Learning attractor models for motor behaviors. *Neural Computation*, 25(2):328–373, 2013. doi:10.1162/NECO_a_00393.

[20] P. Pastor, H. Hoffmann, T. Asfour, and S. Schaal. Learning and generalization of motor skills by learning from demonstration. In *2009 IEEE International Conference on Robotics and Automation*, pages 763–768, 2009. doi:10.1109/ROBOT.2009.5152385.

[21] L. Koutras and Z. Doulgeri. A novel dmp formulation for global and frame independent spatial scaling in the task space. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 727–732, 2020. doi:10.1109/RO-MAN47096.2020.9223500.

[22] S. Calinon. A tutorial on task-parameterized movement learning and retrieval. *Intelligent Service Robotics*, 9(1):1–29, January 2016. ISSN 1861-2776. doi:10.1007/s11370-015-0187-9.

[23] Y. Huang, J. Silvério, L. Rozo, and D. G. Caldwell. Generalized task-parameterized skill learning, 2018. URL `https://arxiv.org/abs/1707.01696`.

[24] J. Zhu, M. Gienger, and J. Kober. Learning task-parameterized skills from few demonstrations, 2022. URL `https://arxiv.org/abs/2201.09975`.

[25] S. Hu and K. J. Kuchenbecker. Hierarchical task-parameterized learning from demonstration for collaborative object movement. *Applied Bionics and Biomechanics*, 2019:9765383, 2019. doi:10.1155/2019/9765383.

[26] J. J. Gibson. *The Ecological Approach to Visual Perception: Classic Edition*. Psychology Press, 1st edition, 2014. doi:10.4324/9781315740218. URL `https://doi.org/10.4324/9781315740218`.

[27] L. Yen-Chen, P. Florence, A. Zeng, J. T. Barron, Y. Du, W.-C. Ma, A. Simeonov, A. R. Garcia, and P. Isola. Mira: Mental imagery for robotic affordances, 2022. URL `https://arxiv.org/abs/2212.06088`.

[28] S. Bahl, R. Mendonca, L. Chen, U. Jain, and D. Pathak. Affordances from human videos as a versatile representation for robotics, 2023. URL `https://arxiv.org/abs/2304.08488`.

[29] J. Chen, D. Gao, K. Q. Lin, and M. Z. Shou. Affordance grounding from demonstration video to target image, 2023. URL https://arxiv.org/abs/2303.14644.

[30] K. Cheng, R. Wu, Y. Shen, C. Ning, G. Zhan, and H. Dong. Learning environment-aware affordance for 3d articulated object manipulation under occlusions, 2023. URL https://arxiv.org/abs/2309.07510.

[31] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In A. Faust, D. Hsu, and G. Neumann, editors, *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 991–1002. PMLR, 08–11 Nov 2022. URL https://proceedings.mlr.press/v164/jang22a.html.

[32] D. Hadjivelichkov, S. Zwane, M. Deisenroth, L. Agapito, and D. Kanoulas. One-Shot Transfer of Affordance Regions? AffCorrs! In K. Liu, D. Kulic, and J. Ichnowski, editors, *Proceedings of The 6th Conference on Robot Learning (CoRL)*, volume 205 of *Proceedings of Machine Learning Research*, pages 550–560, 14–18 Dec 2023.

[33] P. Vitiello, K. Dreczkowski, and E. Johns. One-shot imitation learning: A pose estimation perspective, 2023. URL https://arxiv.org/abs/2310.12077.

[34] X. Zhang and A. Boularias. One-shot imitation learning with invariance matching for robotic manipulation. In *Proceedings of the Robotics: Science and Systems (RSS)*, 07 2024. doi: 10.15607/RSS.2024.XX.134.

[35] Z.-H. Yin and P. Abbeel. Offline imitation learning through graph search and retrieval. *Robotics: Science and Systems*, 2024.

[36] A. Ajay, S. Han, Y. Du, S. Li, A. Gupta, T. Jaakkola, J. Tenenbaum, L. Kaelbling, A. Srivastava, and P. Agrawal. Compositional foundation models for hierarchical planning, 2023. URL https://arxiv.org/abs/2309.08587.

[37] D. E. Gopinath and B. D. Argall. Active intent disambiguation for shared control robots. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 28(6):1497–1506, 2020. doi: 10.1109/TNSRE.2020.2987878.

[38] E. J. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, Apr. 2015. ISSN 1557-9654. doi:10.1109/tit.2015.2399924. URL http://dx.doi.org/10.1109/TIT.2015.2399924.

[39] Y. Chi, Y. M. Lu, and Y. Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, Oct. 2019. ISSN 1941-0476. doi:10.1109/tsp.2019.2937282. URL http://dx.doi.org/10.1109/TSP.2019.2937282.

[40] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[41] OpenAI. Hello gpt-4o, 2024. URL https://openai.com/index/hello-gpt-4o/.

