Vision-Language-Action Model and Diffusion Policy Switching Enables Dexterous Control of an Anthropomorphic Hand

Cheng Pan[1], Kai Junge[1] & Josie Hughes[1]

https://vla-diffu-switch.github.io/

***Abstract*— To advance autonomous dexterous manipulation, we propose a hybrid control method that combines the relative advantages of a fine-tuned Vision-Language-Action (VLA) model and diffusion models. The VLA model provides language commanded high-level planning, which is highly generalizable, while the diffusion model handles low-level interactions which offers the precision and robustness required for specific objects and environments. By incorporating a switching signal into the training-data, we enable event based transitions between these two models for a pick-and-place task where the target object and placement location is commanded through language. This approach is deployed on our anthropomorphic ADAPT Hand 2, a 13DoF robotic hand, which incorporates compliance through series elastic actuation allowing for resilience for any interactions: showing the first use of a multi-fingered hand controlled with a VLA model. We demonstrate this model switching approach results in a over 80% success rate compared to under 40% when only using a VLA model, enabled by accurate near-object arm motion by the VLA model and a multi-modal grasping motion with error recovery abilities from the diffusion model.**

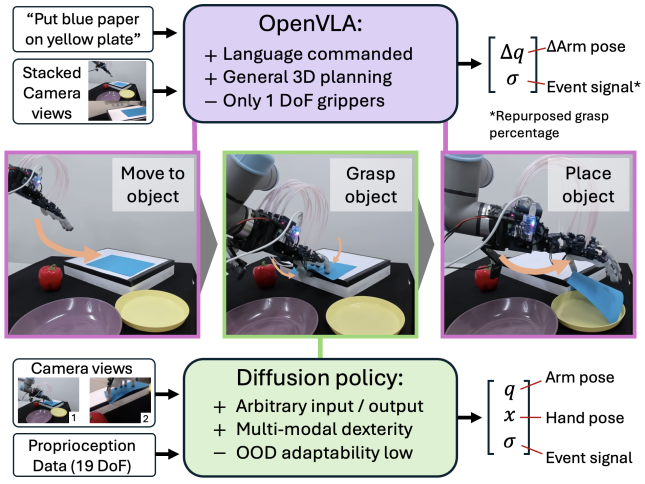

Fig. 1. Combined VLA and Diffusion policy approach for dexterous manipulation which uses an event signal to transition between the different policies, enabling text input to be translated to hand and wrist commands for a anthropomorphic manipulator.

I. Introduction

Dexterous manipulation requires not only a robot hand with the physical capacities, but also advanced task planning, wrist trajectory generation, contextual selection of grasp types and precise hand control [1]. This requires trained models or controllers with different goals or capabilities to seamlessly integrate together to achieve a start-to-end solution. Recent developments in Large Language models [2] and Visual-Language-Action Models for Robots (VLAs) [3] have shown that they offer an effective approach for high-level planning based upon language inputs. These models require large training data-sets and make us of robotic manipulation data-sets such as the open-X-embodiement which contains 1.5M trajectories [4], or others with an order of 60-70k trajectories [5], [6]. The trained models have been effectively applied to a number of pinch grippers with different embodiments to allow translation of text input to motion planning [4], [7]. The complexity of these generalist models for robotic manipulation is growing, from the notable RT-1 transformer model in 2022 [8], to the RT-2 model [9] in 2023 with 55 Billion parameters, and recently the generalised manipulation Octo model with 93 Million parameters [10], However, VLAs have been only demonstrated its use with 1 DoF pinch grippers. For more dexterous, anthropomorphic manipulators it is unclear if these large models can provide the precision and complex actuation signals required.

There are a number of existing methods which could be combined with the high level planning capabilities of VLAs for dexterous manipulators. The use of visual servoing [11] is a widely applied method for approaching and grasping objects, and has been used effectively for three finger hands and anthropomorphic hands. However, this approach struggles to utilize contextual or environmental information to inform a grasping strategy or policy. One approach for dexterous control of more complex manipulators is the use of diffusion models [12] which are well suited for robot motion learning and planning. These have been successfully applied to manipulation [13], [14], and extended to work on a multi-fingered hand[15], [16]. These models typically require data-collection and training for a specific task, however, they have the capacity for smooth and precise control, capturing of contextual information and direct generation of motor control signals. The advantages of diffusion models correspond to some of the weaknesses of VLAs, such that a hybrid approach could be advantageous for anthropomrophic dexterous manipulation.

