

VacuumVLA: Boosting VLA Capabilities via a Unified Suction and Gripping Tool for Complex Robotic Manipulation

Hui Zhou^{1,*}, Siyuan Huang^{2,*}, Minxing Li^{3,*}, Hao Zhang¹, Lue Fan³, Shaoshuai Shi⁴

Abstract—Vision-Language-Action (VLA) models have significantly advanced general-purpose robotic manipulation by harnessing large-scale pre-trained vision and language representations. Among existing approaches, a majority of current VLA systems employ parallel two-finger grippers as their default end-effectors. However, such grippers face inherent limitations in handling certain real-world tasks—such as wiping glass surfaces or opening drawers without handles—due to insufficient contact area or lack of adhesion.

To overcome these challenges, we present a low-cost, integrated hardware design that combines a mechanical two-finger gripper with a vacuum suction unit, enabling dual-mode manipulation within a single end-effector. Our system supports flexible switching or synergistic use of both modalities, expanding the range of feasible tasks. We validate the efficiency and practicality of our design within two state-of-the-art VLA frameworks: DexVLA and π_0 . Experimental results demonstrate that with the proposed hybrid end-effector, robots can successfully perform multiple complex tasks that are infeasible for conventional two-finger grippers alone. All hardware designs and controlling systems will be released.

I. INTRODUCTION

Thanks to advances in vision-language models [1], [24] and the accumulation of large-scale manipulation datasets [3], [4], embodied AI has made significant progress in recent years. By training on vast amounts of aligned visual and textual data, VLMs have acquired strong generalization capabilities and transferable representations, enabling robots to rapidly adapt to new tasks with little or even no task-specific supervision. At the same time, large-scale manipulation datasets provide rich examples of action demonstrations. Through imitation learning, robots can efficiently learn from these examples and master complex motor skills.

Building on recent progress, OpenVLA [5] unifies a pre-trained vision encoder, a pretrained large language model (LLM), and an action prediction module into a cohesive Vision-Language-Action (VLA) framework, enabling more intelligent and adaptable robotic manipulation. By leveraging the rich, general-purpose representations learned by the pretrained models, OpenVLA can generalize to novel tasks and achieve zero-shot task transfer through natural language instructions alone, without requiring task-specific fine-tuning. Further advancements are exemplified by π_0 [6] and $\pi_{0.5}$ [7], which incorporate more powerful VLM

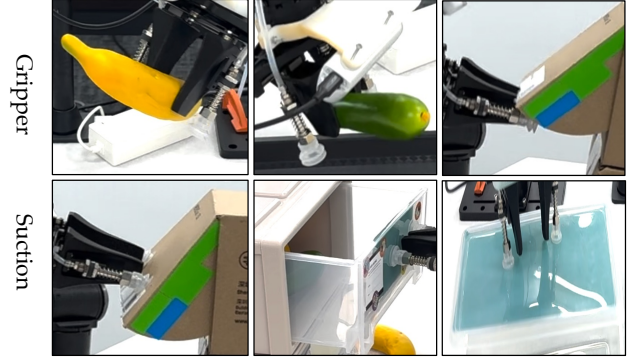


Fig. 1: Illustration of the our end-effector with integrated gripper and suction cup.

backbones and a flow-based action decoder, achieving robust and high-frequency control across a variety of robotic tasks.

However, effective complex manipulation requires more than semantic understanding and spatial reasoning—it is ultimately constrained by physical hardware, particularly the design and capabilities of the end effector. Current Vision-Language-Action (VLA) models predominantly rely on visual and linguistic inputs for task interpretation and planning, yet often overlook the end effector as a critical modality for real-world interaction. This sensory-motor asymmetry limits a robot’s ability to perform diverse and dexterous physical tasks, underscoring the need for co-design of intelligent control strategies and versatile hardware.

To address these limitations, we develop a low-cost end effector that integrates suction and gripping functionalities, as shown in Fig. 1, and complement it with a dedicated control and data collection system. To evaluate the effectiveness of our design, we conduct experiments using two distinct Vision-Language-Action (VLA) frameworks. The results demonstrate that our integrated end effector enables robots to successfully perform a range of household tasks that are previously unachievable with conventional grippers. Our key contributions are as follows:

- We develop a novel, low-cost end effector that combines suction and gripping capabilities. This hybrid design enables robots to perform challenging household tasks—such as opening handleless drawers—while maintaining strong performance in standard grasping operations.
- We establish and validate a comprehensive data acquisition and control system. To demonstrate its effectiveness, we design and execute four distinct tasks: clear-

*Equal contribution.

†Project leader.

¹The Chinese University of Hong Kong, Hong Kong SAR, China

²Shanghai Jiao Tong University, Shanghai, China

³Institute of Automation, Chinese Academy of Sciences, Beijing, China

⁴DiDi Global, China

ing a tabletop, opening a handleless plastic container, opening a handleless drawer, and opening a delivery cardboard box. We successfully validate our system using two open-source Vision-Language-Action (VLA) frameworks—DexVLA [8] and π_0 [6]—achieving consistent and promising results in both.