9 Appendix

9.1 Loss Landscape

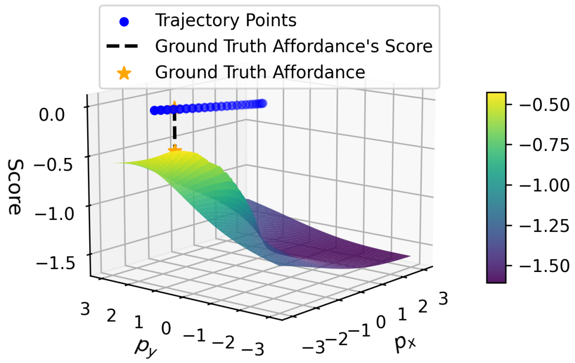


Figure 5: Score Landscape in 2D case for a trajectory length of $T = 25$. Notice the large flat gradient areas around regions far from the trajectory.

9.2 Definition of Baselines

A. Cosine Similarity Score

We define the **cosine similarity score** as:

$$\text{Score}_{\text{cos}}(p) = \frac{1}{T}\sum_{t=1}^{T}\left|\frac{(p - x_t)^\top \ddot{x}_t}{||(p - x_t)|| \cdot ||\ddot{x}_t|| + \epsilon}\right| \tag{5}$$

While still **non-convex**, cosine similarity score is effective in 2D due to a constrained directional space and sharp gradient alignment. Its performance degrades in 3D settings where directional ambiguity, vanishing gradients, and the lack of magnitude information significantly impair its ability to infer correct influence points.

B. Quadratic Residual Score

We define the **quadratic residual score** as:

$$\text{Score}_{\text{quad}}(p) = \frac{1}{T}\sum_{t=1}^{T}||(p - x_t) - \ddot{x}_t||^2 \tag{6}$$

Despite this objective being **convex** and easy to optimize, it assumes the magnitude of the observed acceleration matches that of the position-based prediction, which is often violated in realistic demonstrations and results in low accuracy even in simulated environment.

C. Inverse Dynamics Triangulation

Under the Newtonian assumption $\vec{f}_t = m \cdot \vec{a}_t$, the **inverse dynamics triangulation** method interpret the direction of the acceleration vector at each timestep as a ray pointing toward the latent affordance point. Then, the triangulation process estimates the point in space that minimizes the orthogonal distance to all such rays:

$$\min_{p \in \mathbb{R}^3} \sum_{t=1}^{T} \left|(I - \hat{a}_t \hat{a}_t^\top)(x_t - p)\right|^2 \tag{7}$$

where $\hat{a}_t$ is the unit acceleration direction at timestep t, and $(I - \hat{a}_t \hat{a}_t^\top)$ is the projection matrix orthogonal to that direction.

While effective under clean conditions, this method is highly sensitive to noise in the direction of $\vec{a}_t$, especially in higher dimensions where ray intersections are less geometrically constrained. As such, it performs well at low noise levels but deteriorates rapidly when acceleration signals are noisy or inconsistent.