We propose a hybrid framework (Fig. 1) for dexterous

[1]The authors are with Faculty of Mechanical Engineering, Swiss Federal Institute of Technology Lausanne, 1015 Lausanne, Switzerland cheng.pan@epfl.ch

manipulation which combines a fine-tuned openVLA pre-trained model [7] for language conditioned arm and hand positioning, with a diffusion policy for dexterous grasping [12] that is able to apply on a multi-fingered hand. Although we propose this as a generalizable approach, we deploy and evaluate on our compliant, series elastic actuated anthropomorphic ADAPT Hand [17], [18] which can robustly and dexterously interact with the environment, demonstrate the first use of VLAs on a multi-fingered hand to our best of knowledge. By fine-tuning the openVLA model for approaching the object or end target, and using the output grasp percentage as the trigger to switch to the diffusion model for object interactions, we can achieve dexterous manipulation capabilities in response to a language input. This approach is tested on a small sub-set of objects for the full grasping pipeline, showing a success rate of over 80%, compared to the VLA-only baseline. Likewise, specific data collection and training methodology alongside the robotic hardware leads to a number of desired behaviours: (1) multi-modal grasping, with the environmental context or location of an object on the table changing the grasping strategy, (2) failure recovery when an object is dropped or slips from the hand. (3) the compliance in the hand providing stability and resilience to contact and environmental collisions.

In the remainder of this paper we first introduce the methods, providing details of the robotic setup and the hybrid approach proposed. The experimental setup and results are given, before finishing with a discussion of the contributions and future outlook.

II. Methods

In this section, we describe the methodology of the proposed hybrid VLA-diffusion model and experimental setup used to perform an autonomous pick-and-place task involving multiple objects and placement locations.

A. VLA & Diffusion Model Switching Framework

Our framework switches between VLA and diffusion models, to leverage the benefits of both. Due to the differences and structural limitations of both, this is non-trivial. For high level text to robot arm motion, our framework uses openVLA, a model pre-trained on robotic arms with pinch grippers [7]. Given a language input at the beginning, based on visual feedback this model outputs 6 values for the robot arm end effector pose, and one signal to control the grasp percentage. For grasping control, a diffusion policy model [12] is used, capable of learning motion controllers from arbitrary sensor data input to an arbitrary number of output commands. As such, the two models have different inputs and outputs and providing different robot control signals. Secondly, to integrate the two methods, the switching between the two models must be automatic. This requires a 'stopping condition' signal to mark the end of the motion execution from each model: not a property of both models a priori.

Fig.2 illustrates how the two models are used synergistically together and how they are switched during execution by tracking the evolution of a float parameter, the event signal σ. The VLA is used to generate the movement of the arm, to move the hand to the proximity to the object. The scalar output of the VLA is repurposed to indicate different events in the time sequence; to indicate both when the hand is in proximity to the target location for both picking and placing. For the diffusion policy, an additional signal was recorded during training to indicate the successful end of grasp, and the transition back to the VLA. When combined, the robot should move autonomously following a input command, as illustrated in Fig.1.

One limitation of this hybrid system is directly tied to the limitation of diffusion policy controllers, where a unique diffusion policy per object is necessary. In this work, a lookup table is used to select one based on the language input to the VLA (see Fig.2).

B. Robotic setup

1) ADAPT Hand 2: To demonstrate our framework's capabilities to perform dexterous manipulation on a multi-finger robotic hand, we use the custom built ADAPT Hand 2[17]; an anthropomorphic robotic hand controlled with 13 motors. The hand incorporates series compliance at the base joints (metacarpophalangeal, carpometacarpal) which allow for safe interaction with the environment[18]. This provides some physical filtering to potentially noise actuation signals generated by the transformed based controllers. Fig.3 show the ADAPT Hand 2 and its notable features including the anatomical kinematic design and continuous skin.

2) Experimental Setup & Teleoperation: Fig.4 depicts the robotic setup used for recording training data and execute the learned motions. The ADAPT Hand 2 is attached directly to the Universal Robot 5 (UR5) robot arm to perform a pick and place task. The data collection is purely performed with the real robot, using a custom teleoperation system developed in[17]. The teleoperation system uses the hand tracking system of the Apple Vision Pro headset to capture the users hand poses, and performs a joint-to-joint transfer of the arm and hand to translate this to robot motion. This is made possible by the highly anthropomorphic design of the hand. To capture the visual data which is used as sensor signals used to control the hand the setup also includes two cameras: static to the world (cam1) and on the hand (cam2).

