VLA Model-Expert Collaboration for Bi-directional Manipulation Learning

Tian-Yu Xiang[†], Ao-Qun Jin[†], Xiao-Hu Zhou[*], Mei-Jiang Gui, Xiao-Liang Xie, Shi-Qi Liu,
Shuang-Yi Wang, Sheng-Bin Duang, Si-Cheng Wang, Zheng Lei, Zeng-Guang Hou[*]

Abstract— The emergence of vision-language-action (VLA) models has given rise to foundation models for robot manipulation. Although these models have achieved significant improvements, their generalization in multi-task manipulation remains limited. This study proposes a VLA model-expert collaboration framework that leverages a limited number of expert actions to enhance VLA model performance. This approach reduces expert workload relative to manual operation while simultaneously improving the reliability and generalization of VLA models. Furthermore, manipulation data collected during collaboration can further refine the VLA model, while human participants concurrently enhance their skills. This bi-directional learning loop boosts the overall performance of the collaboration system. Experimental results across various VLA models demonstrate the effectiveness of the proposed system in collaborative manipulation and learning, as evidenced by improved success rates across tasks. Additionally, validation using a brain-computer interface (BCI) indicates that the collaboration system enhances the efficiency of low-speed action systems by involving VLA model during manipulation. These promising results pave the way for advancing human-robot interaction in the era of foundation models for robotics. (Project website: https://aoqunjin.github.io/Expert-VLA/)

Index Terms— Human-Robot Collaboration; Human Factors and Human-in-the-Loop; Learning from Demonstration

I. INTRODUCTION

Motivated by the successful application of large-scale data to enhance generalization and robustness in computer vision [1] and natural language processing [2], recent efforts in robot learning have focused on leveraging extensive manipulation data to develop robotic foundation models [3]–[6]. These studies design algorithms trained on diverse tasks, environments, and robotic embodiments, aiming to develop generalized policies across settings and platforms. Beyond dataset scale, robotic foundation models incorporate principles from vision and language models, using language instructions—processed via pre-trained models such as LLaMA 2 [7]—to guide manipulation, while vision models
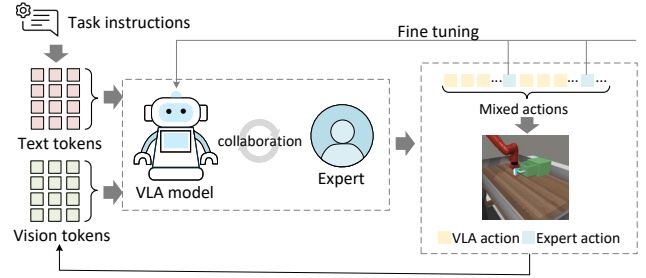
Fig. 1. The proposed VLA model-expert collaboration system integrates a VLA model and expert interactions to enhance manipulation. The VLA model generates actions by processing task instructions as text tokens and environmental inputs as vision tokens. Meanwhile, the expert makes decisions at a lower frequency, assisting the VLA model. Expert-executed actions are collected to fine-tune the VLA model, improving system performance.

like SigLIP ViT [8] condition action sequences on visual inputs. These architectures, termed **V**ision-**L**anguage-**A**ction (VLA) models [9], integrate prior task knowledge from vision and language, surpassing traditional robotic learning approaches [10].

Although VLA models have advanced autonomous manipulation abilities over traditional robotic learning algorithms, challenges persist in developing universe policy across the environment due to the scarcity of high-quality demonstrations and the diversity of manipulation tasks. Compared to computer vision and natural language processing, the scale of robotic manipulation datasets remains relatively limited. For example, the Open X-Embodiment dataset [3], the largest open-source robotic manipulation dataset, contains approximately 2.5 million demonstrations—even significantly fewer than pre-large-model era datasets in other domains, such as ImageNet with over 10 million images and GPT-2, trained on 8 million web pages [11], [12]. Furthermore, manipulation tasks exhibit greater heterogeneity and abstraction than vision and language processing. Unlike perception tasks, manipulation skills are inherently difficult to transfer, even for biological intelligence such as humans. In addition, a single manipulation task can be executed through multiple strategies, expanding the state space in policy learning.

