

# Hybrid-Diffusion Models: Combining Open-loop Routines with Visuomotor Diffusion Policies

Jonne Van Haastregt<sup>\*1</sup>, Bastian Orthmann<sup>\*1</sup>, Michael C. Welle<sup>1,2</sup>, Yuchong Zhang<sup>2</sup>, Danica Kragic<sup>2</sup>

**Abstract**—Despite the fact that visuomotor-based policies obtained via imitation learning demonstrate good performances in complex manipulation tasks, they usually struggle to achieve the same accuracy and speed as traditional control based methods. In this work, we introduce *Hybrid-Diffusion* models that combine open-loop routines with visuomotor diffusion policies. We develop Teleoperation Augmentation Primitives (TAPs) that allow the operator to perform predefined routines, such as locking specific axes, moving to perching waypoints, or triggering task-specific routines seamlessly during demonstrations. Our Hybrid-Diffusion method learns to trigger such TAPs during inference. We validate the method on challenging real-world tasks: Vial Aspiration, Open-Container Liquid Transfer, and container unscrewing. All experimental videos are available on the project’s website: <https://hybriddiffusion.github.io/>

## I. INTRODUCTION

Advances in Imitation Learning [1]–[4] have propelled autonomous manipulation capabilities to tackling complex tasks such as spreading sauce on a pizza [1], opening a capped bottle [5], inserting a hanger into a T-shirt [3], and mounting a gear on a bike [4]. Performance is heavily dependent on demonstration quality [4], [6] and thus on the combination of teleoperation interface and operator skill. To democratize the collection of expert demonstrations, there has been significant work on teleoperation systems, ranging from a simple 6D space mouse [1], to virtual reality/augmented reality (VR/AR)-powered approaches [7], [8], hardware puppeteer approaches [2], [9], [10], and whole-body setups [11], [12].

One fundamental limitation of teleoperation systems is the morphological mismatch between the teleoperator and the robot. For example, a robot such as Atlas [13] has a joint that allows it to rotate its torso by 180°, an action a human body cannot perform with the same ease — so a motion-captured teleoperator cannot map such a motion 1:1 to the robot, thus not fully using the morphological advantages of the robot embodiment. In other cases, the teleoperation setup might be ideal for free-space 6D motion, but if the task requires continuous rotation around an axis (e.g., unscrewing a bottle), humans are not have the functionality of robotic manipulators that can simply actuate the last joint to obtain yaw rotation in the end-effector (tool) frame. Hence, robots’ morphological advantages can be mapped and integrated into a teleoperation system to allow the operator to fully leverage the particular embodiment.

<sup>\*</sup>These authors contributed equally (listed in alphabetical order).

<sup>1</sup>INCAR Robotics AB, Sweden

<sup>2</sup>KTH Royal Institute of Technology, Sweden

## Expert demonstrations

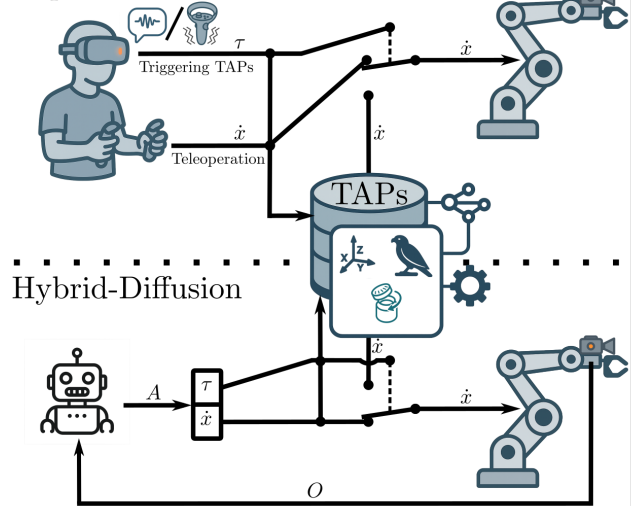


Fig. 1. Overview of our Hybrid-Diffusion Model; during teleoperation, the expert can trigger a Teleoperation Augmentation Primitive (TAP) either via speech or (AR) controller inputs. The Hybrid-Diffusion model learns to also trigger such TAP routines during execution, making use of routines during tasks.

In this work, we introduce *Hybrid Diffusion Models* and a principled teleoperation structure to obtain relevant training data. We present a teleoperation system that allows the operator to perform predefined (sometimes task-specific) routines with different interface modalities such as AR-enhanced controller inputs or voice commands. We introduce three types of *Teleoperation Augmentation Primitives (TAPs)*:

i) *Axis locking* — allows the operator to lock one or several axes (X, Y, Z, roll, pitch, yaw) in either the tool or base frame of the robot, keeping selected axes fixed regardless of teleoperation inaccuracies. This is particularly useful when a certain end-effector rotation must be maintained over parts of the task or when straight motion along a principal axis is desired.

ii) *Perching-waypoints* — allows the operator to predefine a set of waypoints that cover particular stages of (long-horizon) tasks. The operator can trigger a preset waypoint at any time, commanding the robot to move to it. This is particularly useful when only end-effector cameras are used and relevant objects may leave the field of view during part of the task — the triggered perching-waypoint brings the desired scene back into view and offers a unified starting position for the next step of the task.