II. RELATED WORKS

A. End Effector Design

In grasping tasks, end-effector design is a crucial aspect that determines the upper limit of the system’s capability to handle objects. Among these, parallel grippers are a very common type of end-effector, which grasp objects using driven jaws that can open and close. This type of gripper can handle various object shapes and properties, and its grasping capability can be enhanced by integrating additional sensors, such as tactile [9], [37] and force [36] sensors. Multi-finger grippers and dexterous hands are also popular end-effectors. Previous studies have explored different designs for multi-finger grippers [38]–[40] and dexterous hands [27]–[29], with some work specifically dedicated to developing tailored algorithms [33]–[35] and datasets [30].

In contrast, vacuum suction grippers [10], [11] grasp objects by generating suction force through vacuum pressure, enabling them to accomplish tasks that conventional grippers cannot perform. However, their effectiveness is restricted to objects with flat and smooth surfaces that can be sealed by the suction cup, such as glass. Moreover, vacuum suction grippers struggle to grasp porous or cloth-like objects effectively, as they cannot generate sufficient suction force in such cases.

Beside single-function grippers, Multi-functional grippers [12]–[16] which have proven effective in various challenges have been introduced to overcome the limitations of single-function grippers. The MIT-Princeton team [15] took 1st place in the stowing task at the 2017 Amazon Robotics Challenge with their integrated suction-grasping hardware. Recently, the champion [16] of the 9th Robotic Grasping and Manipulation Competition (RGMC) held at ICRA 2024 also utilized an integrated suction-grasping hardware design.

B. Vision Language Action Models

Recent research on Vision-Language-Action (VLA) models [5]–[8], [17]–[22] has primarily focused on leveraging large-scale multimodal pretraining to achieve policy generalization across multiple instructions and diverse embodiments. These models can typically be decomposed into two modules: the first is a multimodal model responsible for encoding visual and linguistic information into tokens, and the second is an action expert that maps the encoded features directly to low-level control signals. The first module often incorporates reasoning mechanisms [7], [8] to improve instruction understanding. Currently, two common architectures are used for the action expert: flow-matching [32] based and diffusion [31] based. Flow-matching based architectures, such as π_0 [6], combine pretrained vision-language encoders with fast action decoders to enable high-frequency action

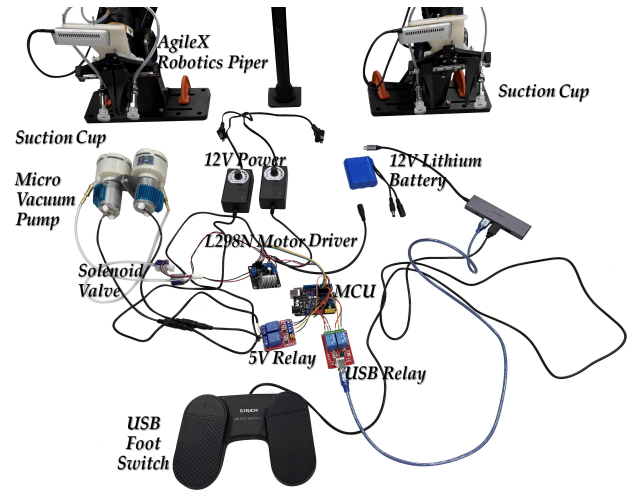


Fig. 2: Hardware Details.

outputs. Diffusion based models [8] support diverse and long-horizon behaviors through stochastic generation, but usually entail high training and inference costs. Despite significant progress, most existing VLA methods remain limited to visual and language inputs and are heavily constrained by hardware limitations.

III. METHODS

In this section, we first introduce a multifunctional gripper that combines suction and finger grasping capabilities, enabling it to handle a wide variety of objects in household tasks. Second, we present our vision-language-action model, which is primarily enhanced by incorporating a new dimension for the suction tool.

A. Limitations of Previous VLA End-effectors

As discussed in Section II-A, the design of end-effectors is one of the key issues in the field of robotics. The parallel gripper is the most popular choice for VLAs.

Parallel grippers are effective and easy to control, but due to their simple structure, they are unable to perform some relatively complex tasks. They are limited by object size, as they require sufficient space to form a stable grasp with their jaws. Furthermore, the grasping stability for certain objects cannot be guaranteed during manipulation, for example, a table tennis ball. Dexterous hands, on the other hand, are a comprehensive imitation of the human hand. Although capable of handling more complex tasks, they are more difficult to control due to the high degree-of-freedom (DoF). Furthermore, they still struggle with many household tasks, including opening a lid or drawer without a handle.

Table I lists some typical tasks that parallel grippers and dexterous hands struggle to perform.