To effectively deploy VLA models in target environments, their manipulation capabilities need to be enhanced for downstream tasks [3], [6], [13]–[16]. One approach is fine-tuning VLA models with task-specific manipulation data, a widely adopted strategy in large-model. Pre-trained VLA

models across multiple embodiments and environments exhibit positive transfer in downstream application, as evidenced by improved performance on target tasks [3], [13]. Another approach involves integrating expert decisions with the policy model to semi-autonomous systems. By delegating limited actions to experts and assigning the majority of routine operations to the policy model, this collaboration reduces the expert workload while mitigating the limitations of the VLA models in complex cases. Although expert-in-the-loop frameworks are common in semi-autonomous robotics [17], [18], their integration with VLA models is still an open problem.

To overcome these limitations, this study integrates expert-in-the-loop and fine-tuning techniques to enable collaborative manipulation and learning between experts and VLA models. The VLA model is first fine-tuned with a small amount of task-specific data, followed by collaboration with experts to accomplish target tasks. During the collaboration, human experts become more familiar with the system and more skillful in manipulation. Manipulation data collected during VLA model-expert collaboration is stored in a buffer for subsequent fine-tuning, enabling continuous performance improvement (See Fig. 1). The contributions of this study can be summarized as follows:

- Semi-autonomous manipulation is achieved via collaboration between the VLA model and experts. To the best of our knowledge, this work is a pioneer study in investigating VLA model-expert collaboration.
- The collaborated process enables bi-directional learning: VLA models can be further fine-tuned using manipulation data, while experts adapt to the VLA model.
- Experimental results in the MetaWorld environment confirm the effectiveness of collaboration. With an action ratio of the VLA model to the expert set to $4:1$, the success rate of the VLA model in different tasks improves by $6.2\%/13.5\%$ for MT10/MT50 benchmark, and the number of action steps for human experts decreases by 82.24%.

II. RELATED WORKS

A. VLA Models for Robot Learning

Building on the success of vision and language foundation models, VLA models have emerged as a promising approach for developing generalist robot policies [10], [19]. These models leverage visual and language representations to provide high-level task instructions and contextual cues for low-level actions.

Based on their input-output structures [19], VLA models can be categorized into four types: One-Step input with Discrete-Action output (OSDA) [4], [9], [20], Historical-Step input with Discrete-Action output (HSDA) [21]–[23], One-Step input with Continuous-Action output (OSCA) [14], [24]–[26], and Historical-Step input with Continuous-Action output (HSCA) [5], [6], [27].

The key distinction between one-step and historical-step models is whether actions are predicted solely from the current observation or incorporate historical context. While robot action spaces are inherently continuous, the VLA model's action space can be either continuous or discretized, depending on the design of the action head. A discrete action head leverages the structure of the language decoder to discretize the continuous action space, assigning specific values to each token in the output layer, transforming action prediction into a classification problem [4], [9], [22]. In contrast, continuous actions can be generated using methods such as a diffusion-based head [5], [28] or a flow-matching-based head [14]. This study evaluates representative VLA models from different categories within the proposed VLA model-expert collaboration system [4], [5], [14].

B. Fine-tuning Techniques for Robot Manipulation Models

Fine-tuning plays a crucial role in adapting pre-trained robot models to downstream applications [3]–[6], [13], [14]. The most straightforward approach involves fine-tuning the models with a limited number of target manipulation trials, which has shown effectiveness in several models [3]–[5], [14]. Due to the challenges and time-consuming nature of data collection, self-improvement techniques have emerged. These techniques allow models to fine-tune using synthetic data generated by the model itself [6], [13]. However, erroneous manipulations within the generated data can degrade model performance. Another fine-tuning approach involves reinforcement learning (RL), where high-level planning or low-level control policies are optimized using designed rewards [15], [16], [29]. While effective, RL requires extensive agent-environment interactions, making it a challenge to deploy. This study proposes an alternative fine-tuning paradigm where expert interactions with the model serve as an optimal policy. By leveraging historical manipulation data during the collaboration with experts, the VLA model can refine its capabilities, improving performance through interaction.

III. METHOD

A. General Structure of VLA models

The VLA models take vision input and language instructions as conditioned states to predict actions, which can be represented as $\mathcal{V} \times \mathcal{L} \to \mathcal{A}$. Here, $\mathcal{V}$ represents the visual input space, $\mathcal{L}$ represents the language instruction set, and $\mathcal{A}$ represents the action space of the robot. The VLA model establishes a mapping between vision-language input and the action space.