9.3 Demonstrations for Each Task

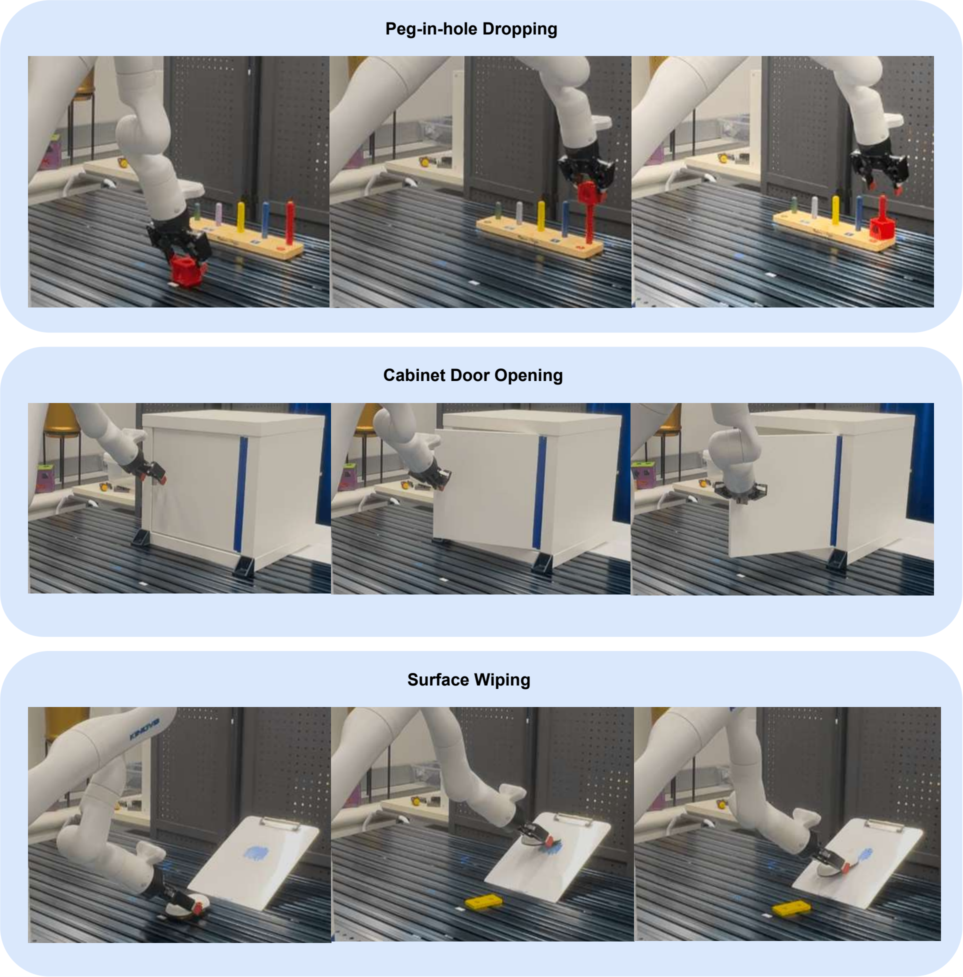


Figure 6: **Single Demonstration per Task.** Each row shows the provided demonstration for one of the three tasks: peg-in-hole dropping (top), cabinet door opening (middle), and surface wiping (bottom). These single demonstrations are the only inputs used for learning; our method infers task-relevant spatial reference frames from each to enable downstream generalization to novel object configurations and orientations.

9.4 Simulation Experiment Details

9.4.1 3D Baseline Analysis

Directional Consistency Score (ours) achieved the best overall performance across all noise levels tested. Even as the force signal became increasingly corrupted, it maintained both low error and low variance: MEDE raises modestly from 0.0560 ± 0.0276 at 10% noise to 0.5814 ± 0.3722 at 80%

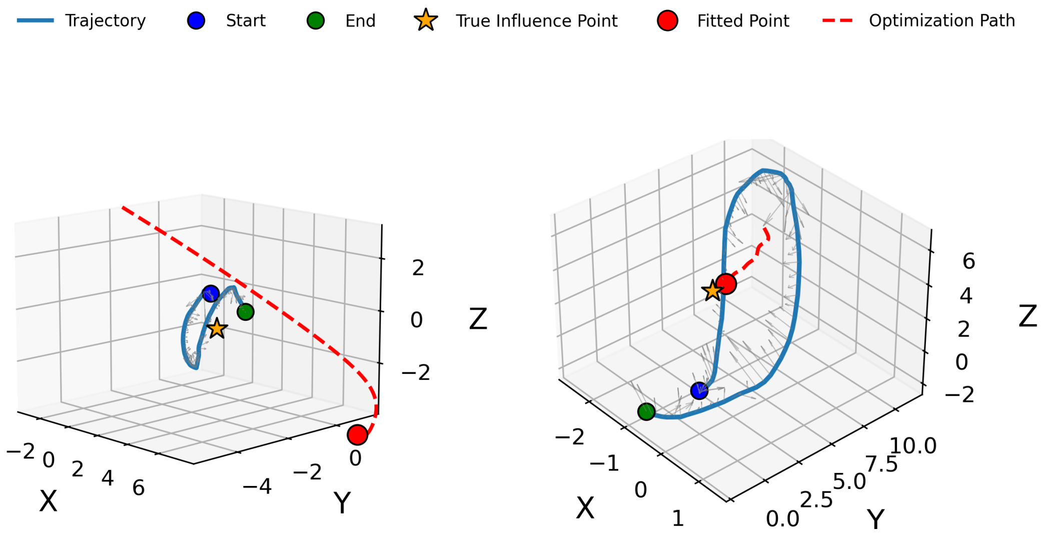


Figure 7: **Qualitative comparison between random and structured initialization under 50% noise.** Gray arrows represent $\ddot{x}$, which appear noisy due to 50% noise level, and the dashed red line illustrates the optimization trajectory. (Left) Random initialization is trapped in a flat-gradient region. (Right) Structured initialization converges close to the true influence despite noisy acceleration.

Table 1: Comparison of Spatial influence Inference Methods under Different Noise Levels over 50 Random Seeds in 3D. Report in MEDE: Mean ± Std

Method	**0% Noise**	**10% Noise**	**30% Noise**	**50% Noise**	**80% Noise**
Directional Consistency Score	0.0263 ± 0.0392	$\mathbf{0.0560 \pm 0.0276}$	$\mathbf{0.1688 \pm 0.0831}$	$\mathbf{0.3044 \pm 0.1808}$	$\mathbf{0.5814 \pm 0.3722}$
Directional Consistency Score (Random Initialization)	0.1884 ± 0.8217	0.2794 ± 1.0735	0.3076 ± 0.6416	0.4828 ± 1.1278	0.7945 ± 1.2573
Inverse Dynamics Triangulation	$\mathbf{0.0127 \pm 0.0157}$	0.1111 ± 0.0818	0.4009 ± 0.2589	0.6540 ± 0.4020	0.9474 ± 0.5689
Cosine Similarity	1.8483 ± 2.2455	1.8892 ± 2.3067	1.4667 ± 2.0082	1.1981 ± 1.7585	1.7815 ± 2.0691
Quadratic Residual	0.3311 ± 0.2217	0.3397 ± 0.2225	0.4147 ± 0.2441	0.5191 ± 0.2868	0.7482 ± 0.4037

noise. Despite its non-convexity, our method yielded stable and accurate estimates when paired with structured initialization and optimization.

