GeoPredict: Leveraging Predictive Kinematics and 3D Gaussian Geometry for Precise VLA Manipulation

Supplementary Material

6. Robocasa Simulation Benchmark

In this section, we provide a detailed definition of the sub-tasks used in our evaluation, supplementing Table 1. We conduct our experiments on the RoboCasa simulation framework [34], a large-scale simulation benchmark tailored for training generalist robots in everyday kitchen environments. RoboCasa features high visual and physical realism, offering over 2,500 high-quality 3D assets across more than 150 object categories and dozens of interactable furniture and appliances. Following the standard `Human-50` protocol, we utilize datasets collected via human teleoperation. For our specific experimental setup, we select a subset of 24 atomic tasks from the benchmark to evaluate the manipulation capabilities of our policy. Consistent with focused manipulation assessments, we exclude the `NavigateKitchen` task to isolate stationary manipulation skills from mobile navigation.

The selected 24 atomic tasks represent foundational visuomotor skills that serve as building blocks for complex long-horizon activities. As detailed in the RoboCasa tasks and datasets overview [34], these tasks span diverse skill families, including pick-and-place operations, opening and closing doors and drawers, twisting knobs, turning levers, pressing buttons, and insertion. These tasks require the robot to interact with various articulated objects (e.g., microwaves, cabinets, stoves, sinks) and manipulate a wide range of objects (e.g., vegetables, mugs, bowls) within realistic, clutter-filled kitchen scenes.

Table 6 presents a comprehensive breakdown of these 24 sub-tasks. The first column lists the abbreviations used in Table 1, the second column provides the corresponding atomic task name as defined in RoboCasa, and the third column offers a detailed description of the task objective.

7. Additional Simulation Experiments

7.1. Robocasa Generated-300

Table 7 presents the evaluation results on the RoboCasa `Generated-300` setting. In this regime, the model is trained on a dataset synthesized by MimicGen [34], comprising 300 demonstrations per task (7,200 total trajectories), representing a significant increase in data scale compared to the `Human-50` setting (1,200 total trajectories). Comparing the results in Table 7 with Table 1, we observe that scaling the training data yields a notable performance improvement, particularly within the Pick-and-Place skill family (e.g., `PickPlaceCounterToStove` improves from 20.4% to 42.0%). This performance boost can be attributed to the nature of Pick-and-Place tasks, which involve manipulating high-diversity object categories with varied affordances [34]. As discussed in the original RoboCasa benchmark, atomic tasks with high object diversity are challenging to learn in the few-shot `Human-50` regime. The `Generated-300` setting exposes the policy to a broader distribution of object instances and poses, allowing GeoPredict to leverage its geometric priors more effectively and achieve superior generalization on these tasks.

7.2. Robocasa Human-50

In Table 1 of the main manuscript, the reported performance metrics for both the fine-tuned π_0^* baseline and our GeoPredict framework represent the average success rates derived from five independent evaluation runs using distinct random seeds. Due to spatial constraints in the main text, the statistical variance for these experiments was omitted. To provide a more comprehensive analysis of stability and robustness, we present the complete benchmarking results, including the corresponding standard deviations for all 24 sub-tasks in Table 8.

7.3. LIBERO

Table 2 summarizes the comparative results on LIBERO benchmark suites. To ensure a fair and rigorous comparison with prior state-of-the-art methods [37, 47], we adhere to the standard evaluation protocol. Specifically, the reported success rates for our method are averaged over three independent evaluation runs initialized with different random seeds. GeoPredict achieves an average success rate of 96.5% over four task suites, which validates the stability and robustness of our geometry-aware framework.

8. Additional Real-World Experiments

In-Distribution Results: To further validate whether the proposed method facilitates genuine generalization rather than merely overfitting to the training data, we present the in-distribution (ID) evaluation results in Table 9. As shown, GeoPredict consistently outperforms the π_0 baseline across all ID settings, achieving near-perfect success rates (100.0% in Spatial and Geometry, 95.0% in Robustness). More importantly, comparing these ID results with the out-of-distribution (OOD) performance reported in Table 5 reveals a widening performance margin. For instance, in the Geometry task, our method's advantage over the baseline increases from +15.0% in the ID setting to +45.0% in the

Table 6. **Detailed Definition of 24 Atomic Tasks from RoboCasa Simulation Benchmark.**