B. Gripper Designs

Based on the above analysis, we propose an innovative hardware design to overcome the limitations. The goal of the first step for our end effector is to meet the requirements

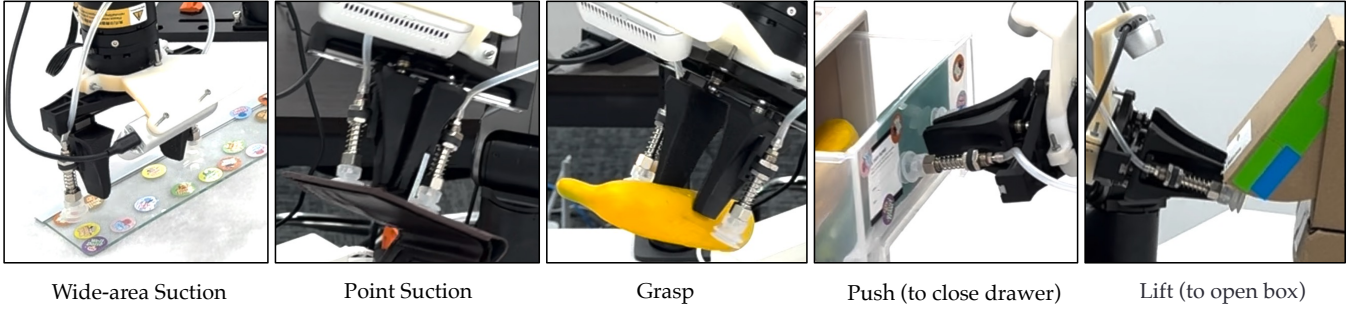


Fig. 3: Prime actions: Adjustable gripping width for varying object sizes (first two). The remaining three primitives are based on the standard two-finger gripper.

TABLE I: Typical examples of challenging tasks that popular end-effectors struggle to perform. These tasks can be accomplished by our hybrid end-effector.

ID	Description
1	Picking up an extremely thin piece of paper or glass
2	Holding a large object whose size exceeds the effector's max range
3	Opening a lid or drawer without a handle
4	Open a closed cardboard box

of common household tasks. To this end, we target a variety of different household objects: elongated glass items, banana props, cucumber props, wallets, sealed plastic containers, handleless drawers, and delivery cardboard boxes. We define prime actions that are complementary to each other in terms of utility across different object types and scenarios, and ensured successful execution through teleoperation, guaranteeing that at least one primitive can successfully complete the task.

1) *Hardware Details*: The overall hardware visualization is shown in Fig. 2. Each specific component and the system control is described as follows.

- **Two-Finger Gripper Base**: AgileX Robotics Piper [23]
- **Micro Vacuum Pump**:
 - Voltage: 12 V DC
 - Flow Rate: > 15.0 L/min
 - Vacuum Pressure: -60 kPa
 - Power: 12 W
- **MCU**: Arduino Uno R3
- **Solenoid Valve**: Generic 0520F 12 V DC
- **Suction Setting**: Silicone suction cup (15 mm diameter), metal connector (60 mm length)
- **Mounting Parts**: 3D printed
- **Others**: USB Relay, Silicone tube, 12V lithium battery, USB foot switch (for data collection), L298N Motor Driver

System control. We employ the USB protocol to control relays for generating distinct signal codes. Upon receiving a "turn-on" command from the computer via the relay, the MCU activates the L298N driver chip through GPIO interfaces, closing the solenoid valve. This action isolates the silicone tube from the atmosphere, establishing an airtight state. Concurrently, the MCU triggers another relay via a

GPIO interface to switch on the vacuum pump, thereby initiating the suction operation. Conversely, upon receiving a "turn-off" command, the MCU uses GPIO interfaces to open the solenoid valve, connecting the silicone tube to the atmosphere to release pressure, and simultaneously deactivates the relay to turn off the vacuum pump. The final status of all system devices is subsequently transmitted back to the computer via the UART protocol.

2) *Prime Actions*: We discover that many household tasks can be decomposed into three prime actions: **Suction**, **Grasp** and **Move**.

Suction can be applied to various challenging tasks, especially those including large or handleless objects. Unlike other methods [14], [16], which only use one suction cup, each gripper is equipped with two suction cups driven by the gripper jaw. Thus, we can adjust the distance between the two suction cups to fit different objects. For example, for large glass slides, we can set the gripper to its maximum stroke and then proceed with suction (wide-area suction). For wallets of square shapes, we can set the gripper to its minimum stroke and then proceed with suction (point suction).

Grasp. For common objects with handles (including cucumbers and bananas), we use the normal two-finger gripping function to grasp them.

Move refers to a series of actions that move a certain part of an object, including *push*, *pull*, *lift* and *press*. For example, to close a drawer and lift the lid of a delivery box, the *push* function and the *lift* function of the gripper are utilized.

Our hybrid end-effector can successfully execute the aforementioned prime actions, and therefore accomplish household tasks by combining these actions.

C. Vision Language Action Model

VLA Formulation. For learning VLA (Vision-Language-Action) models, we employ a common dual 6-axis-arm manipulation hardware platform. Unlike other mobile manipulation setups, we use a fixed base. The viewpoint includes a fixed top-view camera and two wrist-mounted cameras, one on each arm. The robot's observation at timestep t consists of base and hand visual inputs V_t^b , V_t^{left} , and V_t^{right} , the state of each robotic arm $s_t \in \mathbf{R}^7$ (including gripper state), and the suction status $f \in \{\text{True}, \text{False}\}$.