We focus on the pick-and-place of three representative objects with varying geometry (red pepper, tape, and piece of paper) which are to be be placed in one of two locations (yellow or purple plate) shown in the image on the right. Although this is a small sub-set of objects is reflects a range of different grasp types or modalities.

To run both the trained VLA and diffusion models, we use the Nvidia RTX 4090 GPU to achieve a control frequency of approximately 5 Hz for the respective models.

C. Data Collection

The data collection process must capture the relevant data for the two different models and the switching framework. Shown in Fig.5, the data used to train the VLA and the

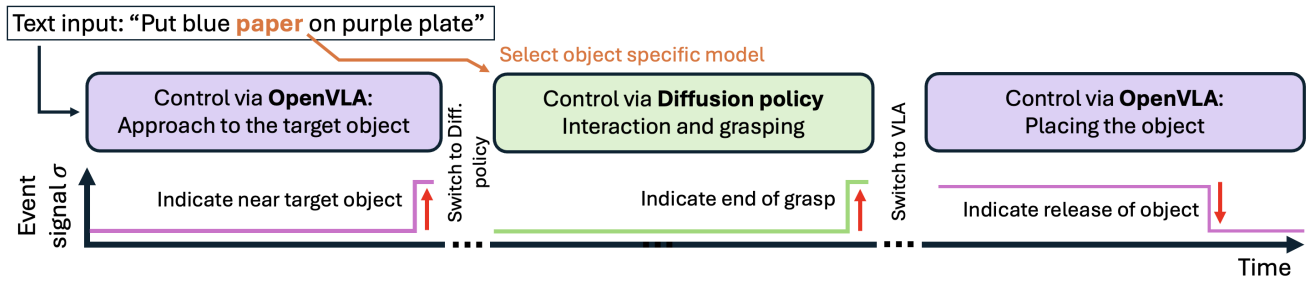

Fig. 2. Depiction of the concept to switch between the VLA and diffusion model using a common event signal σ that tracks key moments in the pick-and-place task.

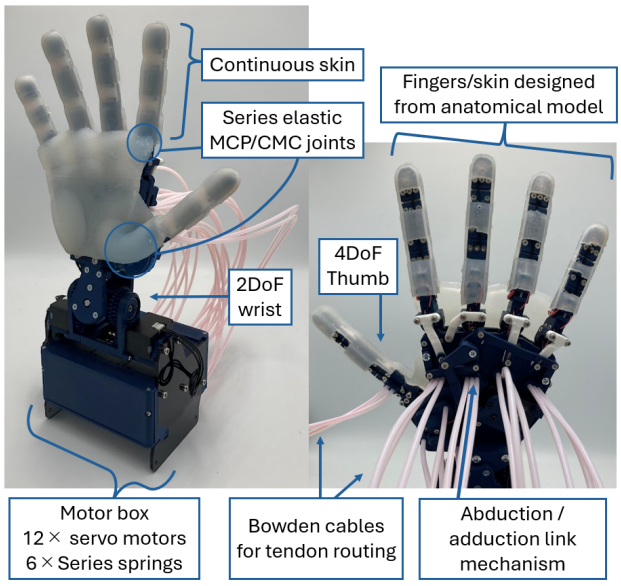

Fig. 3. ADAPT Hand 2, highlighting the soft continuous skin, compliant series elastic finger joints, and the anatomically driven design.

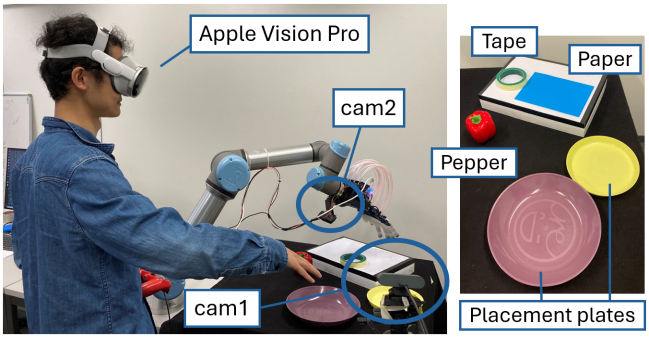

Fig. 4. Left) Robot setup for gathering training-data through teleoperation, showing the use of the Vision pro, and the location of the two cameras for capturing training data. Right) The test objects and environment used for data-capture and testing.