As a multimodal model, the vision and language inputs are first encoded through separate encoders, each designed for its respective modality. The deep representations extracted by these encoders are then fused, either through networks like FiLM [30] or by directly concatenating. Since this study does not focus on the design of the VLA model, here a general form for the input-output relationship of VLA models is provided:

$$a_i = \pi_{\text{VLA}}(l_i, v_i) \tag{1}$$

where a_i is the action or action sequence generated by the VLA model, $l_i \in \mathcal{L}$ is the language instruction, $v_i \in \mathcal{V}$ is
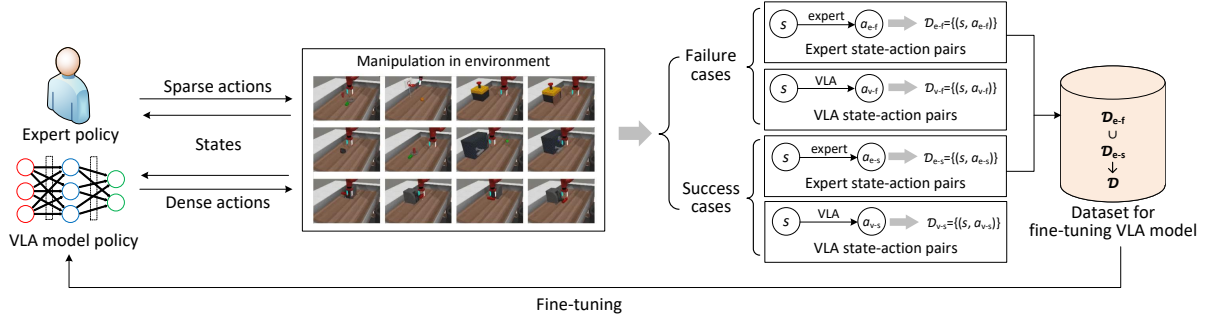
Fig. 2. Collaboration pipeline between VLA model and expert for manipulation and learning.

Algorithm 1 : Collaborated Learning

Require: VLA model's policy: π_{VLA}, expert's policy π_e, fine-tuning steps: S, and collaboration epoch: T

1: Load pre-trained VLA model, initialize a data buffer $\mathcal{D}_{\text{buffer}}$ and a dataset $\mathcal{D}$ for fine-tuning VLA.
2: **for** $x = 1$ to T **do**
3:     **while** Buffer $\mathcal{D}_{\text{buffer}}$ is not full **do**
4:         Carry out tasks based the collaborated manipulation pipeline with π_{VLA} and π_e.
5:         Collect actions generated based on π_e and corresponding vision/language states in to $\mathcal{D}_{\text{buffer}}$.
6:     **end while**
7:     Save $\mathcal{D}_{\text{buffer}}$ into dataset $\mathcal{D}$ and clear $\mathcal{D}_{\text{buffer}}$.
8:     **for** $s = 1$ to S **do**
9:         Sample mini-batch from D.
10:        Update π_{VLA} with supervised learning paradigm.
11:     **end for**
12: **end for**

---

the visual input, and π_{VLA} is the policy of the VLA model. The policy π_{VLA} predicts the action at the i-th step based on the vision observation v_i and language instruction l_i. Note that depending on the model's structure, the vision input may either include historical information or only the current observation.

Given the continuous nature of the robot action space, when the VLA model employs a discrete action head, the predicted discrete action must be mapped back to the original continuous space:

$$a_i = \frac{\hat{a}_i}{\text{vocab\_size} - 1}(a_{\max} - a_{\min}) + a_{\min} \qquad (2)$$

where $\hat{a}_i$ denotes the discrete action value corresponding to the predicted token, vocab_size is the vocabulary size, and $a_{\max}$ and $a_{\min}$ represent the upper and lower bounds of the action, respectively. This transformation maps discrete actions back to the continuous action space.

B. Expert Policy

Two expert policies are considered in this study.

1) Rule-based Policy: The rule-based policy is implemented within the MetaWorld simulation environment [31].

The policy is defined by the contributors, with the robot being aware of both the task goal and the targets' position. The robot directly controls the arm to reach the desired position. Given that the inverse dynamics of the robot are known, this policy provides a near-optimal solution for accomplishing the tasks.

