Any-ttach: Quick End-effector Swapping Enables Manipulation Dexterity with Simplicity

Anonymous Submission

***Abstract*— Robotic manipulation dexterity is typically pursued through high-DoF multifingered hands with substantial complexity. While many robotic hands aim to replicate human hands, the functional role of human hands suggests that much of this complexity is to support tool using and making. This observation motivates *Any-ttach*: we leverage the flexibility of mechanical systems and directly employ end-effector swapping as a central mechanism for dexterity. *Any-ttach* consists of 1) a low-cost, automatic, and easily deployable end-effector swapping mechanism that interfaces with a 1-DoF parallel gripper, 2) a handheld device for collecting human demonstration data, and 3) a task planning framework that integrates learned, parameterized, and planned skills to enable flexible use of diverse everyday tools. In the experiments, we evaluate the effectiveness of the swapping mechanism, showing that this kinematic constrained coupling provides unique advantages in robust skill demonstration and tool-changing efficiency. The system demonstrates broad compatibility across different skills with different end-effectors, spanning regular everyday tools, articulated tools (e.g., scissors), and a low-cost anthropomorphic hand. We verify the system on two long-horizon tasks (making sandwich, preparing cucumber) with 6 subskills, demonstrating that hierarchical tool–skill decomposition improves complex task reliability. We hope *Any-ttach* offers an alternative path to manipulation dexterity with simplicity — dexterity can emerge from composing skills across end-effectors and exploiting contact interactions within each skill. In this view, even a simple 1-DoF gripper can achieve greater manipulation dexterity. More details and videos are available at `https://any-ttach.github.io/`.**

I. Introduction

General-purpose dexterity remains a central challenge in robotic manipulation. Most existing approaches treat dexterity as an intrinsic property of the end effector and seek broader capability by increasing mechanical complexity through high degree of freedom hands, richer contact surfaces, and dense sensing [1]–[4]. Although these systems demonstrate impressive performance, they tightly couple manipulation dexterity to hardware complexity and complex contact-rich control. Even simple tasks require precise regulation of high dimensional contact dynamics, and the challenge becomes greater when long horizon tasks demand both low level contact control and high level skill sequencing.

An alternative perspective highlights extrinsic dexterity, where simple grippers achieve diverse behaviors by exploiting environmental constraints [5]. This perspective suggests that manipulation capability can emerge from intelligent contact interactions using simple end-effectors. Furthermore, evolutionary perspectives suggest that many distinctive features of the human hand emerged to support tool use and tool making [6], [7]. Combining the evolution perspective with extrinsic dexterity suggests that manipulation dexterity

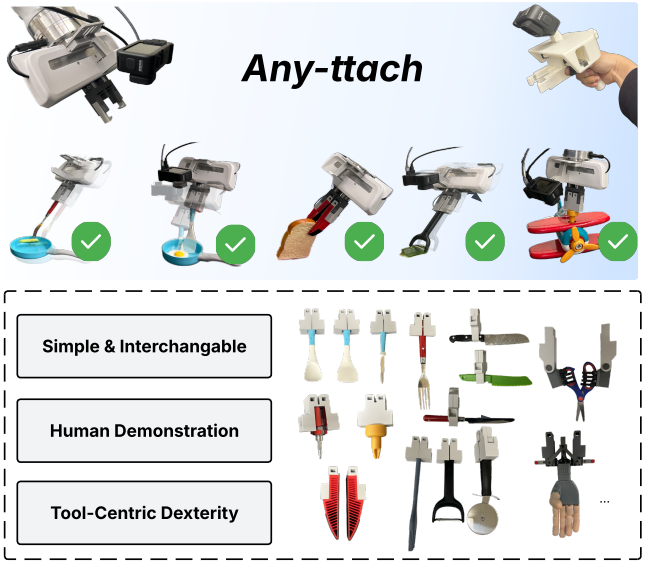

Fig. 1. **Tool-centric design achieves dexterity and end-effector swapping with simplicity.** Any-ttach enables robots to perform diverse manipulation skills by rapidly switching between interchangeable tool modules through a standardized mechanical interface. By externalizing task-specific contact geometry into tools, the system reformulates manipulation dexterity as tool selection and skill execution rather than complex end-effector control.

may arise from switching across simple end-effectors/tools and executing skills with those simple end-effectors with intelligent contact interactions. Switching tools via grasping is challenging because a grasped tool creates a second contact interface between the manipulator and the tool. Disturbances at either the manipulator tool interface or the tool object interaction can lead to failure [8]–[10]. Prior work addresses this challenge either by using high degree of freedom hands to actively regulate the manipulator tool interface through force redistribution or regrasping [11]–[13], or by permanently mounting task specific end effectors that limit flexibility across tasks.

We propose *Any-ttach*, a manipulation system built around an automatic quick-swap mechanical interface that enables a 1-DoF parallel gripper to directly attach and exchange everyday tools. Instead of relying on dexterous grasping or permanent attachment to use tools, *Any-ttach* provides efficient automatic end-effector swapping and reliable mechanical attachment, supporting stable tool use in contact-rich interactions. The system adopts a hierarchical architecture in which a task planner decomposes high-level instructions into ordered tool–skill pairs, and skill modules execute each behavior in closed loop. These skills can be implemented through learned policies, parameterized controllers,

or planning-based methods. We also design a handheld tool interface that shares the same coupling mechanism to facilitate human demonstration collection and reduce the human–robot embodiment gap. Through this design, we aim to explore an alternative path of manipulation dexterity with simplicity: dexterity can emerge from composing skills across end-effectors and exploiting contact interactions within each skill, allowing even a simple 1-DoF gripper to achieve rich manipulation capabilities. Our contributions are summarized as follows:

- An automatic quick-swap mechanical interface that enables a 1-DoF parallel gripper to directly attach and exchange diverse everyday tools, including 0-DoF tools, 1-DoF articulated tools (e.g., scissors and tweezers), robot fingers, and a low-cost anthropomorphic hand. With our designed handheld interface, it provides a flexible and standardized coupling mechanism for both execution and demonstration.
- A hierarchical framework that integrates task planning, tool selection, skill composition, and skill execution. The framework supports heterogeneous skill implementations and flexible integration for different end-effectors.
- A comprehensive experimental study demonstrating reliable tool swapping, compatibility with diverse tools and skill types, and long-horizon manipulation through composed tool-use behaviors.