diffusion policy is captured separately. For the VLA model (Fig.5A) the robot is teleoperated to perform the full pick-and-place task, while the operator manually records the event signal. However, only the motion before and after the grasping task (as indicated by the event signal) is used to train the model. A key feature in this process is to deliberately close the hand in mid-air when the event signal is first applied (see second to left image on Fig.5A). Since the event signals is originally pre-trained to correspond to the gripper motion, the model most successfully performs if the hand closes when the robot is in proximity to the target. When using the VLA model, a pre-programmed generic power grasp motion sequence is used as a proxy for a 1 DoF gripper. For every object and placement combination, 20 trials are recorded. To capture the event signal, the operator can manually press a button on a PS4 game controller.

The training data for the diffusion policy is collected only for the grasping motion. For every object, the hand is placed approximately 5cm above the object with a total of 40 trials for tape and paper, and 30 trials for pepper are recorded. For these demonstrations, the robot is teleoperated to perform different grasping strategies (e.g. sliding off the table or directly picking) and recovery motions for two or three trials (e.g. purposefully failing to grasp the object and re-grasping) alongside the event signal which is triggered when the hand lifts the object. The event signal for the diffusion model is zero for the total duration except for the final 10 steps of each trial.

D. Model training

The VLA model is fine-tuned based the pre-trained open-VLA model [7] on the dataset excluding the grasping periods. Instead of using images from a single camera as shown originally[7], we combine two images from cam1 and cam2. The two images are resized to 224x144 and 224x80, to then be vertically concatenated into a single image(see Fig.1). We use the default fine-tuning parameter settings of openVLA model but with a batch size of 22, and set image augmentations of fine-tuning as False (primary camera position and light condition are kept same during data collection and testing). The fine-tuning continues until the training action accuracy exceed 95 % and converges. The fine-tuning is loaded on a cluster virtual machine with a single A100-80GB.

The training of diffusion policy model is based on the collected grasping demonstration. Images from two cameras are resized into 320x240 as inputs of the model, during training, images are randomly cropped images into 288x216 as data augmentation. The training parameter settings are same as in the CNN-based diffusion policy in [12]. The

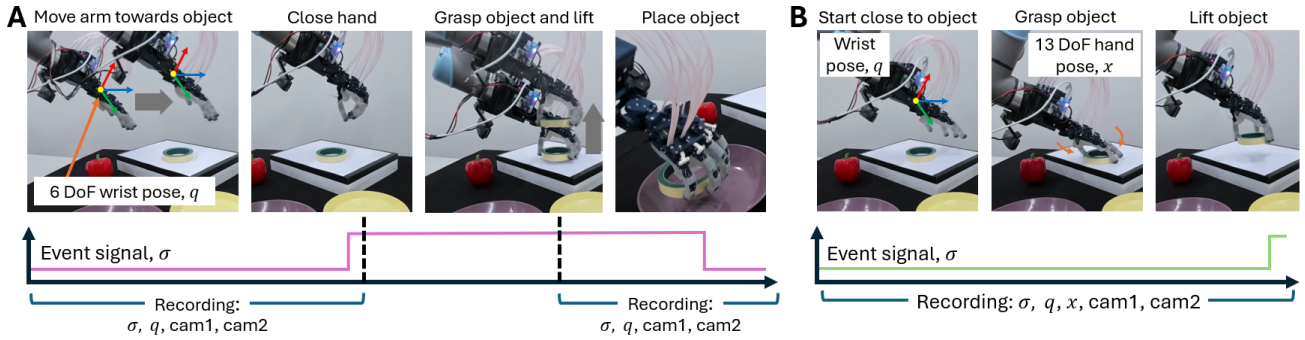

Fig. 5. A) Data-collection process for the VLA which includes the full grasping process, and the event signal recording. B) Data-collection process for the diffusion model which includes only the grasping portion of the demonstration.

model is trained for 1500 epochs on a custom GPU.

E. Experimental Plan

A series of experiments are conducted to evaluate the proposed framework, starting with assessing the VLA and diffusion policy individually before combining altogether for one autonomous motion. The VLA is evaluated on its ability to move the robot hand towards the target object. Here, the robot is started from a fixed initial position and commanded to move.