iii) *(Open-loop) Routines* — allows the operator to trigger

a predefined (parameterized) (open-loop) primitive at a chosen point in the task that exploits morphological advantages of the robot. This is particularly useful when the robot has a kinematic advantage over the human, e.g., unscrewing a bottle, where a primitive can comprise: close gripper → rotate counter-clockwise → open gripper → rotate clockwise → close gripper → rotate counter-clockwise → move straight up (lifting the lid). The number of rotations, gripper force, etc., can be set via parameters. Note that while the routines presented in this work are open-loop, the framework is not limited to these and triggering routines that have closed-loop feedback or sub policies powered by reinforcement learning or similar is also possible.

We include operator-triggered TAP events in the training data of the baseline visuomotor diffusion models and extend the model [1] with the ability to trigger a TAP autonomously during inference—yielding *Hybrid Diffusion Models* that learn to invoke open-loop routines as part of policy execution, as shown in Fig. 1. We validate the usefulness of TAPs on three real-world tasks: Vial Aspiration, Open-Container Liquid Transfer, and container unscrewing. We compare against a baseline that does not explicitly trigger TAPs but instead continues to diffuse the recorded actions as if they were perfectly teleoperated. Note that not all TAPs are necessarily beneficial for policy rollout, but they can facilitate expert demonstrations in the first place. In summary, we see that the use of TAPs for Vial Aspiration, and open-container liquid transfer does not degrade or improve the policy performances, while the performances increase substantially for the container unscrewing task.

## II. RELATED WORK

### A. Teleoperation setups

Robot teleoperation using extended reality (XR) has been extensively studied over the past decades. In immersive VR, Galarza et al. [14] presented a teleoperation system that immerses users in a fully virtual environment to learn the operation and control of a KUKA youBot. Moletta et al. [7], [15] introduced a VR framework for cloth-folding manipulation that maps VR controller input to a physical robot rather than a simulated one. For AR, or mixed reality (MR)-based teleoperation, Arboleda et al. [16] designed some AR visual cues [17] that enrich the operator’s view with hints about the end-effector’s pose relative to the target, aiming to improve depth perception and task performance; they evaluated these designs against a baseline in pick-and-place tasks. In 2024, a new paradigm in a ‘puppeteer’ metaphor was introduced: Van Haastregt et al. [8] demonstrated the first system that lets operators control a physical robot by completely manipulating its virtual counterpart in AR using headset controllers. Building on this, Zhang et al. [18], [19] upgraded the approach by eliminating handheld controllers and incorporating LLM-driven voice commands and hand-gesture interaction with the virtual robot in AR, to improve the efficiency of teleoperating the physical robot.

Anything we can say how our work relates to this?

### B. Imitation Learning for Robot Manipulation

Early imitation learning, establishing the supervised behavioral cloning (BC) principle [20], mapped observations directly to actions, but suffered heavily when the observation distribution shifted. The reduction-based view of DAgger [21] addressed this by iteratively aggregating expert corrections under the learner-induced state distribution, becoming a standard method for data-efficient robot learning.

With stronger perception, visual Imitation learning, specifically Diffusion Policy, recast visuomotor BC as conditional denoising over action sequences, improving stability and handling multimodality [1]; its formulation now underpins many IL systems [22]–[24]. Recently, Large Behavior Models [4], multi-task diffusion-style policies, where introduced together with a rigorous evaluation pipeline and highlighting data/architecture factors for real-world performance at scale.

A complementary thread introduces hybrid action spaces to shorten horizons while retaining precision. HYDRA [25] alternates between sparse waypoints and dense per-timestep control, with relabeling to improve consistency, showing promising performances on real long-horizon tasks. Subsequent work integrates hybridization with richer perception: in [26] the authors present a method that learns salient 3D task points to anchor waypoint prediction from point clouds and switches to wrist-image dense control near contact. Hybrid policies also extend beyond fixed arms: HOMER [27] combines hybrid IL with a whole-body controller that maps end-effector objectives to coordinated base-arm motion, demonstrating imitation based mobile manipulation in a home scenario. While the hybrid concept has been explored before, our work introduces a setup where an operator can integrate the hybrid routines directly into the demonstration as well as a policy framework that learns to trigger these when needed for the task.

## III. METHODOLOGY

In this section, we first describe our teleoperation setup to facilitate Teleoperation Augmentation Primitives (TAPs) and subsequently lay out how the Visuomotor Diffusion policy can trigger TAPs during task execution, enabling Hybrid-Diffusion Models.

### A. Teleoperation Augmentation Primitives

Teleoperation Augmentation Primitives are primitives that a teleoperator can trigger when deemed beneficial for improving expert task demonstrations. Our TAPs aim to address specific shortcomings encountered in imitation learning teleoperation expert during demonstration collection.

**Types of TAPs:** We differentiate between three different types of TAPs and exemplify these with a particular task where we deemed them particularly useful as shown in Fig. 3 - note that TAPs can also be combined if the task demands.