Abbreviation	Atomic Task Name	Description
PnP CTC1	`PickPlaceCounterToCabinet`	Pick an object from the counter and place it inside the cabinet. The cabinet is already open.
PnP CTC2	`PickPlaceCabinetToCounter`	Pick an object from the cabinet and place it on the counter. The cabinet is already open.
PnP CTS1	`PickPlaceCounterToSink`	Pick an object from the counter and place it in the sink.
PnP STC1	`PickPlaceSinkToCounter`	Pick an object from the sink and place it on the counter area next to the sink.
PnP CTS2	`PickPlaceCounterToStove`	Pick an object from the counter and place it in a pan or pot on the stove.
PnP STC2	`PickPlaceStoveToCounter`	Pick an object from the stove (via a pot or pan) and place it on (the plate on) the counter.
PnP CTM	`PickPlaceCounterToMicrowave`	Pick an object from the counter and place it inside the microwave. The microwave door is already open.
PnP MTC	`PickPlaceMicrowaveToCounter`	Pick an object from inside the microwave and place it on the counter. The microwave door is already open.
CM SU	`CoffeeSetupMug`	Pick the mug from the counter and insert it onto the coffee machine mug holder area.
CM SV	`CoffeeServeMug`	Remove the mug from the coffee machine mug holder and place it on the counter.
CB PS	`CoffeePressButton`	Press the button on the coffee machine to pour coffee in to the mug.
TSS	`TurnSinkSpout`	Turn the sink spout.
OSD	`OpenSingleDoor`	Open a microwave door or a cabinet with a single door.
CSD	`CloseSingleDoor`	Close a microwave door or a cabinet with a single door.
ODD	`OpenDoubleDoor`	Open a cabinet with two opposite-facing doors.
CDD	`CloseDoubleDoor`	Close a cabinet with two opposite-facing doors.
OD	`OpenDrawer`	Open a drawer.
CD	`CloseDrawer`	Close a drawer.
TNSF	`TurnOnSinkFaucet`	Turn on the sink faucet to begin the flow of water.
TFSF	`TurnOffSinkFaucet`	Turn off the sink faucet to end the flow of water.
TNS	`TurnOnStove`	Turn on a specified stove burner by twisting the respective stove knob.
TFS	`TurnOffStove`	Turn off a specified stove burner by twisting the respective stove knob.
TNM	`TurnOnMicrowave`	Turn on the microwave by pressing the start button.
TFM	`TurnOffMicrowave`	Turn off the microwave by pressing the stop button.

Table 7. **RoboCasa Simulation Benchmark Results on Generated-300 Setting.** Task success rates (%) across 24 sub-tasks and the Average Success Rate (%). **Bold** indicates the best performing model.

Method	PnP CTC1	PnP CTC2	PnP CTS1	PnP STC1	PnP CTS2	PnP STC2	PnP CTM	PnP MTC	CM SU	CM SV	CB PS	TSS
BC-Transformer [34]	16.0	10.0	16.0	14.0	0.0	4.0	0.0	12.0	4.0	24.0	42.0	58.0
GeoPredict (Ours)	**52.0**	**18.0**	**58.0**	**38.0**	**42.0**	**52.0**	**14.0**	**24.0**	**28.0**	**46.0**	**74.0**	**92.0**
Average Success Rate	**OSD**	**CSD**	**ODD**	**CDD**	**OD**	**CD**	**TNSF**	**TFSF**	**TNS**	**TFS**	**TNM**	**TFM**
35.0	44.0	86.0	22.0	62.0	40.0	98.0	48.0	46.0	44.0	12.0	**76.0**	**62.0**
55.4	**56.0**	**94.0**	**56.0**	**86.0**	**66.0**	**100.0**	**74.0**	**72.0**	**56.0**	**20.0**	54.0	58.0

Table 8. **RoboCasa Simulation Benchmark Results with Standard Deviation.** Task success rates (%) across 24 sub-tasks. *Denotes our fine-tuned experimental results. **Bold** indicates the best performing model.

Method	PnP CTC1	PnP CTC2	PnP CTS1	PnP STC1	PnP CTS2	PnP STC2
BC-Transformer [34]	6.0	2.0	2.0	8.0	2.0	6.0
GWM [30]	4.0	**18.0**	20.0	22.0	2.0	18.0
π_0^* (Baseline) [3]	24.0 ± 4.7	6.8 ± 3.0	26.8 ± 5.0	15.6 ± 2.2	11.2 ± 3.6	12.8 ± 3.9
GeoPredict (Ours)	**32.4 ± 5.0**	8.8 ± 3.9	**27.6 ± 5.9**	**31.2 ± 7.2**	**20.4 ± 7.9**	**28.8 ± 5.4**
Method	**PnP CTM**	**PnP MTC**	**CM SU**	**CM SV**	**CB PS**	**TSS**
BC-Transformer [34]	2.0	2.0	0.0	22.0	48.0	54.0
GWM [30]	14.0	**20.0**	16.0	36.0	**76.0**	**72.0**
π_0^* (Baseline) [3]	6.8 ± 3.3	7.6 ± 3.3	18.4 ± 6.1	33.6 ± 6.1	74.0 ± 5.8	69.6 ± 3.8
GeoPredict (Ours)	**14.4 ± 4.6**	18.0 ± 5.1	**28.4 ± 3.8**	**49.2 ± 8.3**	67.6 ± 2.6	70.8 ± 6.3
Method	**OSD**	**CSD**	**ODD**	**CDD**	**OD**	**CD**
BC-Transformer [34]	46.0	56.0	28.0	28.0	42.0	80.0
GWM [30]	**58.0**	54.0	28.0	50.0	56.0	80.0
π_0^* (Baseline) [3]	40.8 ± 7.0	**82.0 ± 7.1**	55.2 ± 13.2	69.2 ± 6.1	66.4 ± 9.4	96.0 ± 2.8
GeoPredict (Ours)	40.4 ± 11.0	78.8 ± 2.7	**87.2 ± 2.3**	**70.0 ± 4.7**	**77.6 ± 7.3**	**96.8 ± 1.1**
Method	**TNSF**	**TFSF**	**TNS**	**TFS**	**TNM**	**TFM**
BC-Transformer [34]	38.0	50.0	32.0	4.0	62.0	70.0
GWM [30]	52.0	44.0	46.0	**22.0**	64.0	70.0
π_0^* (Baseline) [3]	43.6 ± 7.3	86.0 ± 2.8	43.2 ± 2.3	6.0 ± 4.0	59.6 ± 4.3	60.0 ± 5.3
GeoPredict (Ours)	**72.4 ± 6.1**	**94.8 ± 3.0**	**60.0 ± 8.2**	13.2 ± 5.0	**84.8 ± 7.2**	**82.8 ± 4.6**