The diffusion policy is tested for three different performances, which leverage potential of diffusion policy for high-dimensional action space and handling multi-modal action distributions. One is on the success rate when the hand is placed increasingly far away from the target object: 5, 10, and 15cm from the default position for the tape and pepper. The ability to perform multi-modal grasps is also evaluated by shifting the position of the tape and a cuboid of dimensions 7.5x7.5x1 cm (specifically used for this experiment to demonstrate this behavior) from the edge of the platform. The third is if the robot is able to perform recovery behavior when grasping attempts fail.

Finally, the two processes are combined to perform the pick-and-place task as illustrated in Fig.1. Here, the combined process is tested on all objects and placement locations for five times each, while the arm and event signal trajectories are recorded. The full pick-and-place motion is also performed on the VLA model trained on the full grasping process to provide a benchmark for our framework. When using the VLA model for grasping, the pre-programmed grasp sequence used for the VLA data collection is re-used as the 1 DoF gripper proxy. To measure the success beyond a binary metric, the success of the task is given as one of five possible scores including:

- 1.00: Full task success
- 0.75: Failed to place on the correct plate
- 0.50: Failed to place on incorrect plate
- 0.25: Failed to grasp correct object
- 0.00: Approach wrong object

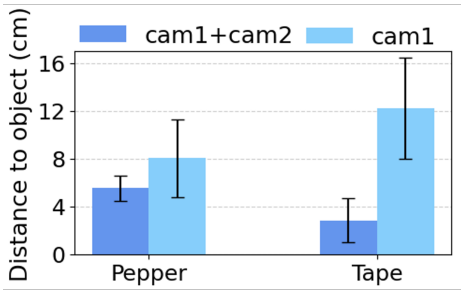

Fig. 6. The x-y offset from the target object when the VLA is used to approach an object. This includes results when the VLA is trained with concatenated images of both primary and wrist camera views vs. VLA trained with single primary camera view

III. Experimental Results

A. VLA Performance

To demonstrate the limitations of a VLA to perform the entire grasping sequence, its success rate is evaluated using two of the test objects. The blue paper (most challenging object to grasp) was excluded from this test as when fine-tuning using data of all three objects, the training fails to converge effectively (with action accuracy reaching only 80 %), The VLA was fine-tuned until it achieves an action accuracy over 95% (requiring 50k fine-tuning steps), and we evaluate its performance on the entire grasping and placing task for 5 repeats for each combination, where the objects are placed in random within the testing area and commanded to place on a specific color plate. As shown in Table. I, the VLA with scalar output for grasping can only occasionally grasp the pepper and is unable to grasp tape, which requires more precise and dexterous manipulation. The score achieved pre-dominatnly relates to the VLA's ability to move the arm to the correct object.

TABLE I
SUCCESS RATE OF THE VLA FOR DIFFERENT OBJECTS FOR GRASPING AND PLACING TO PLATES.

End Target	Pepper	Tape	Blue Paper
Purple plate	0.45	0.20	-
Yellow plate	0.40	0.25	-

Despite the limitations of the VLA to perform the entire grasping sequence, through the appropriate training it can

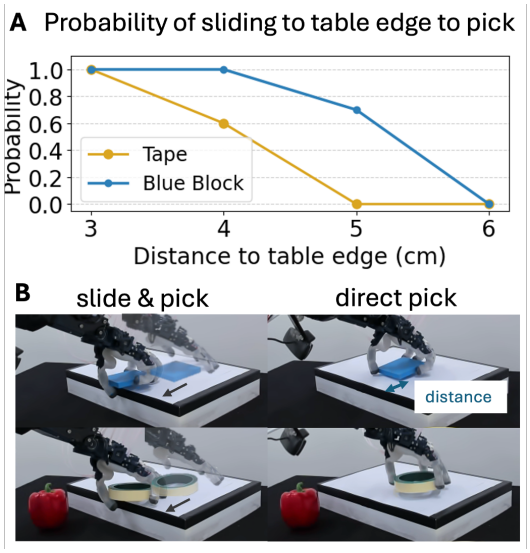

Fig. 7. For two test objects (tape and blue-block) the grasping mode (sliding and picking, or picking) is given for when the object is placed an increasing distance from the table edge. Below, the pictures demonstrate this multi-modal picking behaviour with slide and pick, and direct pick shown.

enable the hand to reach to a location near the object based upon the text input. To evaluate the 'reaching' precision of the VLA, the offset between the centre of the hand and the object is recorded in the x-y plane (i.e. parallel to the desk) after running the VLA. The position is recorded when the event signal is given by the VLA to switch to the diffusion model. This is repeated for pepper and tape objects which are randomly placed within the test area. The VLA is fine-tuned using a dataset that excludes the grasping process, with training continuing until the action accuracy exceeds 95 % (50k fine-tuning steps). To emphasize the advantage of concatenated images of both primary and wrist camera views, we also train a VLA with only images from the primary camera.