2) Human Users Policy: The human users' policy involves participants controlling the robot arm during the task. Although participants are encouraged to manipulate the arm to reach the target, their varying proficiency, coupled with potential mismatches between the 2D and 3D environments, may affect performance. As a result, while the policy can achieve the target, it may not be optimal.

C. Expert-VLA Collaboration

The collaboration between VLA models and experts consists of two key processes: manipulation and learning. In the manipulation step, the expert policy assists VLA models in accomplishing various manipulation tasks. The learning step utilizes data collected during manipulation to fine-tune the VLA model, further enhancing its performance (See Fig. 2).

1) Collaborated Manipulation: Given that VLA models can handle most manipulation tasks except under extreme conditions, this study incorporates a limited number of expert policy actions as a complement to the VLA model. The collaborated manipulation process follows a straightforward design: for a given task, the VLA model autonomously executes actions for N steps, followed by one step from the expert policy, repeating this cycle until the task is either completed or reaches the failure threshold.

Since the majority of steps in a given manipulation task can be successfully completed by the VLA model, the proportion of actions generated by the VLA model can be higher than those from the expert policy (e.g., four times more). This collaborative approach enhances the performance compared to a VLA model-only policy while reducing the burden on user-driven manipulation.

2) Collaborated Learning: Through the interaction between the VLA model and the expert policy, new manipulation data can be collected and used for further fine-tuning of the VLA model. Despite failure cases occur during the collaborated manipulation stage, due to the low proportion

Models	Type	V-model	L-model	Fusion
π_0 [14]	OSCA	PaliGemma (Vision-language model) [32]		
OpenVLA [4]	OSDA	DINOv2 [33]	SigLiP [8]	Concat
Octo [5]	HSCA	Light-CNN	T5 [34]	Concat

of expert-policy actions failing to accomplish tasks. The expert policy in these cases can still be considered a near-optimal action at the corresponding steps. Consequently, this data is also used for fine-tuning the VLA model. The detailed implementation of the collaborative learning process is provided in Algorithm 1. It should be noted that the collaborative learning process is bidirectional. If the expert policy is provided by human participants, interaction with the system can also enhance the participant's proficiency in manipulation.

IV. EXPERIMENTS AND RESULTS

The experiments are designed to validate the following hypotheses:

- VLA models and experts benefit from the collaboration by enhancing the performance of VLA models and reducing the experts' workload.
- VLA models and experts can adapt to manipulation tasks through collaboration, enabling bi-directional learning.

A. Implementation Details

1) VLA models: In this study, three representative state-of-the-art models are selected to validate the proposed framework [4], [5], [14]. These models differ in input (one-step/historical steps) and output (continuous/discrete action), representing different types and routes of VLA models (Table I).

2) Experimental Environment: The proposed framework is validated in the MetaWorld environment using the ML10 and ML50 benchmarks [31], which evaluate multi-task learning algorithms with 10 (ML10) or 50 (ML50) tasks. Since VLA models are designed as foundation models for manipulation tasks, they must handle a variety of tasks, not just a single one. Therefore, this study is evaluated under the multi-task paradigm. For each task, 50 trajectories are collected using a rule-based policy for fine-tuning the VLA models, with each trajectory limited to a maximum of 500 steps. Data are captured using a fixed camera setup with an elevation angle of -25° and an azimuth of 145°.

3) Training Details: A series of data augmentation techniques are applied to the vision inputs during fine-tuning to enhance model generalization. These augmentations include random resized cropping (90% of the original size), random brightness adjustment ($\pm 20\%$), random contrast adjustment (within $[0.8, 1.2]$), random saturation adjustment (within $[0.8, 1.2]$), and random hue adjustment (± 0.05).

TABLE II

COMPARISON BETWEEN OCTO COLLABORATED WITH RULE-BASED EXPERT POLICY AND HUMAN EXPERT POLICY IN MT10 ($N = 4$).