II. Related Work

A. Dexterity in Robotic Manipulation

A dominant paradigm in manipulation dexterity pursues intrinsic dexterity, embedding capability directly in the end-effector morphology. High-DoF manipulators [4], [14] operate in high-dimensional contact spaces and typically rely on large-scale simulation and learning infrastructure [15], [16]. The previous work [14] demonstrated in-hand cube rotation via reinforcement learning, and subsequent work advances dexterous grasping [3], object-centric manipulation [2], and functional grasping policies [12]. In contrast, extrinsic dexterity [5] shows that simple grippers can achieve rich behaviors, with extensions through structured primitives [17], [18] and contact-mode control [19]. While intrinsic approaches embed capability within the hand itself, extrinsic approaches leverage structure available in the environment [20]. Our approach differs from both directions. Instead of increasing in-hand complexity or relying on environmental fixtures, we externalize task-specific contact geometry into interchangeable tool modules. This shifts part of the interaction structure from continuous control into simple geometric design, enabling dexterity through controlled reconfiguration of the end-effector rather than higher morphological complexity.

B. Tool Manipulation

Tool use has been studied as a practical means of extending manipulation capability beyond the native end-effector geometry [11]. Existing work addresses different components of the tool pipeline, including affordance learning [13], pose estimation [21], [22], and contact-rich skill acquisition via imitation learning [23], dynamics learning [24], and compliant control [25]. More recent approaches incorporate LLMs for tool selection and sequencing [26]. Despite this progress, most systems assume that tools are grasped by general-purpose grippers. Under this assumption, execution forces can perturb the grasp, leading to tool pose variability that reduces repeatability—particularly for learned policies trained under fixed tool geometry. Existing solutions either employ high-DoF hands to regulate the grasp during execution or permanently mount task-specific end-effectors, trading off flexibility.

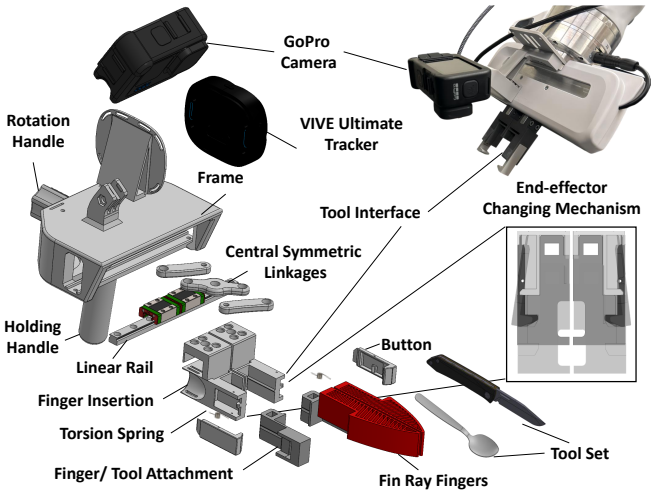

Fig. 2. **Hardware Design.** A self-aligning quick-swap interface enables repeatable attachment and autonomous exchange among interchangeable tool modules, allowing the robot to switch end-effectors during multi-step tasks. Standardized coupling is shared by robot end-effector and handheld demonstrator for consistent tool geometry during data collection.

C. Industrial Tool Changers

Industrial tool changers enable quick swapping through various mechanisms such as pneumatically-actuated cam-and-ball locks [27], [28], bayonet locks driven by robot wrist motion [29], magnetic couplings [30], [31], 3D-printed kinematic couplings [32], but they typically require extra actuation, are designed for pre-programmed industrial workflows, only pair with specially designed tools, and are hard to be applied for everyday tools. In contrast, our approach provides a passive and automatic quick-swap interface that does not require extra actuation, is compatible with planning and learning-based manipulation frameworks, and can be easily mounted on various everyday tools/fingers and allow 1-DoF actuation on the tools.

D. Hierarchical Planning and Skill Composition

Hierarchical planning and skill learning are widely used to address long-horizon manipulation [33]. Classical symbolic planning relies on manually specified models and precondition–effect representations [34], [35], while more recent work leverages foundation models for task planning and grounding [36]–[41], enabling flexible semantic decomposition. At the execution level, imitation learning and sequence