The results, given in Fig. 6 demonstrated that the VLA with combined images from cam1 and 2 is capable of bringing the robot hand to the target object within an offset of 6 centimeters. The addition of the visual input from both cameras significantly improves the localization precision for the tape, reducing the mean of offset reduced from 12*cm* to 3*cm*. The improvement offered by the two camera approach is that with only one, there is a lack of depth information, however two cameras starts to compensate for this.

B. Diffusion Model

To highlight the capabilities of the diffusion policy model to perform precise manipulation and learn multi-modal behavior, we perform grasping tests only using the diffusion policy model. To demonstrate the multi-modal behavior and capacity to leverage the environment for grasping, we focus on picking up the tape and the blue-block. The diffusion model policies have been trained with human demonstration of picking the objects using both slide & pick and direct pick (see Fig. 7 (b)) demonstrations, with an equal number of trajectories for each grasping type. The slide & pick motion

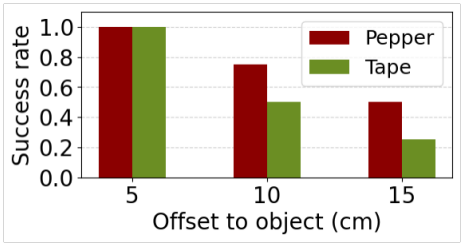

Fig. 8. Grasping success rate of the diffusion model when the hand is initialized with an offset between the centre of the hand and the centre of the object.

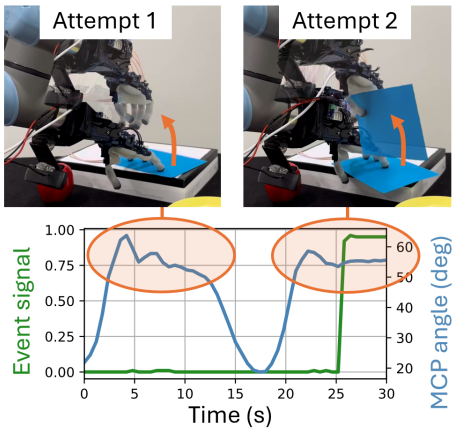

Fig. 9. Demonstration of the diffusion model's ability to recover from failures and the corresponding event signal that indicates a successful grasp.

moves the object to the edge of table, making it becomes easier for thumb to hold the thin object from bottom, while the direct pick requires all fingers precisely align with the height of the object to achieve a successful grasp. In the grasping test, we place the object at varying distances from the table edge and the grasp type evaluated for 5 repeats at each location. Fig. 7 (a) shows that as the object is placed closer to the table edge, the probability of using slide & pick grasping type increases; this reflects how as a human you might slide an object close to the edge of a table, but pick it, if it is in the middle. For all grasping trials, 5 trials for each distance, both objects succeed for all locations, highlighting the effectiveness of diffusion policy in challenging grasping tasks.

Although diffusion policy can control the hand precisely, it requires that the hand is located nearby the objects before execute the manipulation (the valid area depends on the coverage of demonstration for training). To show this limitation, and hence the need for the VLA, we offset the arm from the center of object in a x-y plane parallel to the desk as use this as the starting position of grasping tests. We perform the grasping tests for pepper and tape. Fig. 8 shows that if offset between hand and object exceeds 15 cm, the grasping success rate drops below 50 % for both objects. This emphasizes the necessity of VLA for moving the robot hand close enough to the target object, and that the VLA typical error of 6cm is within the tolerance of the diffusion model.