*Axis locking:* A common setup for robotic teleoperation is to mimic the 6D pose of the teleoperator’s hand (if by tracking a VR controller, using Mocap, or operating a Puppeteer setup) to the robot’s 6D end effector pose. Operating a robot in such a setup in a precise manner is difficult, and

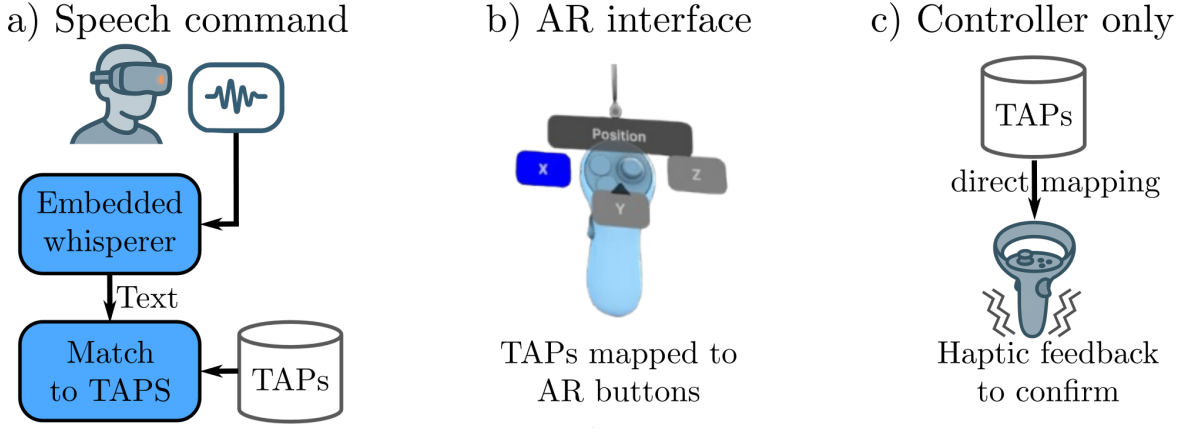


Fig. 2. Different ways of triggering TAPs; Via speech command (a), AR button interfaces(b), or via direct mapping with haptic pattern as confirmation (c -expert users)

performance results can vary widely between operators [28]. It has been shown that specifically the addition of rotation to 3D movement is difficult for human operators [29] and different ways to address such limitations in teleoperation settings [30], [31]. The Axis locking TAPs address this problem by providing the user the option to *lock* one or more of the axis - either in Tool- or the base-frame of the robot. For example, the user can first obtain a desired orientation of the end-effector, lock the rotational axis to avoid having to further spend cognitive energy on holding it in place while doing translational movements to complete the tasks. If one wants to ensure that the robot moves only up or downwards without any movement in X or Y, those axes can be locked to facilitate precise movements in unplugging scenarios. Similarly, all rotational axis can be locked once the end-effector is aligned with the target configuration allowing more precise insertion motion. The operator further has the choice to perform the locking either in the tool or the base frame of the robot, giving more flexibility to the TAP - note that in our experiments, we focus on axis locking in the base frame only.

**Perching-waypoints:** In Imitation learning settings, a combination of end-effector and stationary cameras is often deployed [4], [26], [32]. The end-effector cameras are useful for close up fine-grained observation of the manipulation performed - this egocentric view is often essential for task performance but can lead to failures if the task goal is no longer in view while performing the actions [5]. In this case, the stationary camera(s) become essential to the task performances and are therefore often used in more complex and longer horizon tasks to have a global view of the task progress. However, having stationary cameras means that the workspace for the robot has to be specifically equipped with these and that their viewpoint should be consistent for the trained policy to work, meaning any workstation has to have fixed cameras which is limiting regarding flexible robot deployment. Our perching-waypoint TAP addresses this by providing the user with the ability to set a number of predefined waypoints for a given task and trigger a return

to them whenever it feels beneficial for a task. A common scenario is that the location of an object in a task can be roughly known a-priori, such as "it's on the table". Here, the operator can set a waypoint that has a view of the entire table, giving the robot an overview of the whole scene. It is also useful to effectively *split* long horizontal tasks, as the operator can trigger the perching-waypoint once a subtask is completed  $\rightarrow$  the robot then will start the second part of the task at a defined position, splitting the task essentially in two shorter horizontal tasks that the policy learns to transition to by triggering the perching-waypoint.