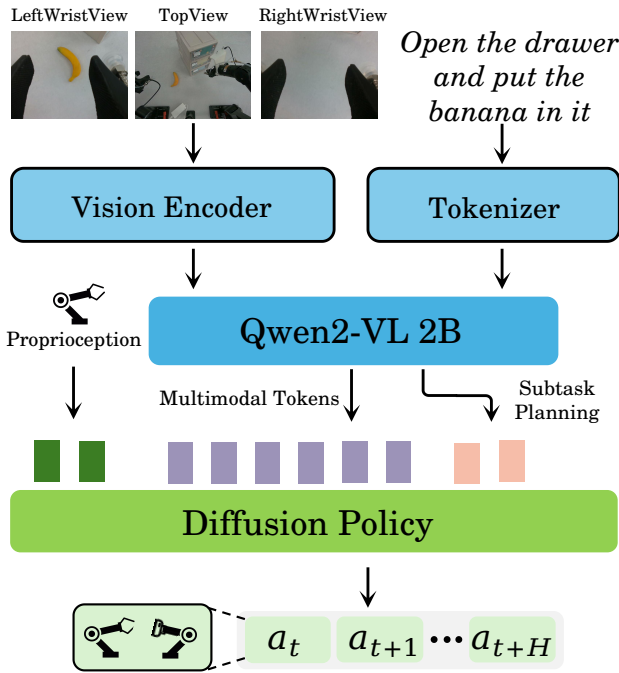


Fig. 4: VacuumVLA (based on DexVLA) architecture.

Given a language instruction L , the objective is to learn an end-to-end policy $\pi(A_t | V_t, L)$ that outputs a low-level, executable action chunk $A_t = \{a_t, a_{t+1}, \dots, a_{t+H}\}$, where $a_t \in \mathbf{R}^{16}$ is formed by concatenating the 6-degree-of-freedom joint vectors of the two robotic arms, the gripper opening width, and the suction status.

Shortcut learning in binary suction input. In existing VLA methods, the current robot state is often included as an observation input. This allows for continuous prediction based on the current state, ensuring that the predicted values within a future action chunk do not deviate too far from the current state. However, the suction status is not well-suited to this paradigm. Our suction status $f \in \{\text{True}, \text{False}\}$ is typically consistent between ground truth and observations across the majority of task learning steps. The status only changes during the specific action chunk when the suction is turned on or off. Such transitions constitute a relatively small proportion of all action chunks in the dataset, making the model prone to the “shortcut” problem (i.e., simply copying the input state). Therefore, in our VLA design, the input is similar to previous VLA approaches, but the output is extended by two dimensions, corresponding to the suction status of the left and right arms, respectively.

VacuumVLA is an end-to-end multimodal robotic policy specifically designed for a suction-gripper hybrid end-effector. To evaluate its effectiveness, we conduct experiments using two distinct state-of-the-art frameworks: π_0 [6] and DexVLA [8].

When built upon the π_0 framework, VacuumVLA integrates visual inputs, natural language instructions, and the robot’s proprioceptive state to generate a continuous distribution over actions—including suction and grasping—

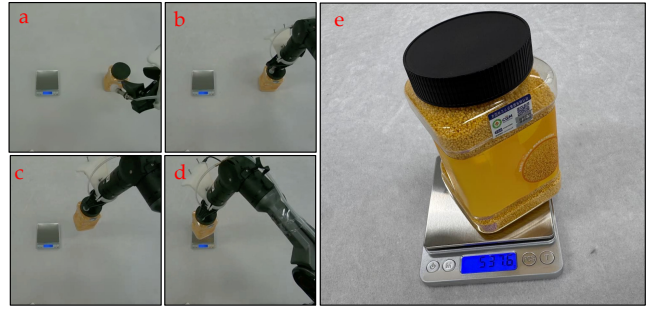


Fig. 5: Weight suction test, where figures (a)–(e) correspond to five moments of the test, respectively, and the last image shows an enlarged view of the item placed on the electronic scale.

using a conditional flow matching model. In this setup, π_0 is initialized from PaliGemma [24]. For action generation, Flow Matching can produce highly precise and realistic outputs. It effectively captures the complex structure and fine details of data, generating motion sequences that are smooth, coherent, and physically plausible.

Alternatively, based on the DexVLA framework—as illustrated in Fig. 4—VacuumVLA adopts Qwen2-VL as the base vision-language model (VLM). The pretrained image encoder from Qwen2-VL [25] is applied to project the robot’s visual observations—comprising three concatenated images in our setup—into the shared embedding space with language tokens. Unlike π_0 , the VLM generates additional reasoning tokens for subtask planning. Consequently, the diffusion-based action expert can generate action chunks conditioned on three components: multimodal hidden states generated by the VLM, reasoning tokens for subtask planning, and current proprioceptive states. This design enables coherent and hierarchical control.

IV. EXPERIMENTS

This section consists of three main parts. The first part is hardware testing, including testing the stability of picking and placing a 500g object and analyzing suction force on different materials. The second part presents a success rate comparison between two versions of VacuumVLA, evaluated on the four predefined gripper primitive tasks. The third part provides visualizations of successful examples.

A. Hardware Function Test

After designing the hardware structure, we conduct two experiments to test the suction power of our hardware.

Weight suction test. The first is a 500g object lifting test in Fig. 5. We select a 537-gram jar of rice and test whether it can be successfully picked up from a desktop and placed onto an electronic scale.