Table 9. **Real-World In-Distribution (ID) Experiment Results.** Task success rates (%) across three distinct settings: Spatial, Geometry and Robustness.

Method	Spatial	Geometry	Robustness
π_0 (Baseline) [3]	90.0	85.0	70.0
GeoPredict (Ours)	100.0	100.0	95.0

OOD setting. This growing gap confirms that our predictive 3D Gaussian geometry module effectively provides robust structural priors, ensuring strong generalization to novel geometric and spatial variations.

Temporal Reasoning: To empirically validate the efficacy of the Track Encoder in resolving temporal ambiguities, we introduce a dedicated task requiring explicit memory of past events. The setup consists of two plates placed side-by-side on the workspace with a green cube initialized in one of them. The robot is tasked with transferring the object to the opposing plate given the instruction: `move the green cube from one plate to the other plate`. This task creates a significant state aliasing challenge because the visual observations during the transport phase are symmetric. Specifically, as illustrated in Figure 5, the images captured when the gripper holds the cube between the plates are visually indistinguishable between a Left-to-Right trajectory (top row) and a Right-to-Left trajectory (bottom row). Consequently, a purely reactive policy relying solely on current observations fails to infer the correct transport direction. Success in this task strictly requires the agent to leverage motion history to maintain directional consistency.

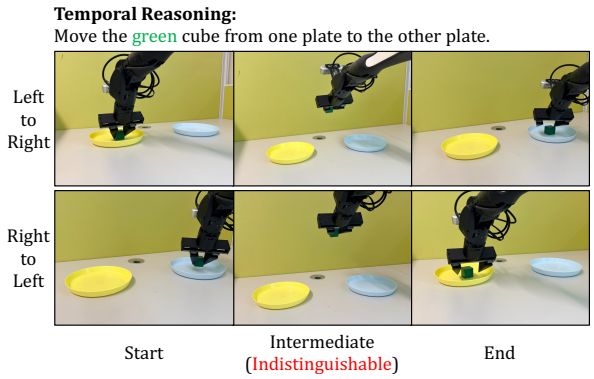

Figure 5. **Temporal Reasoning Evaluation.** Illustration of two distinct trials with opposite transport directions (top: Left-to-Right, bottom: Right-to-Left). The intermediate observations are visually indistinguishable, requiring the policy to leverage motion history to infer the correct target plate.

Table 10. **Additional Real-World Experiment Results.** Task success rates (%) on Temporal Reasoning setting.

Method	Temporal
π_0 (Baseline) [3]	25.0
GeoPredict (Ours)	70.0

We utilized 50 expert trajectories for training to establish this behavior. For each demonstration, the cube was randomly initialized in either the left or right plate and then transported to the opposite side. This randomization ensures that the policy cannot overfit to a single direction and must rely on the context of the starting position. During the evaluation phase, we tested each model over 20 trials. A trial is considered successful only if the robot infers the correct destination plate based on the motion history and successfully deposits the cube into it.

As presented in Table 10, GeoPredict achieves a success rate of 70.0%, substantially outperforming the π_0 baseline, which attains only 25.0%. This significant performance gap underscores the inherent limitation of purely reactive VLA policies when confronting state aliasing; without temporal context, the baseline fails to disambiguate the symmetric visual observations encountered during the transport phase. In contrast, GeoPredict effectively utilizes the Track Encoder to explicitly model motion history, enabling the policy to resolve this visual ambiguity and maintain directional consistency throughout the trajectory.

Supplementary Videos: We provide comprehensive rollout videos comparing our GeoPredict framework with the π_0 baseline across all four real-world experimental settings. These demonstrations cover the Spatial Generalization, Geometry Generalization, and Visual Robustness tasks presented in the main paper, as well as the additional Temporal Reasoning task detailed in this supplementary material. We refer the reader to the accompanying video folder for visual validation of our method's performance and capabilities.