**(Open-Loop) Routines:** While Human morphology is constant over most teleoperators, robot embodiment can be made with a specific task in mind or enhance the kinematic possibilities over their human counterparts [13]. However, a human is still constrained by their own kinematic limits, and therefore, a 1 : 1 mapping is often not possible. Often, a form of retargeting [12] is employed to bridge the gap between the human and robots morphological differences. Indeed, many teleoperation systems already naturally include such higher abstraction routines, for example, when teleoperating a quadruped robot, the operator often simply presses a 2D stick to indicate the desired movement directions  $\rightarrow$  the robot stack takes care of the precise and successful quadruped locomotion. Furthermore, tasks can have repetitive or hard-to-perform parts, such as screwing or unscrewing, where a static rotation around a single axis in toolframe is required - hard to perform in free motion 6D for operators but easy for a robot, as it often just has to move its last joint to achieve the desired motion. Another case can be rapid motions, such as flipping a garment, where the end effector flips upwards to achieve the desired result. Such a mapping of a more complex action to a single operator command is common place in video games where the playable character routinely performs for human impossible actions. Our (Open-Loop) routines are inspired by these thoughts and crafted for a particular robot embodiment. As a concrete example, we craft an *unscrewing routine* which is composed of close gripper  $\rightarrow$  rotate  $\rightarrow$  open gripper  $\rightarrow$  rotate back  $\rightarrow$  close

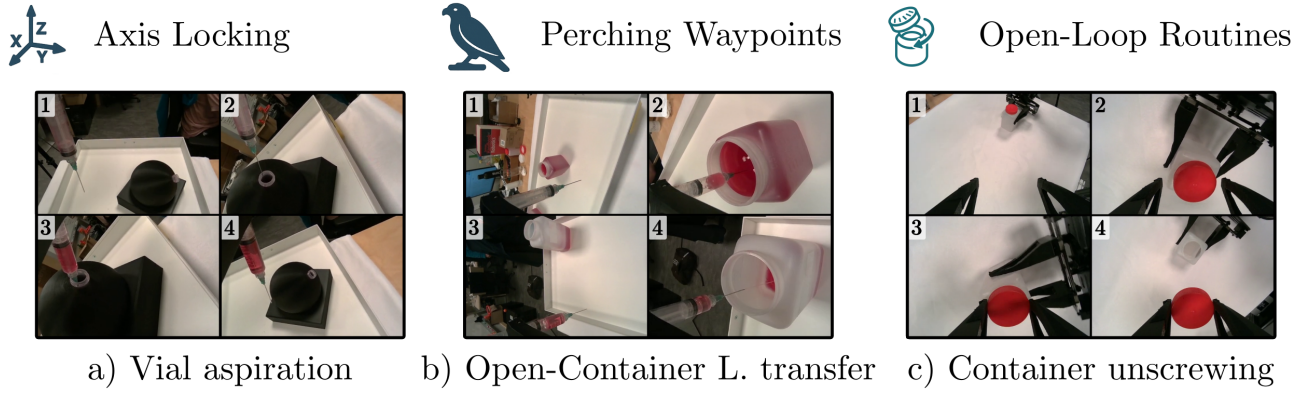


Fig. 3. Different types of TAPs and example tasks. a) shows vial aspiration where rotational axis looking is useful to the operator, b) Open-Container liquid transfer deploys perching waypoints in order to get the respective container in view, and c) - container unscrewing triggers an open-loop unscrewing routine when at the right places.

gripper  $\rightarrow$  rotate  $\rightarrow$  go up. We can parameterize the routine by determining how many times we want the rotation to execute, how far the final joint should rotate, etc. The Routines are embodiment-specific and can be invoked by the operator when it is deemed to be task beneficial. Note that while in this work we only showcase open-loop routines, the framework does support any kind of routines - those could be expressed as closed-loop control-based sub-policies or similar.

**Triggering TAPs during Teleoperation:** The operator needs a clear way to trigger the TAPs given the task at hand. Operators can have different preferences on how to interact with the teleoperation system [33]. We build upon [6]–[8] teleoperation frameworks, which use a VR headset (a Quest 3) to track the controller’s movements and relay it to a cartesian teleoperation controller. Furthermore, we added the functionality to vary the gains of how directly the robot tracks the teloperators’ input - a high gain can be useful for larger motions as the robot performs a multiple of the human’s input, while a lower gain is advantageous when doing precise manipulation steps, an approach very much similar to the sensitivity changes for computer mouse settings. To facilitate a wide range of operators, our TAPs can be triggered via speech or GUI buttons. Note that expert users are able to trigger the buttons without actually wearing the Headset using muscle memory and obtaining confirmation via haptic feedback (Fig. 2).

**Speech:** One way to trigger TAPs is via speech recognition (as shown in Fig. 2 a). The operator presses a predefined button combination on the controller to start the speech recognition process - this avoids background noise or unrelated conversation being misidentified as commands. Once the speech recognition is live, the VR headset is recording via the built-in microphone. Next, the operator specifies from a list of predefined commands such as "Lock X axis", "go to waypoint A", "Execute routine B". Once the command is spoken, the operator again presses the button to end the recording. We use Whisper-Tiny model [34] to perform the speech to text and then match the transcribed string

with the list of predefined TAPs, calculating the Levenshtein distances. The framework then executes the requested TAP. When the recording starts and stops, a short vibration is made from the controller to inform the operator that the recording is active/has ended, and a different pattern for when no valid command was parsed. For the dual arm case the speech recognition can be triggered from the corresponding controller, making it easy and intuitive to use.

**Augmented Reality:** If an operator does not want to use the speech interfaces, one can use the augmented reality controller menu visible through the AR headset. In this mode the operator can trigger a button combination to bring up the AR menu over the corresponding controller, as shown in Fig. 2 b) on the example for axis locking, the AR buttons - activated via the controller stick directions can be labeled with whatever TAPs are active for a given task, pressing the stick downwards switches to the next set of options.