Pressure for different materials. The second experiment involves pressure tests on different materials. We select glass, a leather wallet, and a cardboard box, with atmospheric pressure at 100 Percent. Through these tests as shown in Fig. 6, we observe that, under the same power setting, the

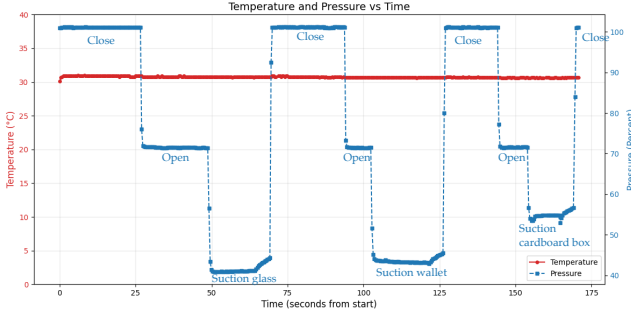


Fig. 6: Suction pressure tests for different materials (relative to ambient atmospheric pressure). The “Close” phase represents that the motor is turned off. The “Open” phase indicates that the motor is on but no object is being sucked. The “Suction” phase refers to the system’s response when the motor is on and actively sucking different target objects.



Fig. 7: Data collection based on homogeneous teleoperation hardware, where the suction cup is controlled by a foot-operated USB device.

suction capability on porous cardboard material is weaker, which aligns with the general phenomenon that suction cups perform poorly on porous surfaces. (Note: We assume that minor temperature fluctuations over a short period do not significantly affect the air pressure inside the tube.)

B. VacuumVLA

As described in the Methods section, we base our experiments on two state-of-the-art VLA frameworks: DexVLA and π_0 . Both DexVLA and π_0 use training datasets identical in size and diversity, collected through homogeneous teleoperation. The key difference is that DexVLA includes an additional annotation of subtask planning in the form of language instructions, which we elaborate on later.

Testing tasks. We define four long-horizon tasks to test our gripper:

- **Task1:** place the following objects into a tray: a 280 mm

× 80 mm glass slide, a banana prop, a cucumber prop, and a wallet.

- **Task2:** open a sealed plastic container, place either the banana or the wallet inside, and close the container.
- **Task3:** open a handleless drawer, place the cucumber inside, and close the drawer.
- **Task4:** open a delivery cardboard box.

All actions can be composed of the prime actions described in Section III-B.2, as shown in Fig. 3.

Data collection. For data collection with the robotic arm, we use homogeneous teleoperation. For suction cup data collection, we employ a foot-operated USB switch. When a trigger signal is detected, the system sends a command to turn on the vacuum pump; pressing the trigger again turns it off, as illustrated in Fig. 7.

The number of trajectories collected varies across different tasks: 200 for Task 1, and 100 each for Task 2, Task 3, and Task 4. During data collection, we switch between suction cup hands.

Since DexVLA requires subtask annotations, we have designed a set of subtask templates for the four tasks. Examples include:

- *Please use the right arm to suction the glass and place it into the brown dinner plate.*

Training details. For the π_0 model, the batch size is set to 16. However, since the open-source π_0 is based on JAX and only single-node training code is released, we trained the model for four days until 80,000 steps.

For the DexVLA model, we use the pre-trained one-stage Action Expert. To accelerate computation, we adopt the 400M-parameter variant `scale_dp_l1`¹. Unlike the original DexVLA, which trains in three stages, we train only Stage 2, as our task does not involve cross-embodiment generalization. We do not use LoRA during training. The batch size is 16 per GPU, and training runs on four A100 servers for two days with a constant learning rate of 2×10^{-5} .

Success rate. For the four long-horizon tasks, we evaluated the success rates of two versions of VacuumVLA, with each task tested 15 times. A trial is counted as successful only if all prime-actions within the task are successfully completed; otherwise, it is recorded as a failure. Additionally, we observe a typical failure case: if the end-effector of the robotic arm continuously oscillates at the same position for more than one minute, the trial is counted as a failure.

TABLE II: Success rates of VacuumVLAs. Traditional end-effectors got success rates of zero primarily due to their requirements for handles and proper object sizes (Section III-A for more details).

Model Base	End Effector	Task1	Task2	Task3	Task4
DexVLA	Gripping	0.0%	0.0%	0.0%	0.0%
DexVLA	Suction-Gripping	73.3%	80.0%	53.3%	33.3%
π_0	Suction-Gripping	53.3%	66.67%	60.0%	53.3%

From the success rate table above, it can be observed that traditional end-effectors fail to handle tasks such as grasping