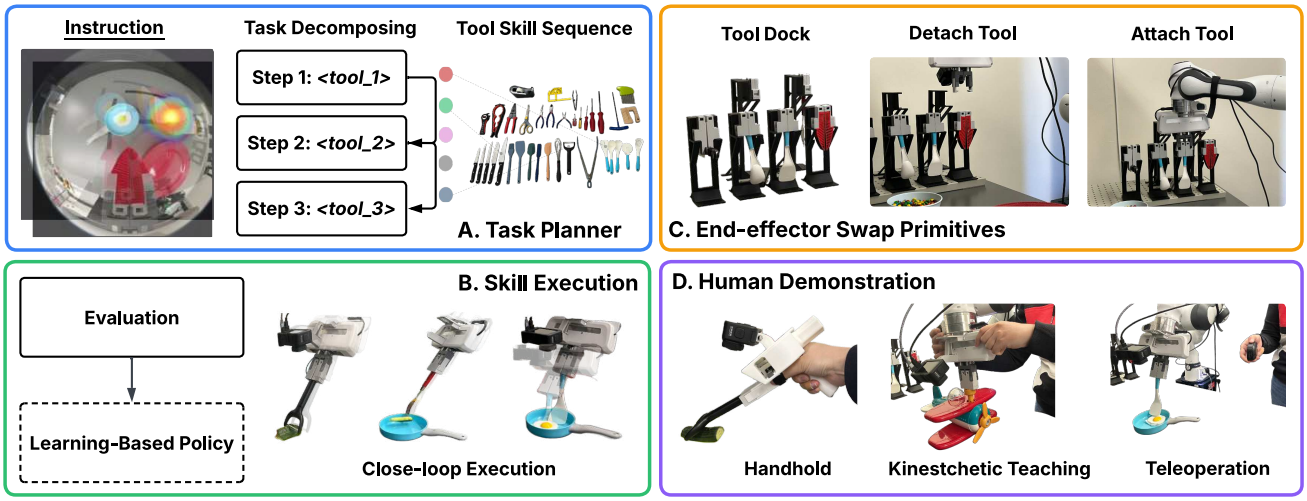

Fig. 3. **System pipeline of Any-ttach.** (A) **Task Planner:** a vision language model decomposes instruction into an ordered sequence of tool–skill pairs. (B) **Skill Execution:** learning-based policies execute each skill in closed loop using visual and proprioceptive observations. (C) **End-effector swap Primitives:** the robot autonomously docks, attaches, and detaches tool modules through the standardized quick-swap interface. (D) **Human demonstration:** a handheld interface compatible with the robot coupling enables demonstration collection through handheld manipulation, kinesthetic teaching, or teleoperation. The unified tool interface ensures consistent tool geometry across demonstration and execution, enabling reliable tool use and long-horizon manipulation.

models such as ACT [42] and Diffusion Policy [43] learn closed-loop manipulation primitives from demonstrations. However, most hierarchical systems assume a fixed end-effector and do not explicitly reason about geometry reconfiguration through tool switching. We integrate planner-level tool reasoning with diffusion-based skill execution under the kinematically constrained tool interface, enabling long-horizon behavior through structured tool–skill composition rather than static embodiment control.

III. Hardware System

Our hardware system consists of the unified coupling interface (Sec. III-A), the quick-swap mechanism (Sec. III-B), the handheld demonstration device (Sec. III-C).

A. Unified Tool Coupling Interface

The core hardware component is a standardized mechanical coupling that mediates attachment between interchangeable tool modules and either the robot end-effector or the human demonstration device (Fig. 2). Each tool module terminates in an identical connector; both embodiments provide the corresponding receptacle geometry. The coupling incorporates passive alignment features of guide rails and tapered constraints that guide insertion to a single engagement configuration. As a result, once attached, the relative transform between the robot flange and tool frame is fixed across trials. By replacing grasp-based tool acquisition with rigid mechanical attachment, the interface eliminates grasp-induced pose variability, a common source of execution instability in learned tool manipulation. This unified tool interface coupling kinematically provides a consistent reference for policy learning and deployment.

To adapt diverse tool handles to this connector, we approximate each handle's geometry with measures and fabricate tool-side adapters via a boolean cut from a shared CAD template, adding small tolerance sizes to ensure a tight fit. This ensures that every tool attaches in a unique, kinematically constrained pose, guaranteeing spatial consistency across demonstrations and executions.

B. Tool Quick-Swap Mechanism

Autonomous long-horizon manipulation requires tool interchange without human intervention. The automatic quick-swap mechanism provides one-motion locking and release: the robot attaches a tool module by inserting it into the coupling mechanism by activating the passive self-locking structure. During detachment, the robot triggers release by knocking into an internal latch structure: contact with the dock-side release feature induces a rotation of the button about its pivot, compressing a spring on the opposite end and lifting the locking side to disengage the passive self-locking structure. No threaded fasteners or auxiliary hardware are required. The mechanism is back-compatible: when no tool module is attached, the robot retains its native parallel-jaw gripper configuration, preserving compatibility with standard parallel grasping.

C. Human Demonstration Device

To collect demonstrations that transfer directly to robot execution, we design a handheld device that shares the same coupling interface as the robot (Fig. 2). Unlike the popular trigger-based or handle-pivoted demonstration tools (e.g., UMI [44]), our device adopts a finger-mounted coupling: operators manipulate the tool module directly with their fingers, preserving intuitive rotational strategies and contact dynamics from human. A holding handle provides wrist-level stability without constraining dexterous finger motion.

To ensure spatial consistency between human demonstration and robotic execution, we integrate a VIVE pose tracker [45] into the handheld interface to provide real-time

6-DoF pose estimation of the coupled tool in a global reference frame. Tool poses are recorded relative to a calibrated robot coordinate frame, enabling direct replay under identical tool geometry and coupling constraints. This interface-level alignment eliminates grasp-induced pose variability, removes the need for kinematic retargeting between dissimilar embodiments, and reduces embodiment-induced variance, improving data efficiency for learning-based skill acquisition.

IV. Hierarchical Planning and Execution

We structure Any-ttach as a three-stage pipeline to decompose long-horizon manipulation tasks (Fig. 3). We define the execution flow as follows: (1) a planner maps task instructions to an ordered sequence of tool-skill pairs; (2) the closed-loop execution of each skill; and (3) a verifier confirms task success, automatically initiating retries upon failure. To define the system formally, we equip the robot with a quick-swap interface $\mathscr{I}$, a discrete library of M tools $\mathscr{Z} = \{z_1, \ldots, z_M\}$, and a set of N skills $\Sigma = \{\sigma_1, \ldots, \sigma_N\}$. Each tool $z \in \mathscr{Z}$ provides a distinct contact geometry, while each skill $\sigma \in \Sigma$ serves as a control primitive optimized for that geometry. Consequently, the system outputs a complete plan as an ordered progression of these (z, σ) pairs.