**Haptic Feedback:** Advanced users are able to parse the augmented reality interfaces without looking at it as they do not wear the Headset over the eyes (which can be straining for longer recording sessions). In this case, the operator receives haptic feedback about which menu is active and which TAPs are available, as shown if Fig. 2 c). The haptic feedback is rendered via different vibrational pulses. This operation mode requires the most operator experience but also with the least restrictions, in practice, the complexity is drastically reduced by pre-selecting the TAPs needed for a given task and only making them available, rendering it possible to trigger a desired TAP with a single button press.

## B. Hybrid-Diffusion Models

The high-level Hybrid-Diffusion Model architecture is shown in Fig. 1, bottom part. We feed our observations, consisting of RGB-Images ( $O_t$ ) and the robot’s 6D end effector pose ( $x_t$ ), into the Visuomotor Diffusion policy block. The policy predicts the actions  $a_t$  to  $a_{t+n}$  as well as the TAPs for the next  $n$  timesteps. We execute as many timesteps as the policy takes to predict again (in our case that correlated to 3–4 steps at 10hz). If during this executions

a TAP is triggered by the policy, the Open-Loop behavior of the triggered TAP takes over, and further actions from the policy are ignored as long as the TAP is being executed. Note that the policy still continues predicting actions given the current observations - this way, once the TAP is finished, the policy takes back control seamlessly without any delay or interruption.

The TAP library has all predefined TAPs that the policy is able to trigger; the policy indicates which TAP should be executed at a given timestep. This setup makes the combination of (open-loop) routines with a visuomotor diffusion policy easily extendable. Note that the policy triggers two TAPs at the same time is an unwanted behavior as the operator doesn't have the capability to trigger two TAPs on the same timestep. In this case we execute the TAP that has a higher priority in the TAP library. We split the Hybrid-Diffusion inference algorithm into two parts; in Algorithm 1 the inference loop continuously predict the new actions  $a$  and TAPs as long as no inference is already running. The action and TAPs for the current timestep are then send to the Robot Controller handling the TAPs (Algorithm 2). Here, if no TAP is active and the receiving TAP is not triggered, the policy actions are executed normally. If a TAP is received, the TAP is executed in an open loop fashion until it's finished, and the policy (which has been silently predicting actions the entire time) is in charge again.

---

**Algorithm 1:** Hybrid-Diffusion via triggered TAPs

---

```

1 while true do
2   if no inference running ( $\sim$  every 3  $dt$ ) then
3      $\lfloor$  predict  $a_t, \dots, a_{t+n}$  and  $TAP_t, \dots, TAP_{t+n}$ ;
4     SendCommands( $a_t, TAP_t$ );
5      $t \leftarrow t + 1$ ;

```

---



---

**Algorithm 2:** Robot Controller handling TAPs

---

```

1 while receiving  $a_t, TAP_t$  do
2   if not  $TAP_{active} == none$  and not  $TAP_t == TAP_{Empty}$  then
3      $\lfloor TAP_{active} \leftarrow TAP_t$ 
4   if not  $TAP_{active} == none$  then
5     execute open-loop behavior of active TAP;
6     if  $TAP_{active}$  finished then
7        $\lfloor TAP_{active} \leftarrow none$ ;
8   else
9      $\lfloor$  ControlRobot( $a_t$ );
10   $t \leftarrow t + 1$ ;

```

---

Note that, as in this work, the TAPs are open loop and the policy can technically trigger at any time; therefore, there is a risk that a triggered TAP can lead to failure of the task and a risk to end outside the support of the training data.

## IV. EXPERIMENTS

We validate our Hybrid-Diffusion model on three distinct tasks that can naturally benefit from activating TAPs during rollouts.

**Experimental setup:** The setup for the three tasks is shown in Fig. 3. In detail, we consider: *i)* Vial Aspiration: The robot is equipped with a syringe and has to draw liquid out of a vial that is mounted on a sphere that enables arbitrary rotations (up to the point no liquid is flowing out of the vial). As the rotational alignment of the syringe needle with the vial is quite challenging the operator makes use of the *Axis locking* TAP - specifically to look all rotational axis once the desired rotation is achieved.

*ii)* Open-Container Liquid Transfer: The robot again is equipped with a syringe. The task is to draw the liquid substances from the right chemical container and deposit them into the left container. The TAP *Perching-waypoints* makes this task easier as it allows the operator to always go to an initial perching waypoint, draw the liquid, active another perching waypoint that brings the target container in clear view to deposit the liquid. *iii)* Container unscrewing task: In this task, a second robot arm (not controlled by the policy) holds a container with a threaded cap. The task is to bring the robot into position over the container, grasp the lid and unscrew it from the container and place it on the table. In this task, an open-loop routine TAP, specifically an unscrewing routine, is essential for task success. The operator simply brings the robot into position, ready to grasp the lid, and triggers the TAP to initiate the routine.