Inverse Dynamics Triangulation performed well at low noise levels: 0.1111 ± 0.0818 at 10% and achieved the best performance over other methods at 0% noise level with MEDE 0.0127 ± 0.0157. However, it degraded more rapidly as the noise increases, reaching 0.9747 ± 0.5689 at 80% noise. We anticipated this behavior since this approach assumes clean Newtonian force signals and becomes unstable when acceleration vectors fluctuate or intersect imprecisely.

Cosine Similarity Score aligned force directions but disregarded magnitude. In 3D, where the directional ambiguity is high and the gradients flatten near $\cos(\theta) \approx 1$, the method struggled to guide the optimization effectively. Interestingly, increasing the level of noise improved its performance, and it reached its best performance at the noise level 50% with the MEDE of 1.1981 ± 1.7585. The cause would be the injected noise that pushes the optimization out of local minima. It produced the highest variance overall and failed catastrophically at times, especially under low noise.

Quadratic Residual Score offered a convex alternative, minimizing the L2 difference between $p - x_i$ and a_t. However, this formulation implicitly assumes that the observed acceleration vectors have magnitudes that are consistent with the inferred direction vectors. In our simulation, we intentionally violated this assumption by scaling the force magnitude as the distance $||p - x_t||$ multiplied by a stochastic coefficient drawn from a uniform distribution. This design mimicked real-world

Table 2: Comparison of Spatial influence Inference Methods under Different Noise Levels over 50 Random Seeds in 2D. Report in MEDE: Mean ± Std

Method	**0% Noise**	**10% Noise**	**30% Noise**	**50% Noise**	**80% Noise**
Directional Consistency Score	0.0391 ± 0.0289	$\mathbf{0.0584} \pm \mathbf{0.0344}$	$\mathbf{0.1473} \pm 0.1903$	$\mathbf{0.2174} \pm \mathbf{0.1725}$	0.4965 ± 1.0663
Inverse Dynamics Triangulation	$\mathbf{0.0128} \pm \mathbf{0.0169}$	0.0739 ± 0.0642	0.2817 ± 0.2396	0.4356 ± 0.3194	0.6179 ± 0.4023
Cosine Similarity	1.7646 ± 2.0060	1.5219 ± 2.0972	1.4158 ± 2.2923	1.6448 ± 2.2317	1.5106 ± 1.8028
Quadratic Residual	0.2275 ± 0.1601	0.2432 ± 0.1475	$0.2753 \pm \mathbf{0.1612}$	0.3218 ± 0.1919	$\mathbf{0.4468} \pm \mathbf{0.2454}$

conditions, where human-applied forces are irregular and not strictly distance-proportional. As a result, the quadratic score performed poorly, with a MEDE of 0.3311 ± 0.2217 even in the noise-free setting.

The results were shown in Appendix Table 1.

9.4.2 Lower-Dimensional Intuition

To better understand the behaviors of different scoring functions, we also conducted a controlled 2D version of the influence inference experiment with setup mirrored the 3D scenario. We evaluated the same four methods (Directional Consistency Score, Inverse Dynamics Triangulation, Cosine Similarity, and Quadratic Residual) across 50 seeds under different directional noise. The results were summarized in Appendix Table 2.

In the 2D setting, both the **Directional Consistency Score (ours)** and **Inverse Dynamics Triagnulation** demonstrated strong performance under low to moderate noise. Notably, the **Directional Consistency Score** achieved the most robust accuracy across 10%, 30%, and 50% noise levels, which confirmed its resilience to force perturbations even in lower-dimensional dynamics. The increased standard deviation of the **Directional Consistency Score** at 80% noise could be attributed to degraded initialization under extreme noise, where the top-k acceleration magnitudes no longer reliably reflected actual force directions. As a result, optimization was more likely to begin near outliers or flat-gradient regions, leading to wider variability in outcomes.

Inverse Dynamics Triangulation again achieved the lowest MEDE of 0.0128 ± 0.0169 in the noise-free setting. Although its performance degraded with increasing noise, the deterioration was more moderate than in the 3D case due to reduced directional ambiguity.