A. Task-Level Planning

We implement a high-level planner to decompose a natural language instruction into executable subtasks. Specifically, a vision–language model [46] is queried with the instruction, current observation, and the available tool library to generate an ordered sequence of subtask–tool pairs. Each subtask is executed using a corresponding learned skill.

B. Skill-Level Execution

Any skill types, including planning-based skills, model-based control, or learned policies, can be integrated within this framework. We implement learned skills and planning-based skills. A learned skill $\sigma^{(i)}$ is implemented as a diffusion policy [43] $\pi_\theta^{(i)}$ trained from demonstrations. The policy maps observation to action chunk. Because the manipulator–tool transform is fixed by mechanical design, the policy operates under a stable tool frame across trials. This reduces cross-trial geometric variance and improves deployment reliability compared to grasp-based tool holding. Receding-horizon execution enables closed-loop correction during contact-rich interaction. We also implement planning-based skills that use existing free-space motion planning methods [47] to reach target poses and execute pre-designed behaviors.

C. Evaluation

We implemented two verifiers to check for the skill pre-condition and post-condition to ensure reliable progression across tool-skill transitions.

Pre-condition: tool attachment. Before executing a skill, the system verifies that the correct tool has been successfully attached. This is assessed using a combination of vision-language reasoning and segmentation-based tool identification. Specifically, the planned tool name serves as a semantic prompt for SAM3 [48], which generates a tool-specific segmentation mask for attachment verification.

Post-condition: skill execution. After skill execution, a VLM-based verifier evaluates whether the intended task outcome has been achieved according to the skill description. The VLM reasons over the current scene observation and the target objective to determine success or failure. If it fails, the skill is retried up to 3 times before proceeding.

V. Experiments

Any-ttach system integrates three key design elements: a kinematically constrained tool interface, a shared human–robot demonstration setup, and a hierarchical planning-execution-evaluation architecture. We therefore organize our experiments around three questions that examine their effects on swapping and demonstration efficiency, skill execution reliability, and long-horizon task robustness:

- **Q1: Swapping mechanism fffectiveness.** How reliable and efficient is the proposed end-effector swapping mechanism compared to grasp-based tool changing?
- **Q2: Tool coverage and skill capability.** What range of tools and manipulation skills can the standardized interface support, and how reliably can these skills be executed?
- **Q3: Hierarchical Robustness.** Does the hierarchical tool–skill planning and evaluation architecture improve robustness in long-horizon multi-tool tasks?

A. Experimental Setup

System setup. We evaluate Any-ttach on a Franka Research 3 (FR3) 7-DoF robot arm equipped with an autonomous tool dock and the proposed kinematically constrained coupling interface. The same coupling geometry is integrated into a handheld demonstration device used for human data collection.

Tool sets. We use two tool sets in our experiments. (i) *Long-horizon tool set:* six interchangeable modules used for autonomous multi-step tasks: gripper, spatula, spoon, peeler, knife, and fork (Fig. 5). (ii) *Extended tool set:* tools attached to evaluate tool coverage and single-skill capability, including daily tools (e.g., brush, pizza wheel, whisk, screwdriver) and unconventional end-effectors (e.g., Fin Ray gripper fingers, toy hand, and scissors). We have in total 15 tools in both sets, and they can all be coupled to either the robot end-effector or the handheld device via the same interface.

Tasks. In Sec V-D , we evaluate the system on multi-step manipulation tasks that require switching tools and executing several skills sequentially. Skill policies operate in closed loop and receive as observations an RGB image of the workspace, the 6-DoF pose of the attached tool interface in the robot frame, and a binary gripper state.

Success metric. We consider two representative tasks (Fig. 5): *Sandwich Assembly* requires three skills: pick-and-place (gripper), flip (spatula), and scoop (spoon), and *Cucumber Preparation* also consisting of peel (peeler), cut (knife), and spear (fork). These tasks require accurate tool