Fig.9 shows the ability of the diffusion policy to recover

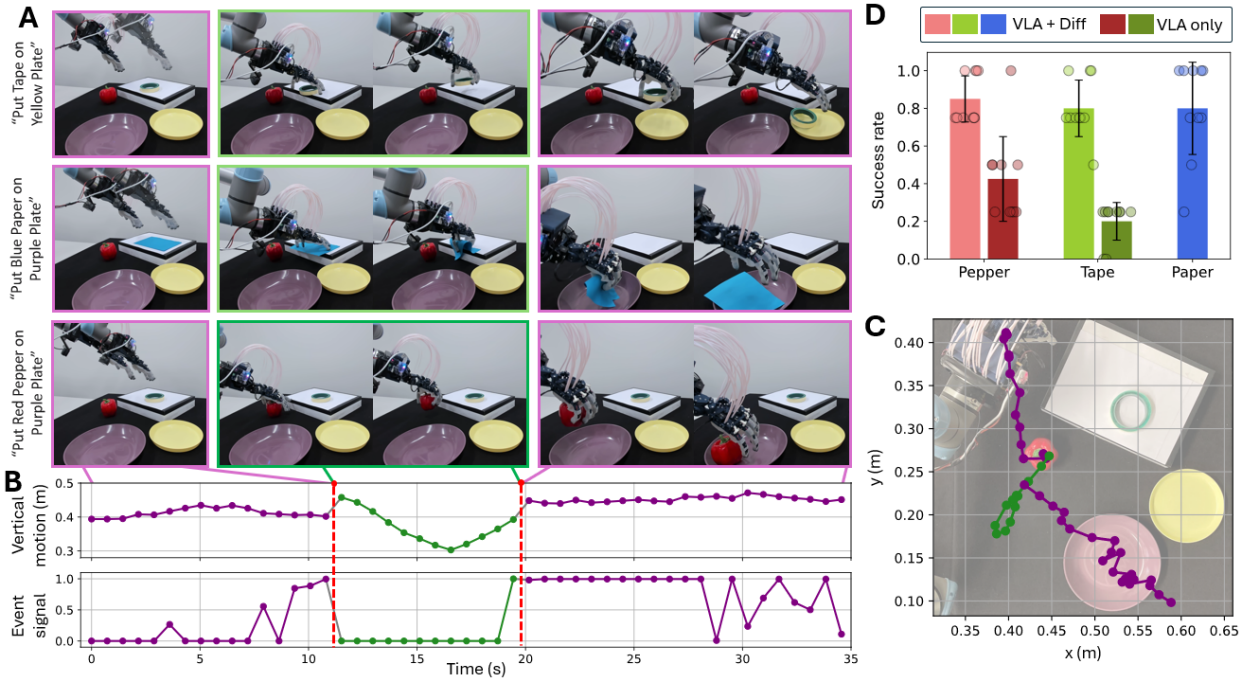

Fig. 10. Demonstration of the VLA-Diffusion switching framework. A) Pictorial sequence of the robot performing the pick-and-place task. B) Trajectories of the vertical motion and event signal over time. C) x-y planar trajectory of the robot arm. For B) and C), the color coding represents which robot controller is in action. D) Success score of the pick-and-place sequence for the VLA+Diffusion model or VLA-only.

from failure. Here, the robot moves to pick up the paper, but fails to fully grasp until the second attempt. The two attempts can be observed by the two peaks in the blue signal showing the angle of the metacarpophalangeal joint (base joint) of the index finger. Importantly, the event signal stays zero on the first attempt when the robot fails, resulting in an automatic reattempt. When the grasp is successful, the event signal does then increase to 1. This shows not only the diffusion policy's recovery ability but the stability of the event signal.

C. Combined VLA & Diffusion Model

Fig.10 shows the result of the combined pipeline of the VLA and diffusion policy. Fig.10A shows three of the six total object-placement combinations. In Fig.10B, the trajectory of the vertical motion alongside the event signal is shown for the pepper picking command. Likewise Fig.10C shows the planar x-y motion of the end effector when viewed from above. The purple and green colors correspond to the recordings from the VLA or diffusion policy guided motions respectively. The trajectories clearly show the switching behaviour. First, the VLA to diffusion model switch occurs when a) close by to the target object and b) when the event signal from the VLA reaches to one. Then, the same phenomena is repeated for switching from the diffusion policy to VLA again. Finally, when the robot is above the bowl, the event signal returns to zero to release the object to complete the task.

Fig.10 D shows the success score for every object-placement combination. For every object the success score is above 0.8. This is a significant improvement when compared to when only using the VLA. The VLA model was purposefully only trained on the pepper and tape, as including the paper increased the task difficulty too much, leading to low training action accuracy and failed convergence. Even then, the VLA-only trials score 0.43, and 0.20 for the pepper and tape respectively. This result strongly highlights the advantage of utilizing the two models to target specific skills in the pick-and-place task.