Tasks	Successful rate			Steps	
	V	V-R	V-H	H	V-H
window open	1.00	1.00	0.98	85.96	18.65
reach	0.34	0.32	0.86	68.98	12.07
peg insert	0.30	0.18	0.52	157.86	28.42
drawer close	1.00	1.00	1.00	42.94	16.42
drawer open	0.92	1.00	0.96	68.54	21.33
push	0.56	0.20	0.80	107.58	13.13
button press	0.30	1.00	0.92	120.10	13.70
window close	1.00	1.00	0.96	88.40	20.81
pick place	0.54	0.46	0.58	112.87	14.21
door open	0.96	1.00	1.00	148.00	19.08
Average	0.69	0.72	0.86	100.12	17.78

[1] 'V': VLA model (Octo); 'V-R': Collaboration between VLA model and rule-based expert policy; 'V-H': Collaboration between VLA model and human expert policy; 'H': Human expert policy.
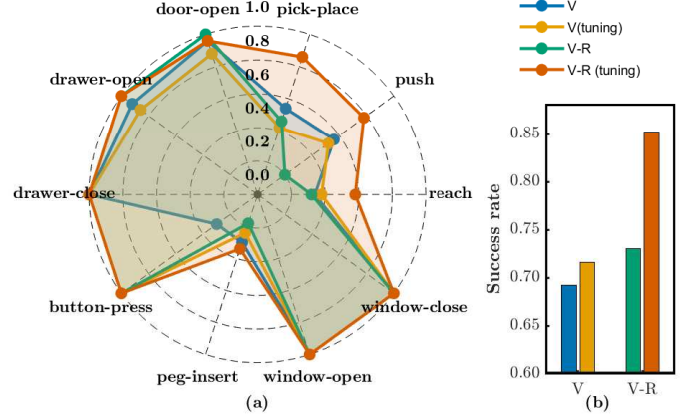

Fig. 3. Comparison of the baseline VLA model (Octo) and the VLA model after collaborative learning (tuning). The success rates of the fine-tuned VLA model—with and without the rule-based expert policy (V vs. V-R)—are presented at the task level (a) and at the average level (b) in the MT10 benchmark.

The action outputs of VLA models are normalized using min-max normalization for the discrete action model and z-score normalization for the continuous action model, following prior work [5], [28]. The models are implemented based on the official code provided by the authors [4], [5], [30]. Each model is fine-tuned with 800K sampled data from rule-based trajectories, optimized using the optimizer specified in the original paper or code.

4) Evaluation: During the evaluation stage, each task in MT10 or MT50 is tested 50 times with randomly initialized states. To ensure a fair comparison, all VLA models across different settings are evaluated using the same set of random seeds for the task initialization.

TABLE III

Benchmark	Model	Baseline	Expert-VLA Collaborated Manipulation					
			$N = 32$	$N = 16$	$N = 8$	$N = 4$	$N = 2$	$N = 1$
MT10	π_0 [14]	0.754	0.746	0.758	0.788	0.846	0.892	0.948
	OpenVLA [4]	0.854	0.862	0.892	0.904	0.924	0.954	0.988
	Octo [5]	0.692	0.680	0.676	0.698	0.716	0.822	0.824
	Improvement		-0.004	+0.008	+0.030	+0.062	+0.122	+0.153
MT50	π_0 [14]	0.566	0.568	0.601	0.651	0.728	0.808	0.918
	OpenVLA [4]	0.844	0.870	0.890	0.916	0.938	0.946	0.965
	Octo [5]	0.446	0.471	0.495	0.539	0.594	0.658	0.756
	Improvement		+0.018	+0.043	+0.083	+0.135	+0.185	+0.261

[1] Baseline models are initialized with pre-trained weights and fine-tuning following the implementation details.
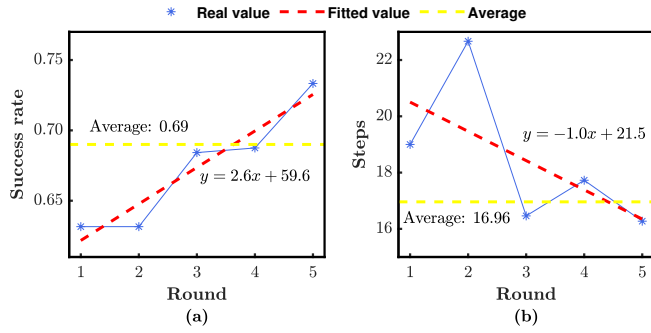[2] Improvement: average improvement from the baseline (Only VLA model manipulation).