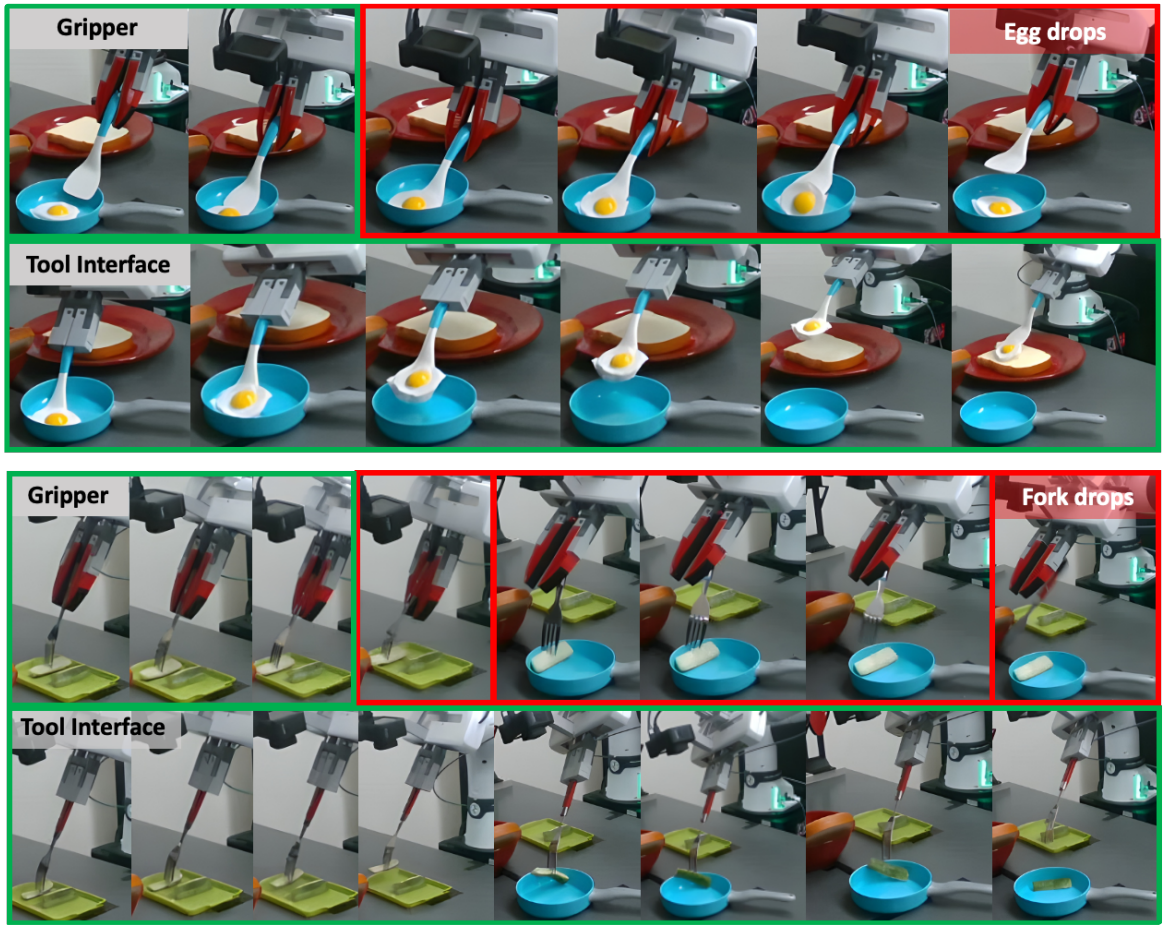

Fig. 4. **Gripper failure cases. Top:** during spatula flipping, contact forces induce tool rotation within the parallel-jaw gripper, causing the grasped tool pose to tilt and the egg to drop (red box). **Bottom:** during fork spearing, similar grasp-induced pose drift accumulates over the skill execution and leads to tool loss and task failure (red box). In contrast, our kinematically constrained tool interface maintains a fixed tool pose under the same interactions, enabling stable contact and successful completion (green boxes). These examples illustrate how grasp-based tool holding introduces a second unstable interface (manipulator–tool) whose errors accumulate under extrinsic, force-transmitting manipulation.

alignment and sustained contact during manipulation. A skill trial is successful if its sub-goal is completed without tool slippage or early termination, and a full task succeeds only if all skills complete in order.

B. Evaluation on End-Effector Swapping Performance

We evaluate the end-effector swapping subsystem from two perspectives: (i) autonomous swapping performance (reliability and time) and (ii) its impact on demonstration efficiency during data collection.

Swapping effectiveness. We evaluate the end-effector swapping mechanism in terms of average swapping time and swapping success rate (Table I). Mechanically, the operation resembles a peg-in-hole insertion: passive alignment features funnel the tool module into a unique engagement configuration, allowing the system to tolerate moderate pose mismatch during approach. In practice, the compliance control of the FR3 robot arm further helps accommodate small residual errors during contact, improving swap robustness.

We compare against two common alternatives for tool changes during demonstrations: gripper-based tool acquisition and manual tool changes. Gripper-based tool changes require grasping and stabilizing the tool handle before use, making the outcome sensitive to grasp pose and contact disturbances. As shown in Table I, grasp-based acquisition achieves a swapping success rate of 55.6% with an average time of 18.59s, while the Any-ttach interface improves success to 87.5% and reduces the average time to 17.00s. Using the same interface with human-assisted attachment/detachment ("Any-ttach with hand") further reduces swapping time to 11.09s with 83.3% success, while a direct grasp of the tool with human hand provide a gold standard reference (3.23s, 100%). Overall, these results show that mechanically constrained coupling substantially improves the reliability of tool changes compared to grasping, and reduces operator overhead during demonstrations by making tool switching a repeatable attachment action rather than a grasp-adjust-secure procedure.

Demonstration effectiveness. We also evaluate demonstration efficiency across three demonstration modes: gripper-based teleoperation, Any-ttach teleoperation with the

TABLE I

SWAPPING EFFICIENCY COMPARISON.

SWAPPING TIME IS MEASURED FROM TOOL DETACHMENT TO REACHING A USABLE POSE AFTER THE NEW TOOL IS ATTACHED; "ANY-TTACH INTERFACE" IS FULLY AUTONOMOUS, WHILE "ANY-TTACH WITH HAND" USES HUMAN-ASSISTED ATTACHMENT/DETACHMENT.

Method	Avg. Swapping Time (s)	Avg. SR
Gripper	18.59	55.6%
Any-ttach interface	**17.00**	**87.5%**
Any-ttach with hand	11.09	83.3%
Human hand	3.23	100%

TABLE II

DATA COLLECTION EFFICIENCY COMPARISON.

REPORTING AVERAGE DEMONSTRATION TIME PER TRIAL AND THE PROPORTION OF DEMONSTRATIONS USABLE FOR POLICY TRAINING.

Method	Avg. Demo Time (s)	Avg. Usable Rate
Gripper-tool teleop	41.24	88.97%
Any-ttach teleop	36.79	96.10%
Any-ttach handheld	10.03	–

standardized interface, and direct handheld demonstrations. We report (i) average demonstration time per trial and (ii) usable demonstration rate, defined as the fraction of collected trajectories that are successful and suitable for diffusion-policy training.

As shown in Table II, interface-based teleoperation reduces demonstration time from 41.24s to 36.79s (10.8%) and increases the usable rate from 88.97% to 96.10%. Direct handheld demonstrations further reduce collection time to 10.03s per trial (3.7× faster than gripper-tool teleoperation) by removing robot-side actuation latency and eliminating retargeting between dissimilar embodiments. Overall, these results indicate that the shared tool interface improves both the efficiency and yield of demonstration collection for tool-centric skills.