**Baseline comparison:** In order to see in which scenarios the policy itself benefits (not just the quality of the demonstrations improve through the use of TAPs), we compare our Hybrid-Diffusion policy with a baseline Diffusion policy that has no access to triggering TAPs - for both models we synthesize the action needed to execute the TAP routines. However, the baseline has no way of triggering the TAPs but has to learn the trajectories resulting from the TAPs as part of the overall policy. For all tasks we randomize the initial conditions seven times (with the same seed for baseline and Hybrid-diffusion) and execute each starting position three times, leading to 21 roll-outs per task per policy.

### A. Experimental results

We visualize the results in Fig. 4, where we show the results for each task using our hybrid diffusion (HD) compared against the baseline (D). We observe that the HD and baseline Vial Aspiration performances is very similar with 62% and 57% success rate, respectively. A similar picture is drawn when analyzing the Open-Container Liquid Transfer task where HD edges the baseline out with 71% vs. 62% - this results are expected as the Axis looking TAP is mostly helpful for the operator improving the overall quality of the demonstration, but a policy can learn to hold a orientation if the training data is of high quality. Similarly, the perching waypoints are essential for task success rate, but a policy can learn to go back to the vantage point without triggering a routine. We retain that the TAP makes this process more

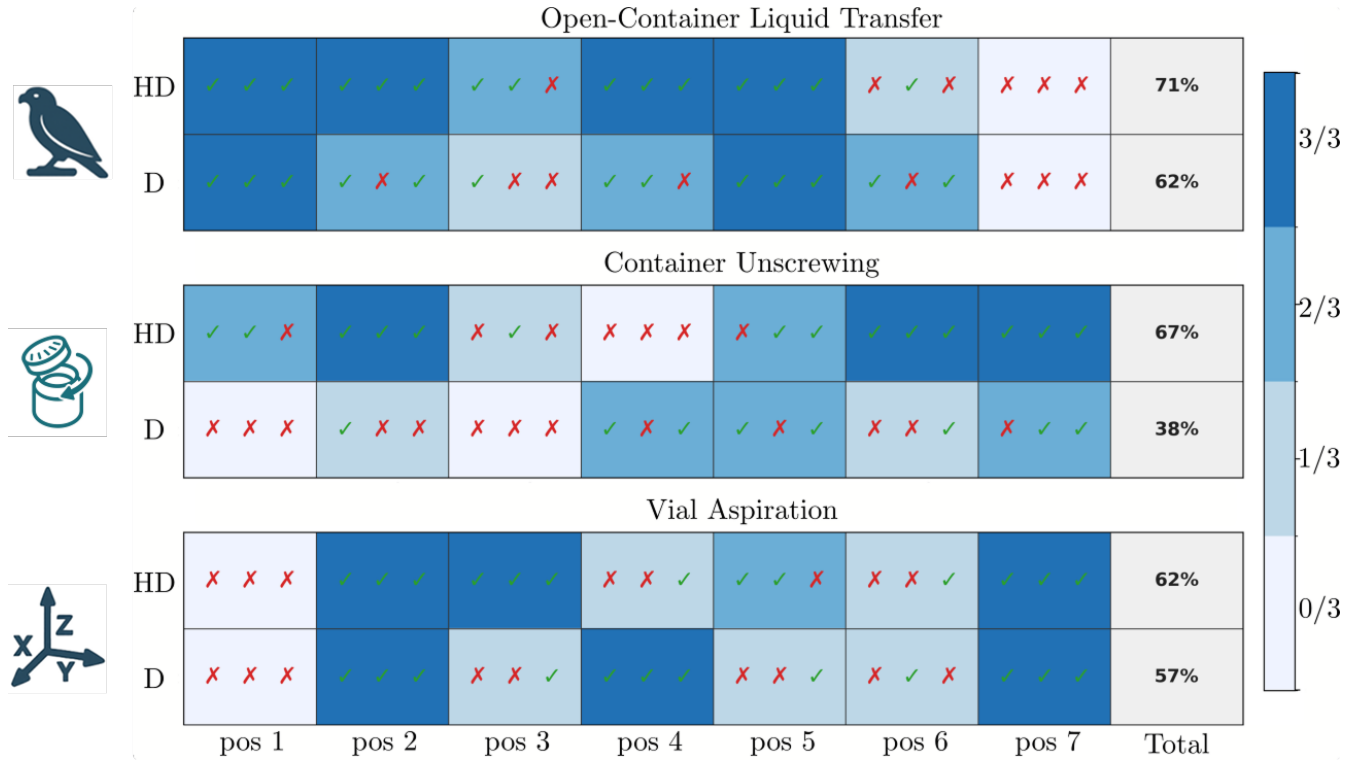


Fig. 4. Experimental results on three tasks, seven novel starting positions each repeated three times. Comparing Hybrid-Diffusion (HD) with the baseline diffusion (D) method.

structured and has the potential to help in longer horizon tasks, but the benefits are limited if the operator moves to perching waypoints anyhow without TAPs.