Fig. 4. Visualization of success rate and action steps executed by human expert over round in hard tasks (successful rate lower than average) under MT10.

B. Collaborated Manipulation

1) Successful Rate Improves for the VLA Models: For VLA models [4], [5], [14], incorporating a small proportion of expert policy—either rule-based or human user policy—consistently improves success rates across different models in both MT10 and MT50 benchmarks. As shown in Table III, increasing the ratio of expert actions based on rules enhances performance. This aligns with intuition, as expert policies generally outperform VLA policies; thus, a higher proportion of expert guidance leads to better outcomes.

It is noteworthy that although the success rate on the MT50 benchmark—which includes more manipulation tasks—is lower, the relative improvement is more pronounced. As the ratio of expert actions increases, the success rates across the two benchmarks converge (Baseline: 0.767 vs. 0.619; $N = 1$: 0.920 vs. 0.880). This observation indicates that, even when the VLA model may fail to complete a task under larger manipulation sets, its actions are not entirely erroneous.

Furthermore, human user policies have been integrated with Octo to evaluate human-VLA collaboration Table II. Five participants interacted with Octo in real-time on the MT10 benchmark (5 participants × 10 times/task, aligning

with rule-based expert experiments) under the setting of $N = 4$ (four VLA model actions followed by one human expert action). The results show a significant improvement in task completion rate compared to the baseline model. Notably, human collaboration yielded better performance gains than rule-based policies, likely because the VLA model is fine-tuned using rule-based data. Human inputs, being more flexible and diverse, complement the original VLA model policy.

2) Manipulation Steps Decrease by Human Users: The benefits of collaboration for human users are evident in the reduction of workload, as reflected in the number of action steps executed by participants. This claim is validated by directly comparing action steps in a pure human-policy setting with those in human-VLA collaboration. Five participants performed tasks in the MT10 benchmark (5 participants × 10 trials per task, following prior settings). The results are summarized in Table II.

With a VLA-to-human action ratio of $4 : 1$ ($N = 4$), the VLA model is expected to take about 80% of the actions, leading to an equivalent reduction in human effort. The observed decrease (82.24%) closely aligns with this expectation and slightly surpasses it. These findings suggest the collaboration not only reduces the number of human-executed actions but also enables the VLA model to sometimes take more optimal actions than human users.

C. Collaborated Learning

1) VLA Models Improving through Historical Manipulation Data: To test whether historical manipulation data can be used to fine-tune the VLA model, the Octo model is re-tuned using collaborated manipulation data with rule-based expert ($N = 4$). As shown in Fig. 3, the success rate of the VLA model improves for most manipulation tasks after re-tuning with collaboration data. The average success rate across different tasks increases by 0.038 (from 0.692 to 0.730). This improvement indicates that the VLA model benefits from learning during collaboration, likely due to the

Experimental setting

Collaboration with VLA model   T=15s/step=77

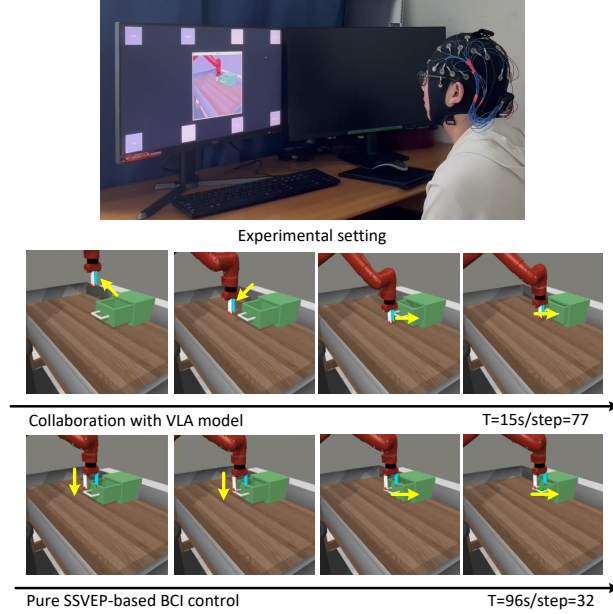Pure SSVEP-based BCI control   T=96s/step=32

Fig. 5. Application of the collaboration framework in SSVEP-based BCI: A comparison between pure SSVEP-based control and the collaboration between the VLA model and the BCI user. Although in some cases the policy of the human participant performs better than the VLA model (steps: 77 vs. 32), the collaboration system significantly improves time efficiency for a given task (time: 15s vs. 96s).

expert policy guiding the model to complete tasks it could not previously accomplish. This manipulation data provides the VLA model with valuable examples of corner cases it would not otherwise encounter, thus enhancing its overall performance.