C. Tool Coverage and Single-Skill Capability

Tool coverage. Any-ttach's tool interface is designed to support a wide range of interchangeable tools with diverse geometries and functional purposes. In our experiments, we attach multiple categories of tools to the standardized interface, including the Fin Ray gripper, kitchen utensils (spatula, spoon, fork, peeler, knife, brush, pizza wheel, and whisk), a screwdriver for assembly tasks, and a variety of other common tools such as a low-cost anthropomorphic toy hand, and scissors. All of these tools can be coupled to both the handheld demonstration device and the robot end-effector using the same mechanical interface. This demonstrates that the interface is not limited to a specific task or tool type, but can accommodate diverse tool geometries and manipulation functions within a unified coupling mechanism.

Comparison with tool-attachment via grasping. We compare against a grasp-based tool baseline in which tools are held by the parallel-jaw gripper, with the attachment mechanism as the sole difference from Any-ttach. As for the attachment/grasping poses, kinematically constrained coupling reduces positional variance by up to 5.1 times and rotational variance by up to 5.5 times relative to the grasp baseline, providing the geometric consistency that underlies the observed improvements in policy reliability.

TABLE III

SKILL SUCCESS RATE (15 TRIALS PER SKILL).

× INDICATES THE SKILL CANNOT BE EXECUTED WITH THE METHOD DUE TO UNACHIEVABLE DEMONSTRATIONS WITH THIS METHOD FOR THE SKILL, PREVENTING POLICY TRAINING AND DEPLOYMENT.

Skill	Gripper-tool	Any-ttach
Pick & Place	10/15	10/15
Flip	7/15	**12/15**
Scoop	6/15	**8/15**
Spread	12/15	13/15
Peel	×	**9/15**
Cut	×	**14/15**
Spear	×	**8/15**

On the skill execution, we train identical diffusion-based skill policies under two conditions: (i) grasp-based holding and (ii) tool interface coupling (Any-ttach). Each skill is trained from 30 demonstrations and evaluated over 15 trials. We use *Deployment Success Rate* to evaluation the skill-level performance, defined as the proportion of successfully completed task trials. This metric measures the effectiveness of policies trained using data collected with the Any-ttach system. As shown in table III, Any-ttach's tool interface coupling improves average sandwich assembly skill success from 44.4% to 71.1%. It also enables three skills that cannot be achieved with the gripper-tool interface (Peel, Cut, Spear).

Peeling and cutting require regulating tool–object contact forces. To approximate this without explicit force control, we use a simple impedance control heuristic, implemented as a 5–7mm positional offset along the tool normal. Success is defined as completing a continuous tool motion along the intended interaction direction (e.g., peeling from the top to the end of the cucumber surface or executing a full cutting stroke), while failures occur when the tool repeatedly slides on the surface without making progress.

Failure mode. In addition, as shown in Fig. 4, we analyze failure modes in a gripper-tool baseline and show how Any-ttach's tool-centric interface mitigates common issues such as tool pose drift and unstable contact during tool use. For tasks like flipping with a spatula or spearing with a fork, the interaction typically introduces contact forces normal to the spatula surface. These forces are transmitted to the handle-gripper contact and generate a tangential component along the gripper's grasping surface. As a result, the tool can slip or rotate within the gripper, leading to tool pose drift and unstable contact during execution. By rigidly coupling the tool through a standardized interface, Any-ttach reduces tool motion complexity dependent on grasping pose, and improves repeatability under contact-rich interactions. In contrast, for tasks like spreading with a brush or scooping with a spoon, the dominant interaction forces are more

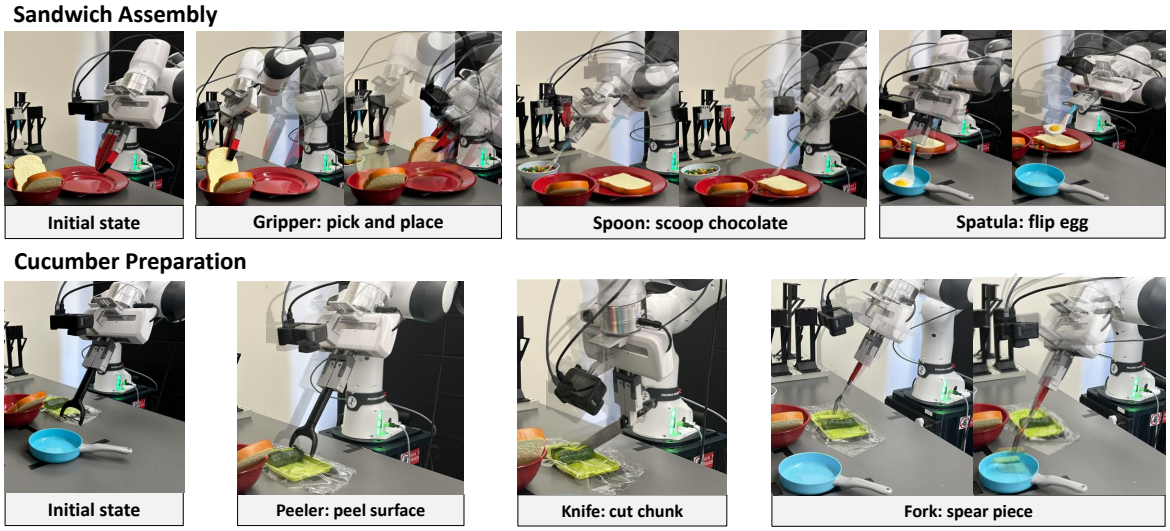

Fig. 5. We evaluate our system on two long-horizon tasks. **Sandwich Assembly:** the robot 1) picks and places bread, 2) scoops filling, 3) flips the fried egg onto the bread. **Cucumber Preparation:** the robot 1) uses the peeler to peel the cucumber, 2) cuts it in half with a knife, 3) uses a fork to spear the cucumber chunk into the pot.