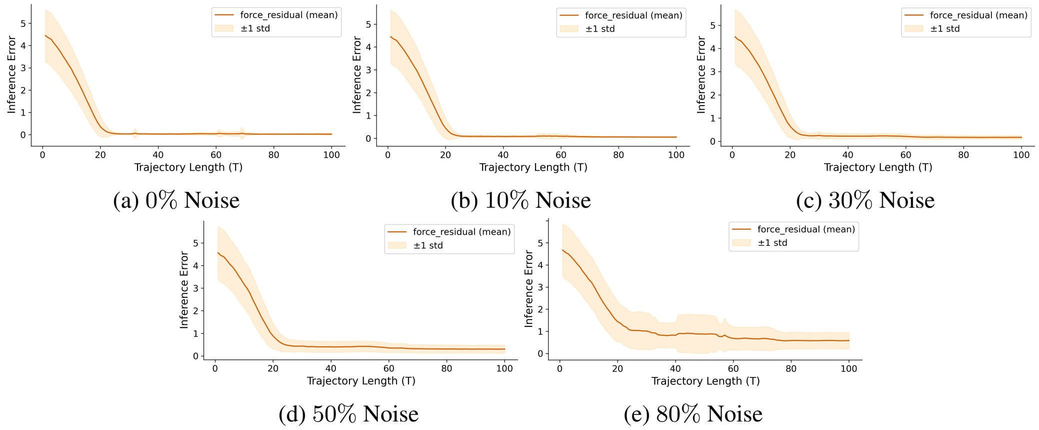
Quadratic Residual Score, as expected, remained stable due to its convex formulation. While it outperformed other methods at noise level 80% with a MEDE of 0.4468 ± 0.2454, this performance was not significantly better than that of the **Directional Consistency Score**, which achieved a comparable MEDE of 0.4965 ± 1.0663.

9.4.3 Inference from Partial Trajectories

To assess how early the spatial intent can be reliably inferred, we evaluated our method under the constraint of partial observation. Specifically, we provided only the first T timesteps of the demonstration trajectory $\{x_t, \ddot{x}_t\}_{t=1}^{T}$ to the inference algorithm, and vary T from 1 to 100. This setting emulated early recognition in human intentions when interacting with robots where only partial trajectory is observed.

Appendix Figure 8 plotted the mean inference error (±1 std) of the fitted point across 50 random seeds under varying trajectory lengths, for five different noise levels: 0%, 10%, 30%, 50%, and 80%.

Across all noise levels, we observed a significant inference error drop within the first $20 - 30$ timesteps and gradually plateaus afterward. At low noise (0%, 10%), as few as 25 steps sufficed to reliably identify the influence point. Under heavier noise (50%, 80%), longer observation windows were required, but the Directional Consistency Score still converged to reasonable estimates under 40 steps.



(a) 0% Noise

(b) 10% Noise

(c) 30% Noise

(d) 50% Noise

(e) 80% Noise

Figure 8: **Effect of Partial Trajectory on Inference Accuracy over 50 random seeds.** Mean inference error (± 1 std) of the Directional Consistency Score evaluated using the first k timesteps ($\{x_t, a_t\}_{t=1}^{k}$), with $k \in [1, 100]$. Our method achieves accurate inference using less than $1/4$ of the complete trajectory under moderate noise, and remains robust even under severe corruption.

As expected, the variance was significantly higher when fewer steps are available, even under a noise-free setting, showing the ambiguity in intent when the motion was just beginning. As more of the trajectory was revealed, the system became increasingly accurate, suggesting that the method captured the cumulative structure of intention encoded in second-order motion.

9.4.4 Sequential Influence Inference

We further evaluated our method in a setting where a trajectory was governed by two spatial influence points p_1 and p_2 in sequence. The transition between the two points was defined by a known switch step, allowing the inference algorithm to segment the trajectory accordingly. For each segment, we fitted a separate influence point using our method and reported the MEDE for each inferred point.

Appendix Table 3 presented the results across three noise levels (0%, 10%, and 30%) and three switch ratios: $30/70$, $50/50$, and $70/30$ corresponding to the percentage of trajectory length influenced by p_1 vs. p_2. Each configuration was averaged over 50 random seeds.

Across all settings, we observed a longer influence period of a point generally yields a lower MEDE. This result aligned with intuition that longer segments provide more consistent motion patterns to constrain the optimization.

Interestingly, we found that $30/70$ **split consistently achieved the lowest overall mean error** across noise levels. This behavior could be explained by the inherent asymmetry in how motion dynamics evolved across segments. In our simulation, the first segment began with an initial velocity uniformly sampled from $[-0.5, 0.5]^3$, representing a moderate and unbiased starting momentum. In contrast, the second segment inherited the velocity from the first, which may accumulate significant directional bias by the time of switching. Consequently, estimating p_2 required more trajectory context.

In real-world applications such as manipulation or teleoperation, this challenge may be less significant. For instance, a human operator may pause or reorient the robot between sub-tasks, effectively resetting the velocity and making both phases more distinguishable. Our simulation represented a worst-case continuity scenario and that influence inference in practice could be even more robust given appropriately segmented demonstrations.

Table 3: Performance of Sequential influence Inference under Varying Noise Levels and Switch Ratios. Reported in MEDE (Mean Euclidean Distance Error) over 50 seeds.