Another comparison evaluates the performance of collaborative manipulation before and after collaborative learning. An improvement in the average success rate is observed (0.136, from 0.716 to 0.852). Notably, after collaborative learning, the improvement due to collaborative manipulation is much higher (0.122 vs. 0.024). This may be because, in cases where the VLA model fails when operating independently, re-tuning brings it closer to success. It may also imply that the enhancement in VLA performance is more pronounced than what is reflected by the success rate.

2) Human Users Become More Skillful During the Collaboration: During collaboration, human users become more familiar with the VLA system and may adjust their policy to improve performance with the VLA model. This analysis focuses on the challenge tasks in MT10 where the success rate is lower than average ('pick place', 'push', 'peg insert', 'reach'). As shown in Fig. 4, this adaptation is observed across different users in the challenge tasks during the first five rounds of interaction with the VLA model. The average success rate shows a strong positive correlation with the number of interaction rounds (Pearson correlation: 0.95). Success rates improve from below average to above average. Similarly, the number of action steps taken by human users to complete tasks shows a negative linear correlation with

TABLE IV
PRELIMINARY RESULTS OF THE APPLICATION OF THE COLLABORATION
SYSTEM VLA MODEL IN SSVEP-BASED BCI ($N = 16$).

Tasks	Participant 1		Participant 2	
	success rate	steps	success rate	steps
window open	1.00 (5/5)	100.20	1.00 (5/5)	160.75
drawer close	1.00 (5/5)	66.40	1.00 (5/5)	79.80
button press	1.00 (5/5)	82.60	1.00 (5/5)	67.80
door open	1.00 (5/5)	83.60	1.00 (5/5)	83.80

rounds (Pearson correlation: -0.63). Initially, action steps are higher than average but decrease over time. Since the VLA model remains unchanged in these settings, this improvement is attributed to learning by the participants.

V. DISCUSSION

The proposed expert-VLA collaboration system has been empirically validated in the SSVEP-based BCI system. The SSVEP-based BCI paradigm typically requires a sustained period of visual stimulation to evoke a steady response in the brain, as reflected by EEG signals for decoding. In this context, the input action signal from the human participant (expert) is slow, limiting the system's responsiveness. However, by leveraging the VLA model for most actions, the proposed collaboration system enhances the whole system's ability to respond at higher speeds. This improvement not only increases efficiency but also significantly reduces the user's workload.

The EEG cap used in the experiment is the Emotiv Flex, collecting signals at 128 Hz, with electrodes near the occipital lobe selected for decoding. The decoding algorithm employed is Canonical Correlation Analysis (CCA). As this study is not focused on the BCI system itself, the experiment primarily aims to validate the feasibility of the VLA model-expert collaboration system.

Two participants are involved in this preliminary experiment, performing four different tasks in the MT10 benchmark. The results are presented in Table IV, showing that both participants can complete the tasks with the collaboration of the VLA model. Additionally, participants were asked to complete the manipulation task using pure SSVEP-based BCI control. Although the total number of steps may be reduced in some tasks, a significant reduction in time is achieved through collaboration with the VLA model, as it performs most actions at a much faster speed compared to the BCI system input (see the supplementary video and Fig. 5).

VI. CONCLUSION

This study presents a collaboration framework between VLA models and experts for bi-directional manipulation learning. During manipulation, the proposed framework benefits both VLA models and experts by improving VLA performance while reducing the workload of human experts.

Beyond manipulation, the framework facilitates bidirectional learning: the VLA model can be re-fine-tuned using manipulation data, while human experts improve their skills through interaction. This work provides a novel perspective on human-machine interaction, offering an effective approach to enhancing the efficiency of low-frequency human action input systems. Furthermore, it enables continuous improvement of VLA model performance through real-world application. Future work will focus on fine-tuning VLA models through online interaction and deploying the system in real robotic environments.