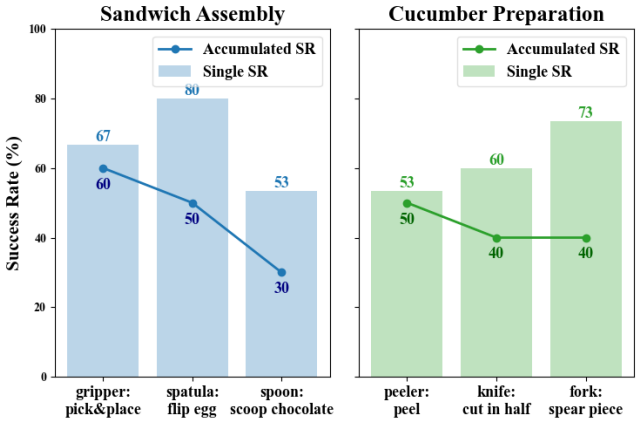

Fig. 6. Single-skill vs accumulated success rates in long-horizon tasks. The difference between Single SR and Accumulated SR highlights how intermediate failures compound over skill sequences in long-horizon tasks.

aligned with directions constrained by the gripper's closing force and frictional support. Consequently, tool pose variation in the gripper is less likely to accumulate, and performance differences between the gripper baseline and the tool-centric interface are smaller for this skill. These observations motivate selecting contact-rich skills such as flipping and spearing as representative benchmarks for evaluating tool-centric interface significance.

D. Hierarchical Robustness on Long-horizon Tasks.

Fig. 6 compares the *Single Skill Success Rate* (Single SR) and the *Accumulated Success Rate* (Accumulated SR) across the two long-horizon tasks. While individual skills achieve relatively high success rates when executed in isolation, the accumulated task success decreases as failures propagate across sequential steps. This gap between single-skill and accumulated success highlights a key challenge in long-horizon manipulation: small failures at intermediate stages can propagate and cause downstream task failure. The proposed two-step condition evaluation mitigates this issue by detecting incomplete outcomes and triggering skill re-execution when necessary. As a result, the system can recover from intermediate errors and maintain higher task robustness.

VI. Conclusion

In this work, we present *Any-ttach*, a tool-centric hierarchical manipulation system that integrates a shared human–robot tool interface with a VLM-based planner and diffusion-based skill execution. By standardizing tool coupling and enabling structured tool–skill composition, *Any-ttach* achieves improved reliability and data efficiency in long-horizon tasks. Our evaluations demonstrate that the kinematically constrained mechanical interface coupling enhances skill execution success, while the shared interface streamlines demonstration collection. The hierarchical architecture further enables robust task-level performance through evaluation and retry. Any-ttach offers a promising direction for decoupling dexterity from morphological complexity through standardized tool interfaces and structured planning.

References

[1] H. Qi, A. Kumar, R. Calandra, Y. Ma, and J. Malik, "In-hand object rotation via rapid motor adaptation," in *Conference on Robot Learning*. PMLR, 2023, pp. 1722–1732.

[2] Y. Chen, C. Wang, Y. Yang, and C. K. Liu, "Object-centric dexterous manipulation from human motion data," *arXiv preprint arXiv:2411.04005*, 2024.

[3] R. Wang, J. Zhang, J. Chen, Y. Xu, P. Li, T. Liu, and H. Wang, "Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation," *arXiv preprint arXiv:2210.02697*, 2022.

[4] K. Shaw, A. Agarwal, and D. Pathak, "Leap hand: Low-cost, efficient, and anthropomorphic hand for robot learning," *arXiv preprint arXiv:2309.06440*, 2023.

[5] N. C. Dafle, A. Rodriguez, R. Paolini, B. Tang, S. S. Srinivasa, M. Erdmann, M. T. Mason, I. Lundberg, H. Staab, and T. Fuhlbrigge, "Extrinsic dexterity: In-hand manipulation with external forces," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 1578–1585.

[6] M. W. Marzke, "Tool making, hand morphology and fossil hominins," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 368, no. 1630, p. 20120414, 2013.

[7] M. M. W, "Precision grips, hand morphology, and tools," *American Journal of Physical Anthropology*, vol. 102, no. 1, pp. 91–110, 1997.

[8] A. Bicchi, "Hands for dexterous manipulation and robust grasping: A difficult road toward simplicity," *IEEE Transactions on robotics and automation*, vol. 16, no. 6, pp. 652–662, 2000.

[9] M. Qin, J. Brawer, and B. Scassellati, "Robot tool use: A survey," *Frontiers in Robotics and AI*, vol. 9, p. 1009488, 2023.

[10] R. Holladay, T. Lozano-Pérez, and A. Rodriguez, "Force-and-motion constrained planning for tool use," in *2019 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2019, pp. 7409–7416.

[11] C. C. Kemp, A. Edsinger, and E. Torres-Jara, "Challenges for robot manipulation in human environments [grand challenges of robotics]," *IEEE Robotics & Automation Magazine*, vol. 14, no. 1, pp. 20–29, 2007.

[12] A. Agarwal, S. Uppal, K. Shaw, and D. Pathak, "Dexterous functional grasping," *arXiv preprint arXiv:2312.02975*, 2023.

[13] A. Z. Ren, B. Govil, T.-Y. Yang, K. R. Narasimhan, and A. Majumdar, "Leveraging language for accelerated learning of tool manipulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 1531–1541.