IV. Discussion & Conclusion

In this work we introduce a framework for dexterous manipulation which combines the relative advantages of a VLA for language input to high level planning, with that of a diffusion policy model for contextual, precise generation of motor signals applicable for a multi-fingered hand. Focusing on a limited set of objects we demonstrate the need for both of these models and the means by which they can be effectively integrated. We deploy this framework on our compliant anthropomorphic hand.

One of the key limitations of this work is the need to train a diffusion model for the addition of each object. Developing a more generalized approach for this, or leveraging existing data-sets to train the diffusion model could enable a more generalized approach. Another limitation is on the openVLA model being solely pre-trained on robots with 1DoF pinch grippers, which limits the use of such models for multi-fingered hands. To diversity the types of end-effectors, fundamental changes to the models or large-scale data collection on various robotic hands could be explored.

Opportunities exist to explore how hardware can enhance learning-based controllers. In this work, the compliant hand not only provided collision resilience but also improved

robustness in grasping tasks. Incorporating variable stiffness as a control input could further enhance grasping robustness, particularly in learning-based methods that struggle with repeatable precise motions. On the sensor feedback, a natural progression follows that of integrating tactile feedback into the fingers in addition to the existing cameras for better control on the interaction forces.

Acknowledgements

This research project was supported by the Microsoft Accelerate Foundation Models Research (AFMR) grant program and the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 945363.

References

[1] A. Billard and D. Kragic, "Trends and challenges in robot manipulation," *Science*, vol. 364, no. 6446, eaat8414, 2019.

[2] F. Zeng, W. Gan, Y. Wang, N. Liu, and P. S. Yu, "Large language models for robotics: A survey," *arXiv preprint arXiv:2311.07226*, 2023.

[3] Y. Ma, Z. Song, Y. Zhuang, J. Hao, and I. King, "A survey on vision-language-action models for embodied ai," *arXiv preprint arXiv:2405.14093*, 2024.

[4] A. Padalkar, A. Pooley, A. Jain, *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models," *arXiv preprint arXiv:2310.08864*, 2023.

[5] A. Khazatsky, K. Pertsch, S. Nair, *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," *arXiv preprint arXiv:2403.12945*, 2024.

[6] H. R. Walke, K. Black, T. Z. Zhao, *et al.*, "Bridgedata v2: A dataset for robot learning at scale," in *Conference on Robot Learning*, PMLR, 2023, pp. 1723–1736.

[7] M. J. Kim, K. Pertsch, S. Karamcheti, *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[8] A. Brohan, N. Brown, J. Carbajal, *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[9] A. Brohan, N. Brown, J. Carbajal, *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint arXiv:2307.15818*, 2023.

[10] O. M. Team, D. Ghosh, H. Walke, *et al.*, "Octo: An open-source generalist robot policy," *arXiv preprint arXiv:2405.12213*, 2024.

[11] E. G. Ribeiro, R. de Queiroz Mendes, and V. Grassi Jr, "Real-time deep learning approach to visual servo control and grasp detection for autonomous robotic manipulation," *Robotics and Autonomous Systems*, vol. 139, p. 103 757, 2021.

[12] C. Chi, S. Feng, Y. Du, *et al.*, "Diffusion policy: Visuomotor policy learning via action diffusion," *arXiv preprint arXiv:2303.04137*, 2023.

[13] C. Chi, Z. Xu, C. Pan, *et al.*, "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots," *arXiv preprint arXiv:2402.10329*, 2024.

[14] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," *arXiv preprint arXiv:2401.02117*, 2024.

[15] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and C. K. Liu, "Dexcap: Scalable and portable mocap data collection system for dexterous manipulation," *arXiv preprint arXiv:2403.07788*, 2024.

[16] T. Lin, Y. Zhang, Q. Li, *et al.*, "Learning visuotactile skills with two multifingered hands," *arXiv preprint arXiv:2404.16823*, 2024.

[17] K. Junge and J. Hughes, "Adapt-teleop: Robotic hand with human matched embodiment enables dexterous teleoperated manipulation," Under review, 2024.

[18] K. Junge and J. Hughes, "Robust anthropomorphic robotic manipulation through biomimetic distributed compliance," *arXiv preprint arXiv:2404.05262*, 2024.