Finally, we can see a markable performance differences in the unscrewing tasks where the Hybrid-Diffusion Combining Open-loop Routines with Visuomotor Diffusion shines in comparison. Analyzing the failure cases reveals that the baseline struggles to do multiple turns in the unscrewing tasks and tries to lift the lid too early while it is still fixed to the bottle. This behavior is explained by the minimal differences in observations between a lid fully closed and the lid having been turned one time - even a human has trouble distinguishing these states, the policy has therefore a multimodal distribution in these cases to either turn or lift, leading to a drop in success rate (38%). The Hybrid method, on the other hand, manages to always remove the lid successfully if the routine was triggered at the right position, as the open loop makes sure to turn enough times to fully unscrew the bottle.

## V. CONCLUSION & FUTURE WORK

In this work, we introduced Hybrid-Diffusion Models, a framework that combines open-loop routines with visuomotor diffusion policies for robot manipulation. We presented Teleoperation Augmentation Primitives (TAPs), which allow operators to seamlessly trigger predefined routines during demonstrations via speech commands, AR interfaces, or haptic feedback. Our TAPs comprise three types: axis locking for constrained motion, perching-waypoints for consistent

viewpoint management, and open-loop routines that exploit morphological advantages of robot embodiments.

We highlighted that TAPs not only improve the quality and ease of collecting expert demonstrations but can also be learned and autonomously triggered by the policy during inference. Our experimental validation across three challenging real-world tasks, Vial Aspiration, Open-Container Liquid Transfer, and Container Unscrewing, revealed that the benefits of Hybrid-Diffusion vary depending on the task characteristics. While axis locking and perching-waypoints primarily facilitate demonstration collection without significantly impacting policy performance (62% vs. 57% and 71% vs. 62%, respectively), open-loop routines showed substantial improvements for tasks requiring repetitive, morphologically advantageous actions (67% vs. 38% for unscrewing).

**Future Work:** Several promising directions emerge from this work. First, extending TAPs beyond open-loop routines to include closed-loop feedback controllers or sub-policies learned via reinforcement learning could enhance robustness and adaptability. Second, investigating automatic TAP discovery, where the system identifies beneficial primitives from failed demonstrations or operator behavior patterns, could reduce manual engineering effort. Third, developing hierarchical Hybrid-Diffusion models that can compose multiple TAPs in sequence for longer-horizon tasks presents an exciting avenue for scaling to more complex manipulation scenarios. Finally, studying how TAPs transfer across different robot embodiments and tasks could provide insights into building more generalizable manipulation policies. We

believe that the principled integration of structured routines with learned visuomotor policies represents a promising path toward more capable and deployable robot systems.