[14] O. M. Andrychowicz, B. Baker, M. Chociej, R. Jozefowicz, B. McGrew, J. Pachocki, A. Petron, M. Plappert, G. Powell, A. Ray, *et al.*, "Learning dexterous in-hand manipulation," *The International Journal of Robotics Research*, vol. 39, no. 1, pp. 3–20, 2020.

[15] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, *et al.*, "Isaac gym: High performance gpu-based physics simulation for robot learning," *arXiv preprint arXiv:2108.10470*, 2021.

[16] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 5026–5033.

[17] S.-M. Yang, M. Magnusson, J. A. Stork, and T. Stoyanov, "Learning extrinsic dexterity with parameterized manipulation primitives," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 5404–5410.

[18] N. Chavan-Dafle and A. Rodriguez, "Prehensile pushing: In-hand manipulation with push-primitives," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 6215–6222.

[19] M. Oller, D. Berenson, and N. Fazeli, "Tactile-driven non-prehensile object manipulation via extrinsic contact mode control." in *Robotics: Science and Systems*, 2024.

[20] W. Zhou and D. Held, "Learning to grasp the ungraspable with emergent extrinsic dexterity," 2022. [Online]. Available: https://arxiv.org/abs/2211.01500

[21] H.-S. Fang, M. Gou, C. Wang, and C. Lu, "Robust grasping across diverse sensor qualities: The graspnet-1billion dataset," *The International Journal of Robotics Research*, vol. 42, no. 12, pp. 1094–1103, 2023.

[22] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 444–11 453.

[23] H. Chen, C. Zhu, Y. Li, and K. Driggs-Campbell, "Tool-as-interface: Learning robot policies from human tool usage through imitation learning," *arXiv e-prints*, pp. arXiv–2504, 2025.

[24] 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.

[25] A. Orsula, M. Geist, M. Olivares-Mendez, and C. Martinez, "Learning tool-aware adaptive compliant control for autonomous regolith excavation," *arXiv preprint arXiv:2509.05475*, 2025.

[26] M. Xu, P. Huang, W. Yu, S. Liu, X. Zhang, Y. Niu, T. Zhang, F. Xia, J. Tan, and D. Zhao, "Creative robot tool use with large language models," *arXiv preprint arXiv:2310.13065*, 2023.

[27] ATI Industrial Automation, "Robotic tool changers," https://www.ati-ia.com/products/toolchanger/robot_tool_changer.aspx.

[28] SCHUNK, "SWS quick change system," https://schunk.com/de/en/automation-technology/tool-changer/sws/c/PGR_1135.

[29] B. Dhakal, "Design of automatic tool changer for universal robots UR5," Master's thesis, Tampere University of Applied Sciences, 2019.

[30] D. Cheong, H. Park, and N. Kim, "Design and maneuver of a tool-changer using switchable magnet for a tool hung by a cable," in *IEEE International Conference on Automation Science and Engineering (CASE)*, 2024, pp. 1252–1257.

[31] H. Song, J. Hur, and S. Jeong, "Coaxial magnetic gear-based tool-changing system," *IEEE Access*, vol. 12, pp. 33 749–33 756, 2024.

[32] D. Mourtzis, J. Angelopoulos, M. Papadokostakis, and N. Panopoulos, "Design for 3D printing of a robotic arm tool changer under the framework of Industry 5.0," in *Procedia CIRP*, vol. 115, 2022, pp. 178–183.

[33] M. T. Mason, "Toward robotic manipulation," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 1, no. 1, pp. 1–28, 2018.

[34] M. Ghallab, D. Nau, and P. Traverso, *Automated Planning: theory and practice*. Elsevier, 2004.

[35] C. R. Garrett, R. Chitnis, R. Holladay, B. Kim, T. Silver, L. P. Kaelbling, and T. Lozano-Pérez, "Integrated task and motion planning," *Annual review of control, robotics, and autonomous systems*, vol. 4, no. 1, pp. 265–293, 2021.

[36] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," *arXiv preprint arXiv:2204.01691*, 2022.

[37] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar, *et al.*, "Inner monologue: Embodied reasoning through planning with language models," *arXiv preprint arXiv:2207.05608*, 2022.

[38] B. Liu, Y. Jiang, X. Zhang, Q. Liu, S. Zhang, J. Biswas, and P. Stone, "Llm+ p: Empowering large language models with optimal planning proficiency," *arXiv preprint arXiv:2304.11477*, 2023.

[39] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," in *2023 IEEE International conference on robotics and automation (ICRA)*. IEEE, 2023, pp. 9493–9500.

[40] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei, "Voxposer: Composable 3d value maps for robotic manipulation with language models," *arXiv preprint arXiv:2307.05973*, 2023.

[41] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Conference on Robot Learning*. PMLR, 2023, pp. 2165–2183.

[42] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.

[43] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, vol. 44, no. 10-11, pp. 1684–1704, 2025.

[44] C. Chi, Z. Xu, C. Pan, *et al.*, "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robot data," in *Robotics: Science and Systems (RSS)*, 2024.

[45] J. Kulozik and N. Jarrassé, "Evaluating the precision of the htc vive ultimate tracker with robotic and human movements under varied environmental conditions," *arXiv preprint arXiv:2409.01947*, 2024.

[46] A. Singh, A. Fry, A. Perelman, *et al.*, "Openai gpt-5 system card," 2025. [Online]. Available: https://arxiv.org/abs/2601.03267

[47] D. Coleman, I. A. Sucan, S. Chitta, and N. Correll, "Reducing the barrier to entry of complex robotic software: a MoveIt! case study," *Journal of Software Engineering for Robotics*, vol. 5, no. 1, pp. 3–16, 2014.

[48] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, *et al.*, "Sam 3: Segment anything with concepts," *arXiv preprint arXiv:2511.16719*, 2025.