¹https://huggingface.co/lesjiescale_dp_l1

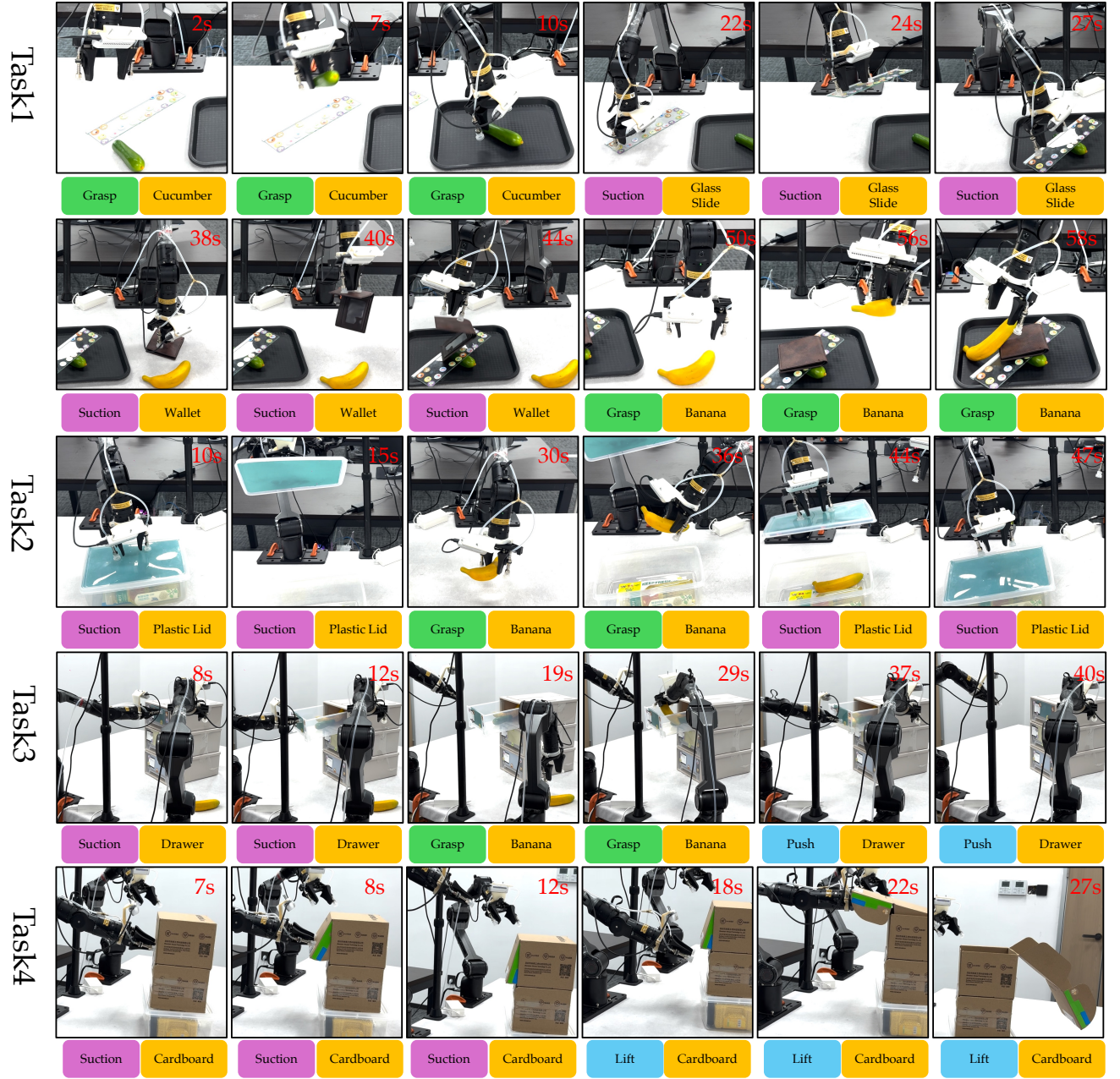


Fig. 8: A visualization of the temporal execution of prime-actions and object types across four tasks in VacuumVLA, where prime-action-grasp is shown in green, prime-action-suction in purple, prime-action-lift or push in blue, and object type in yellow.

a long glass, opening a handleless plastic container, opening a handleless drawer, and opening a delivery cardboard box. This failure is primarily due to their requirements for handles and proper object sizes, with detailed analysis presented in Section III-A. In contrast, our proposed VacuumVLA achieves reasonable results under two different state-of-the-art VLA frameworks. The detailed visualizations are shown in the following section.

As shown in Table II, we present two variants of VacuumVLA based on different model bases. On Task1 and Task2, VacuumVLA with DexVLA as the base model

achieves higher success rates than the version using π_0 as the base. However, the opposite trend is observed on Task3 and Task4. Specifically, for Task2 (putting the lid on the plastic box), the error offset is 5.3 cm for VacuumVLA (DexVLA) and 3.1 cm for VacuumVLA (π_0). Although the π_0 -based model performs worse in terms of success rate, it achieves higher precision in completing the capping task.

C. Visualizations

We separately visualize task1, task2, task3 and task4 of VacuumVLA (dexvla) in Fig. 8.²

V. LIMITATIONS

Although the hardware and end-to-end VLA algorithm we designed achieve certain effectiveness in tasks that conventional parallel grippers cannot complete, several issues remain:

- 1) The designed position of the suction cup may interfere with normal grasping in cluttered scenes, as the presence of the suction cup requires a larger distance between target objects.
- 2) The two versions of the VLA algorithm proposed in this paper do not account for whether a suction event is a true suction (successful attachment) or a false suction (correct positioning but misaligned suction cup). Since the visual features of true and false suction events are nearly identical, this can lead to the robotic arm performing ineffective motions.
- 3) The two suction cups on a single arm proposed in this paper adopt an underactuated design — that is, to simplify the suction end-effector, we use a single motor and a single solenoid valve to control both suction cups on one arm, which may result in one cup achieving proper suction while the other leaks air.