Noise Level	Switch Ratio (p1 / p2)	MEDE (p1)	MEDE (p2)	Overall Mean Error
0%	30 / 70	**0.0551 ± 0.0404**	**0.0541 ± 0.0696**	**0.0546 ± 0.0359**
	50 / 50	0.0622 ± 0.0886	0.0800 ± 0.1255	0.0711 ± 0.0760
	70 / 30	0.0720 ± 0.0654	0.4450 ± 0.5543	0.2585 ± 0.2733
10%	30 / 70	0.0936 ± 0.0424	**0.0949 ± 0.0574**	**0.0943 ± 0.0362**
	50 / 50	0.1131 ± 0.1572	0.1337 ± 0.1504	0.1234 ± 0.1041
	70 / 30	**0.0888 ± 0.0484**	0.4822 ± 0.4958	0.2855 ± 0.2459
30%	30 / 70	0.2360 ± 0.1219	**0.2583 ± 0.1402**	**0.2471 ± 0.1069**
	50 / 50	0.2378 ± 0.1643	0.2799 ± 0.2037	0.2588 ± 0.1331
	70 / 30	**0.1955 ± 0.1029**	0.6322 ± 0.4831	0.4138 ± 0.2442

9.5 Real World Experiment Complete Results

Table 4: Success Rates of Baseline DMP vs. Our Method for Peg-in-hole Dropping Task. Reported over 15 variations.

Metric	Baseline DMP	Ours (Gap)
Overall Success Rate	20.0%	**53.3%** (**+33.3%**)
Grasping Success Rate	80.0%	**86.7%** (+**6.7**%)
Execution Success (Given Grasp)	25.0%	**61.5%** (**+36.5%**)
Success Rate by Object:		
Red Cube	1/5	**4/5** (+**3**)
Blue Cap	1/5	**2/5** (+**1**)
Green Ring	1/5	**2/5** (+**1**)
Success Rate by Rod:		
Red Rod	**3/3**	**3/3** (+0)
Blue Rod	0/3	**1/3** (+**1**)
Yellow Rod	0/3	**3/3** (+**3**)
Purple Rod	0/3	0/3 (+0)
Green Rod	0/3	**1/3** (+**1**)

Table 5: Success Rates of Baseline DMP vs. Our Method for Cabinet Door Opening Task. Reported over 12 variations.

Metric	Baseline DMP	Ours (Gap)
Overall Success Rate	8.3%	**66.7%** (**+58.3%**)
Grasping Success Rate	16.7%	**75.0%** (**+58.3%**)
Execution Success (Given Grasp)	50.0%	**88.9%** (**+38.9%**)
Success Rate by Orientation:		
Frontal	1/4	**3/4** (+**2**)
Left Angled	0/4	**3/4** (+**3**)
Right Angled	0/4	**2/4** (+**2**)
Success Rate by Hinge Position:		
Hinge at Right	1/3	**3/3** (+**2**)
Hinge at Left	0/3	**2/3** (+**2**)
Hinge at Bottom	0/3	**3/3** (+**3**)
Hinge at Top	0/3	0/3 (+0)

Table 6: Success Rates of Baseline DMP vs. Our Method for Surface Wiping. Reported over 9 variations.

Metric	**Baseline DMP**	**Ours (Gap)**
Overall Success Rate	33.3%	**66.7%** (**+33.3%**)
Grasping Success Rate	**100.0%**	77.8% (−22.2%)
Execution Success (Given Grasp)	33.3%	**85.7%** (**+52.4%**)
Success Rate by Stain Color:		
Blue	1/3	**2/3** (+1)
Black	1/3	**2/3** (+1)
Red	1/3	**2/3** (+1)
Success Rate by Surface Tilted Angle:		
45°	**3/3**	2/3 (−1)
90°	0/3	**3/3** (**+3**)
30°	0/3	**1/3** (+1)

9.6 DMP Roll Out Definition

Position DMP:

$$\mathrm{DMP}_{\mathrm{pos}}(x_0^{\mathrm{new}}) = \left\{ y_t \;\middle|\; \begin{array}{l} \tau \ddot{y}_t = \alpha_z \left(\beta_z(g^{\mathrm{deploy}} - y_t) - \tau \dot{y}_t\right) + f(s_t), \\ \tau \dot{y}_t = \ddot{y}_t, \quad \tau \dot{s}_t = -\alpha_s s_t, \\ f(s) = \dfrac{\sum_i \psi_i(s) w_i}{\sum_i \psi_i(s)} s\, g^{\mathrm{deploy}}, \quad g^{\mathrm{deploy}} = \Delta x_T \cdot \dfrac{\|p^* - x_0^{\mathrm{new}}\|}{\|p^* - x_0^{\mathrm{demo}}\|} \end{array} \right\}_{t=1}^{T} \tag{8}$$