## REFERENCES

- [1] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," in *Proceedings of Robotics: Science and Systems (RSS)*, 2023.
- [2] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," in *arXiv*, 2024.
- [3] T. Z. Zhao, J. Tompson, D. Driess, P. Florence, K. Ghasemipour, C. Finn, and A. Wahid, "Aloha unleashed: A simple recipe for robot dexterity," *arXiv preprint arXiv:2410.13126*, 2024.
- [4] J. Barreiros, A. Beaulieu, A. Bhat, R. Cory, E. Cousineau, H. Dai, C.-H. Fang, K. Hashimoto, M. Z. Irshad, M. Itkina, *et al.*, "A careful examination of large behavior models for multitask dexterous manipulation," *arXiv preprint arXiv:2507.05331*, 2025.
- [5] N. Ingelhart, J. Munkeby, J. van Haastregt, A. Varava, M. C. Welle, and D. Kragic, "A robotic skill learning system built upon diffusion policies and foundation models," *arXiv preprint arXiv:2403.16730*, 2024.
- [6] N. Ingelhart, J. Munkeby, M. C. Welle, M. Moletta, and D. Kragic, "Real-time operator takeover for visuomotor diffusion policy training," *arXiv preprint arXiv:2502.02308*, 2025.
- [7] M. C. Welle, N. Ingelhart, M. Lippi, M. Wozniak, A. Gasparri, and D. Kragic, "Quest2ros: An app to facilitate teleoperating robots," in *7th International Workshop on Virtual, Augmented, and Mixed-Reality for Human-Robot Interactions*, 2024.
- [8] J. van Haastregt, M. C. Welle, Y. Zhang, and D. Kragic, "Puppeteer your robot: Augmented reality leader-follower teleoperation," *arXiv preprint arXiv:2407.11741*, 2024.
- [9] H. Fang, H.-S. Fang, Y. Wang, J. Ren, J. Chen, R. Zhang, W. Wang, and C. Lu, "Airexo: Low-cost exoskeletons for learning whole-arm manipulation in the wild," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 15031–15038, IEEE, 2024.
- [10] Y. Jiang, R. Zhang, J. Wong, C. Wang, Y. Ze, H. Yin, C. Gokmen, S. Song, J. Wu, and L. Fei-Fei, "Behavior robot suite: Streamlining real-world whole-body manipulation for everyday household activities," *arXiv preprint arXiv:2503.05652*, 2025.
- [11] T. He, Z. Luo, X. He, W. Xiao, C. Zhang, W. Zhang, K. Kitani, C. Liu, and G. Shi, "OmniH2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning," *arXiv preprint arXiv:2406.08858*, 2024.
- [12] Y. Ze, Z. Chen, J. P. Arasjo, Z.-a. Cao, X. B. Peng, J. Wu, and C. K. Liu, "Twist: Teleoperated whole-body imitation system," *arXiv preprint arXiv:2505.02833*, 2025.
- [13] Boston Dynamics, "An electric new era for atlas." Online. Available: <https://bostondynamics.com/blog/electric-new-era-for-atlas/>, Apr. 2024. Accessed: Sep. 15, 2025.
- [14] B. R. Galarza, P. Ayala, S. Manzano, and M. V. Garcia, "Virtual reality teleoperation system for mobile robot manipulation," *Robotics*, vol. 12, no. 6, p. 163, 2023.
- [15] M. Moletta, M. K. Wozniak, M. C. Welle, and D. Kragic, "A virtual reality framework for human-robot collaboration in cloth folding," in *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*, pp. 1–7, IEEE, 2023.
- [16] S. Arevalo Arboleda, F. Rcker, T. Dierks, and J. Gerken, "Assisting manipulation and grasping in robot teleoperation with augmented reality visual cues," in *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1–14, 2021.
- [17] Y. Zhang, A. Nowak, A. Romanowski, and M. Fjeld, "An initial exploration of visual cues in head-mounted display augmented reality for book searching," in *Proceedings of the 21st International Conference on Mobile and Ubiquitous Multimedia*, pp. 273–275, 2022.
- [18] Y. Zhang, B. Orthmann, M. C. Welle, J. Van Haastregt, and D. Kragic, "Llm-driven augmented reality puppeteer: Controller-free voice-commanded robot teleoperation," in *International Conference on Human-Computer Interaction*, pp. 97–112, Springer, 2025.
- [19] Y. Zhang, B. Orthmann, S. Ji, M. Welle, J. Van Haastregt, and D. Kragic, "Multimodal" puppeteer": An exploration of robot teleoperation via virtual counterpart with llm-driven voice and gesture interaction in augmented reality," *arXiv preprint arXiv:2506.13189*, 2025.
- [20] D. A. Pomerleau, "Alvin: An autonomous land vehicle in a neural network," *Advances in neural information processing systems*, vol. 1, 1988.
- [21] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635, JMLR Workshop and Conference Proceedings, 2011.
- [22] Z. Xian, N. Gkanatsios, T. Gervet, T.-W. Ke, and K. Fragkiadaki, "Chaineddiffuser: Unifying trajectory diffusion and keypose prediction for robotic manipulation," in *7th Annual Conference on Robot Learning*, 2023.
- [23] Q. Yang, M. C. Welle, D. Kragic, and O. Andersson, "S<sup>2</sup>-diffusion: Generalizing from instance-level to category-level skills in robot manipulation," *arXiv preprint arXiv:2502.09389*, 2025.
- [24] G. Yan, Y.-H. Wu, and X. Wang, "Dnact: Diffusion guided multi-task 3d policy learning," *arXiv preprint arXiv:2403.04115*, 2024.
- [25] S. Belkale, Y. Cui, and D. Sadigh, "Hydra: Hybrid robot actions for imitation learning," in *Conference on Robot Learning*, pp. 2113–2133, PMLR, 2023.
- [26] P. Sundaresan, H. Hu, Q. Vuong, J. Bohg, and D. Sadigh, "What's the move? hybrid imitation learning via salient points," *arXiv preprint arXiv:2412.05426*, 2024.
- [27] P. Sundaresan, R. Malhotra, P. Miao, J. Yang, J. Wu, H. Hu, R. Antonova, F. Engelmann, D. Sadigh, and J. Bohg, "Homer: Learning in-the-wild mobile manipulation via hybrid imitation and whole-body control," *arXiv preprint arXiv:2506.01185*, 2025.
- [28] H. Tugal, I. Tugal, F. Abe, M. Sakamoto, S. Shirai, I. Caliskanelli, and R. Skilton, "Operator expertise in bilateral teleoperation: Performance, manipulation, and gaze metrics," *Electronics*, vol. 14, no. 10, p. 1923, 2025.
- [29] S. Zhai and P. Milgram, "Quantifying coordination in multiple dof movement and its application to evaluating 6 dof input devices," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 320–327, 1998.
- [30] X. Wang, S. Ong, and A. Y.-C. Nee, "Multi-modal augmented-reality assembly guidance based on bare-hand interface," *Advanced Engineering Informatics*, vol. 30, no. 3, pp. 406–421, 2016.
- [31] X. Zhang, W. He, M. Billingham, Y. Qin, L. Yang, D. Liu, and Z. Wang, "Usability of visualizing position and orientation deviations for manual precise manipulation of objects in augmented reality," *Virtual Reality*, vol. 28, no. 3, p. 134, 2024.
- [32] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," *arXiv preprint arXiv:2403.12945*, 2024.
- [33] J. Chen, A. Moemeni, and P. Caleb-Solly, "Comparing a graphical user interface, hand gestures and controller in virtual reality for robot teleoperation," in *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 644–648, 2023.
- [34] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*, pp. 28492–28518, PMLR, 2023.