VI. CONCLUSION

This paper presents a multifunctional end-effector based on a parallel gripper, integrating both suction and grasping capabilities, enabling the completion of tasks that were previously impossible for a single parallel gripper, such as opening handle-less drawers and cardboard boxes. Furthermore, to validate the proposed multifunctional end-effector, we implement a hybrid suction-grasping VLA (VacuumVLA) based on two existing state-of-the-art VLA frameworks, demonstrating that end-to-end control of combined suction and grasping actions can be achieved through the VLA approach. For future work, we plan to focus on addressing the issues outlined in Section V. We anticipate that this work will offer a new perspective on enhancing VLAs' performance through innovative hardware designs.

REFERENCES

- [1] Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J. & Others Qwen2. 5-vl technical report. *ArXiv Preprint ArXiv:2502.13923*. (2025)
- [2] Beyer, L., Steiner, A., Pinto, A., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E. & Others Paligemma: A versatile 3b vlm for transfer. *ArXiv Preprint ArXiv:2407.07726*. (2024)
- [3] O'Neill, A., Rehman, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., Mandlekar, A., Jain, A. & Others Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. *2024 IEEE International Conference On Robotics And Automation (ICRA)*. pp. 6892-6903 (2024)
- [4] Khazatsky, A., Pertsch, K., Nair, S., Balakrishna, A., Dasari, S., Karamcheti, S., Nasiriany, S., Srirama, M., Chen, L., Ellis, K. & Others Droid: A large-scale in-the-wild robot manipulation dataset. *ArXiv Preprint ArXiv:2403.12945*. (2024)
- [5] Kim, M., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P. & Others Openvla: An open-source vision-language-action model. *ArXiv Preprint ArXiv:2406.09246*. (2024)
- [6] Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B. & Others Pi0: A Vision-Language-Action Flow Model for General Robot Control. *ArXiv Preprint ArXiv:2410.24164*. (2024)
- [7] Intelligence, P., Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N. & Others Pi0.5: a Vision-Language-Action Model with Open-World Generalization. *ArXiv Preprint ArXiv:2504.16054*. (2025)
- [8] Wen, J., Zhu, Y., Li, J., Tang, Z., Shen, C. & Feng, F. Dexvla: Vision-language model with plug-in diffusion expert for general robot control. *ArXiv Preprint ArXiv:2502.05855*. (2025)
- [9] Yuan, W., Dong, S. & Adelson, E. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*. **17**, 2762 (2017)
- [10] Eppner, C., Höfner, S., Jonschkowski, R., Martín-Martín, R., Sieverling, A., Wall, V. & Brock, O. Lessons from the amazon picking challenge: Four aspects of building robotic systems.. *Robotics: Science And Systems*. **12** (2016)
- [11] Schwarz, M. & Behnke, S. Data-efficient deep learning for RGB-D object perception in cluttered bin picking. *Warehouse Picking Automation Workshop (WPAW), IEEE International Conference On Robotics And Automation (ICRA)*. pp. 2-4 (2017)
- [12] D'Avella, S., Sundaram, A., Friedl, W., Tripicchio, P. & Roa, M. Multimodal grasp planner for hybrid grippers in cluttered scenes. *IEEE Robotics And Automation Letters*. **8**, 2030-2037 (2023)
- [13] Um, S., Jeong, H., Kim, C., Rhee, I. & Choi, H. Rec-gripper: A reconfigurable combined suction and fingered gripper for various logistics picking and stowing tasks. *IEEE Robotics And Automation Letters*. **9**, 87-94 (2023)
- [14] Zeng, A., Song, S., Yu, K., Donlon, E., Hogan, F., Bauza, M., Ma, D., Taylor, O., Liu, M., Romo, E. & Others Robotic Pick-and-Place of Novel Objects in Clutter with Multi-Affordance Grasping and Cross-Domain Image Matching. *2018 IEEE International Conference On Robotics And Automation (ICRA)*. pp. 3750-3757 (2018)
- [15] Zeng, A., Song, S., Yu, K., Donlon, E., Hogan, F., Bauza, M., Ma, D., Taylor, O., Liu, M., Romo, E. & Others Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. *The International Journal Of Robotics Research*. **41**, 690-705 (2022)
- [16] Son, Y., Um, S., Hong, J., Bui, T. & Choi, H. Corner-Grasp: Multi-Action Grasp Detection and Active Gripper Adaptation for Grasping in Cluttered Environments. *ArXiv Preprint ArXiv:2504.01861*. (2025)
- [17] Yu, J., Liu, H., Yu, Q., Ren, J., Hao, C., Ding, H., Huang, G., Huang, G., Song, Y., Cai, P. & Others ForceVLA: Enhancing VLA Models with a Force-aware MoE for Contact-rich Manipulation. *ArXiv Preprint ArXiv:2505.22159*. (2025)
- [18] Shukor, M., Aubakirova, D., Capuano, F., Kooijmans, P., Palma, S., Zouitine, A., Aractingi, M., Pascal, C., Russi, M., Marafioti, A. & Others Smolvla: A vision-language-action model for affordable and efficient robotics. *ArXiv Preprint ArXiv:2506.01844*. (2025)
- [19] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J. & Others Rt-1: Robotics transformer for real-world control at scale. *ArXiv Preprint ArXiv:2212.06817*. (2022)
- [20] Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohllhart, P., Welker, S., Wahid, A. & Others Rt-2: Vision-language-action models transfer web knowledge to robotic control. *Conference On Robot Learning*. pp. 2165-2183 (2023)
- [21] Cheang, C., Chen, G., Jing, Y., Kong, T., Li, H., Li, Y., Liu, Y., Wu, H., Xu, J., Yang, Y. & Others Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *ArXiv Preprint ArXiv:2410.06158*. (2024)
- [22] Liu, S., Wu, L., Li, B., Tan, H., Chen, H., Wang, Z., Xu, K., Su, H. & Zhu, J. RDT-1B: a Diffusion Foundation Model for Bimanual Manipulation. *The Thirteenth International Conference On Learning Representations*.