Orientation DMP:

$$\mathrm{DMP}_{\mathrm{quat}}(q_0^{\mathrm{new}}) = \left\{ y_t \;\middle|\; \begin{array}{l} \tau \ddot{y}_t = \alpha_z \left(\beta_z(\Delta q_T - y_t) - \tau \dot{y}_t\right) + f(s_t), \\ \tau \dot{y}_t = \ddot{y}_t, \quad \tau \dot{s}_t = -\alpha_s s_t, \\ f(s) = \dfrac{\sum_i \psi_i(s) w_i}{\sum_i \psi_i(s)} s\, \Delta q_T \end{array} \right\}_{t=1}^{T} \tag{9}$$

9.7 Score Visualization

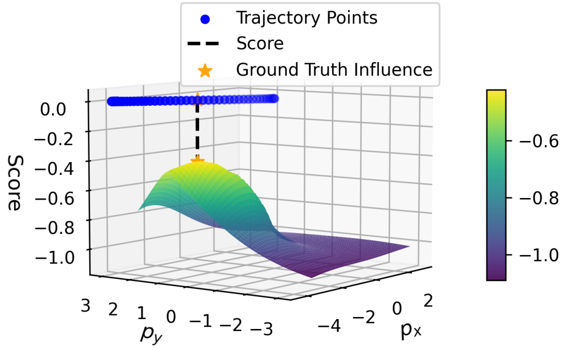


Figure 9: Score Landscape in 2D case for a trajectory length of $T = 50$. Notice the high score region around the ground-truth influence point.

9.8 Prompt to VLM

After we obtained the inferred influence point, it was projected on the initial RGB start state image. To extract semantic information about the task and the relevant environmental features, we adopt a two-phase querying strategy using a Vision-Language Model (VLM).

In the first phase, given the full demonstration trajectory overlaid on the start-state image, the VLM is queried to infer a concise task label that captures the primary objective of the motion. In the second phase, given the projected influence point on the same image and the previously inferred task label, the VLM is queried to identify the fine-grained environmental feature corresponding to the influence point and suggest an appropriate robot interaction location.

Few text-based examples are given to restrict the output to fine-grained details, without image-text pairs or in-context visual examples. The VLM is used solely as a tool for semantic extraction of task-relevant features, without any task-specific retraining or fine-tuning.

For the experiments conducted in this work, we employ GPT-4o [41], one of the latest publicly available Vision-Language Models at the time of the experiments. Due to the rapid advancement of vision-language models, the proposed pipeline is designed to directly benefits from future models with improved visual reasoning capabilities.

We provide the full text of the prompts used for querying the VLM in each phase below.

Phase 1: Task Label Inference

> **Instructions:** You are a motion understanding expert. Provided is an image showing the initial state of a robot task, overlaid with the robot's full demonstration trajectory.
>
> Based on the scene context and the projected trajectory, identify the task that the robot is attempting to perform.
>
> Respond with a concise task label that captures the primary objective of the robot's action. Focus solely on the relevant interaction. Ignore unrelated background objects.
>
> Examples of valid task labels include: "place object on shelf," "connect two components,"" align tool with fixture," "adjust object position," or "slide object along surface."
>
> Your output should be a short phrase describing the task.

Phase 2: Fine-Grained Feature Identification

> **Instructions:** You are an environment reasoning expert. Provided is an image of the environment annotated with a projected spatial influence point, inferred from the robot's demonstration.
>
> The task is {task label}. Your goal is to reason about fine-grained, task-relevant environmental features based on the task and influence point. In particular:
>
> - Detect precise structural features that are critical for completing the task.
> - Avoid vague descriptions like "cabinet" or "box." Instead, refer to specific parts, boundaries, or interaction affordances.
>
> Based on the influence point and the task:
>
> 1. Describe the fine-grained environmental feature that the projected point corresponds to that is related to the task. Be specific about the geometry, material, or function if relevant.
> 2. Specify where the robot should grasp or interact with the object to successfully complete the task, using fine-grained features as references. Be specific about the geometry, material, or function if relevant.

Do not mention unrelated scene elements.

Provide two short phrases: one for the influence point description and one for the grasp location.

Example Outputs

- Task label: *Open cabinet door*
- Fine-grained description:
  - (1) The projected point corresponds to the vertical seam between the cabinet door and the frame, near the edge where the door can be pulled open.
  - (2) The robot should grasp the small metal door handle located along the seam to successfully open the cabinet door.