REFERENCES

[1] A. Kirillov *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[2] J. Achiam *et al.*, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[3] Q. Vuong *et al.*, "Open x-embodiment: Robotic learning datasets and RT-x models," in *Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition@ CoRL2023*, 2023.

[4] M. J. Kim *et al.*, "OpenVLA: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[5] O. Mees *et al.*, "Octo: An open-source generalist robot policy," in *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.

[6] K. Bousmalis *et al.*, "Robocat: A self-improving generalist agent for robotic manipulation," *Transactions on Machine Learning Research*, 2023.

[7] H. Touvron *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[8] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 975–11 986.

[9] B. Zitkovich *et al.*, "RT-2: Vision-language-action models transfer web knowledge to robotic control," in *Proceedings of the Conference on Robot Learning*, 2023, pp. 2165–2183.

[10] 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.

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[12] A. Radford *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[13] M. Dalal, A. Mandlekar, C. R. Garrett, A. Handa, R. Salakhutdinov, and D. Fox, "Imitating task and motion planning with visuomotor transformers," in *Proccedings of the Conference on Robot Learning*, 2023, pp. 2565–2593.

[14] K. Black *et al.*, "π_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

[15] Y. Guo, J. Zhang, X. Chen, X. Ji, Y.-J. Wang, Y. Hu, and J. Chen, "Improving vision-language-action model with online reinforcement learning," *arXiv preprint arXiv:2501.16664*, 2025.

[16] T. Carta, C. Romac, T. Wolf, S. Lamprier, O. Sigaud, and P.-Y. Oudeyer, "Grounding large language models in interactive environments with online reinforcement learning," in *Proceedings of the International Conference on Machine Learning*, 2023, pp. 3676–3713.

[17] W. Wang, R. Li, Y. Chen, Y. Sun, and Y. Jia, "Predicting human intentions in human–robot hand-over tasks through multimodal learning," *IEEE Transactions on Automation Science and Engineering*, vol. 19, no. 3, pp. 2339–2353, 2021.

[18] M. Ma and L. Cheng, "A human-robot collaboration controller utilizing confidence for disagreement adjustment," *IEEE Transactions on Robotics*, 2024.

[19] X. Li, P. Li, M. Liu, D. Wang, J. Liu, B. Kang, X. Ma, T. Kong, H. Zhang, and H. Liu, "Towards generalist robot policies: What matters in building vision-language-action models," *arXiv preprint arXiv:2412.14058*, 2024.

[20] H. Zhen, X. Qiu, P. Chen, J. Yang, X. Yan, Y. Du, Y. Hong, and C. Gan, "3D-VLA: A 3D vision-language-action generative world model," in *Proceedings of the International Conference on Machine Learning*, 2024.

[21] H. Wu *et al.*, "Unleashing large-scale video generative pre-training for visual robot manipulation," in *Proceedings of the International Conference on Learning Representations*, 2023.

[22] A. Brohan *et al.*, "RT-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[23] C.-L. Cheang *et al.*, "GR-2: A generative video-language-action model with web-scale knowledge for robot manipulation," *arXiv preprint arXiv:2410.06158*, 2024.

[24] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023.

[25] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn, "BC-Z: Zero-shot task generalization with robotic imitation learning," in *Proceedings of the Conference on Robot Learning*, 2022, pp. 991–1002.

[26] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell, "Real-world robot learning with masked visual pre-training," in *Proceedings of the Conference on Robot Learning*, 2023, pp. 416–426.

[27] S. Reed *et al.*, "A generalist agent," *arXiv preprint arXiv:2205.06175*, 2022.

[28] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, 2024.

[29] A. Szot, M. Schwarzer, H. Agrawal, B. Mazoure, R. Metcalf, W. Talbott, N. Mackraz, R. D. Hjelm, and A. T. Toshev, "Large language models as generalizable policies for embodied tasks," in *The Twelfth International Conference on Learning Representations*, 2023.

[30] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[31] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine, "Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning," in *Proceedings of the Conference on Robot Learning*, 2020, pp. 1094–1100.

[32] L. Beyer *et al.*, "Paligemma: A versatile 3b vlm for transfer," *arXiv preprint arXiv:2407.07726*, 2024.

[33] M. Oquab *et al.*, "Dinov2: Learning robust visual features without supervision," *Transactions on Machine Learfning Research Journal*, 2024.

[34] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.