²The complete demo can be found at

- [23] Agilex Robotics. [Online]: <https://global.agilex.ai/products/piper>.
- [24] Beyer, L., Steiner, A., Pinto, A., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E. & Others Paligemma: A versatile 3b vlm for transfer. *ArXiv Preprint ArXiv:2407.07726*. (2024)
- [25] Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W. & Others Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *ArXiv Preprint ArXiv:2409.12191*. (2024)
- [26] Liu, H., Li, C., Wu, Q. & Lee, Y. Visual instruction tuning. *Advances In Neural Information Processing Systems*. **36** pp. 34892-34916 (2023)
- [27] Weng, Z. BiDexHand: Design and Evaluation of an Open-Source 16-DoF Biomimetic Dexterous Hand. *ArXiv Preprint ArXiv:2504.14712*. (2025)
- [28] Shaw, K., Agarwal, A. & Pathak, D. Leap hand: Low-cost, efficient, and anthropomorphic hand for robot learning. *ArXiv Preprint ArXiv:2309.06440*. (2023)
- [29] Romero, B., Fang, H., Agrawal, P. & Adelson, E. Eyesight hand: Design of a fully-actuated dexterous robot hand with integrated vision-based tactile sensors and compliant actuation. *2024 IEEE/RSJ International Conference On Intelligent Robots And Systems (IROS)*. pp. 1853-1860 (2024)
- [30] Liu, Y., Yang, Y., Wang, Y., Wu, X., Wang, J., Yao, Y., Schwertfeger, S., Yang, S., Wang, W., Yu, J. & Others Realdex: Towards human-like grasping for robotic dexterous hand. *ArXiv Preprint ArXiv:2402.13853*. (2024)
- [31] Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Advances In Neural Information Processing Systems*. **33** pp. 6840-6851 (2020)
- [32] Lipman, Y., Chen, R., Ben-Hamu, H., Nickel, M. & Le, M. Flow matching for generative modeling. *ArXiv Preprint ArXiv:2210.02747*. (2022)
- [33] Xu, Y., Wan, W., Zhang, J., Liu, H., Shan, Z., Shen, H., Wang, R., Geng, H., Weng, Y., Chen, J. & Others Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 4737-4746 (2023)
- [34] Wan, W., Geng, H., Liu, Y., Shan, Z., Yang, Y., Yi, L. & Wang, H. Unidexgrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. *Proceedings Of The IEEE/CVF International Conference On Computer Vision*. pp. 3891-3902 (2023)
- [35] Lan, F., Wang, S., Zhang, Y., Xu, H., Oseni, O., Zhang, Z., Gao, Y. & Zhang, T. Dexcatch: Learning to catch arbitrary objects with dexterous hands. *ArXiv Preprint ArXiv:2310.08809*. (2023)
- [36] Li, J., Zhu, K., Lu, G., Chen, I., Dong, H. & Others Construction of a Multiple-DOF Under-actuated Gripper with Force-Sensing via Deep Learning. *ArXiv Preprint ArXiv:2506.11570*. (2025)
- [37] Liu, S. & Adelson, E. Gelsight fin ray: Incorporating tactile sensing into a soft compliant robotic gripper. *2022 IEEE 5th International Conference On Soft Robotics (RoboSoft)*. pp. 925-931 (2022)
- [38] Shao, L., Ferreira, F., Jorda, M., Nambiar, V., Luo, J., Solowjow, E., Ojea, J., Khatib, O. & Bohg, J. Unigrasp: Learning a unified model to grasp with multifingered robotic hands. *IEEE Robotics And Automation Letters*. **5**, 2286-2293 (2020)
- [39] Burgess, M. & Adelson, E. Grasp Everything (GET): 1-DoF, 3-Fingered Gripper with Tactile Sensing for Robust Grasping. *ArXiv Preprint ArXiv:2505.09771*. (2025)
- [40] Cutler, E., Xing, Y., Cui, T., Zhou, B., Rijnsoever, K., Hart, B., Valencia, D., Ong, L., Gee, T., Liarokapis, M. & Others Benchmarking Reinforcement Learning Methods for Dexterous Robotic Manipulation with a Three-Fingered Gripper. *ArXiv Preprint ArXiv:2408.14747*. (2024)