9.9 Real World Task Details and Failure Example Visualizations

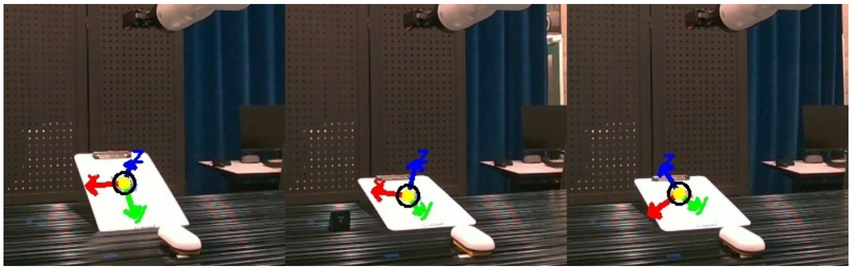
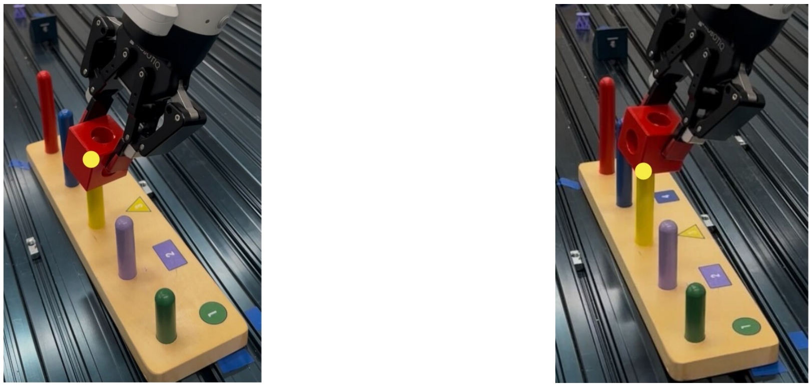


Figure 10: **Comparison of Extracted Local Frames in Surface Wiping Variants.** (Left) Local frame extracted from the demonstration, correctly aligning the z-axis with the surface normal of a tilted board. (Middle) Frame inferred for a successful execution on a flatter tilt, preserving the surface orientation and enabling stable wiping contact. (Right) Frame inferred during a failed trial, where the z-axis does not follow the true surface normal, resulting in a wiping trajectory that drifts off the board or fails to maintain contact.



(a) **Ours (DMP + Frame)** successfully hooks the red cube over the shorter yellow rod by adapting the trajectory downward before release.

(b) **Baseline DMP** fails to adapt to the rod height. Despite visual alignment, the cube misses the hook due to an insufficient downward motion.

Figure 11: **Task Sensitivity to Rod Height Variation.** The yellow dot indicates the top of the yellow rod. A small height change leads to failure if the hook motion is not adapted. Our method infers a spatial reference frame that enables height-aware motion adjustment, while baseline DMP fails despite close visual alignment.

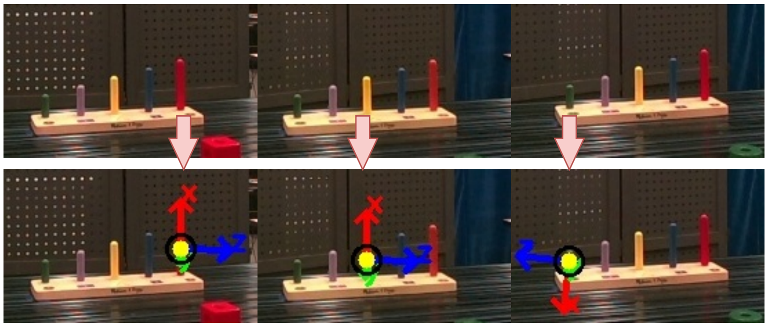


Figure 12: **Comparison of Extracted Local Frames in Drop Task Variants.** (Left) Extracted local frame from the demonstration, aligned with the red rod. (Middle) Inferred frame for a successful execution on the yellow rod, showing good alignment with the rod and consistent trajectory adaptation. (Right) Inferred frame for a failed execution on the green rod, where the extracted frame is misaligned due to depth noise or segmentation error, leading to an incorrect drop trajectory.

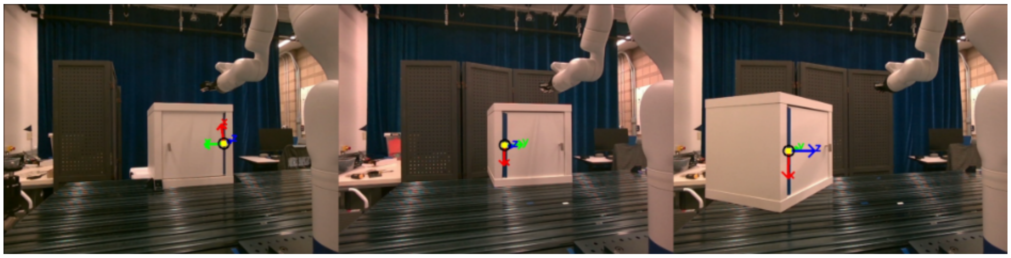


Figure 13: **Comparison of Extracted Local Frames in Cabinet Door Opening Variants.** (Left) Extracted local frame from the demonstration, aligned with the door's hinge and surface in a frontal configuration. (Middle) Frame inferred for a successful mirrored execution, correctly capturing the hinge orientation and aligning the z-axis with the surface normal. (Right) Frame inferred in a failed mirrored execution, where the z-axis is not orthogonal to the surface, leading to an incorrect arc and failed opening trajectory.