2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Rethinking the Practicality of Vision-language-action Model: A Comprehensive Benchmark and An Improved Baseline

Wenxuan Song*[1], Jiayi Chen*[1], Xiaoquan Sun*[1,2], Huashuo Lei[1], Yikai Qin[1], Wei Zhao[3], Pengxiang Ding[3,4], Han Zhao[3,4], Tongxin Wang[1], Pengxu Hou[1], Zhide Zhong[1], Haodong Yan[1], Donglin Wang[3], Jun Ma[1], Haoang Li[1]

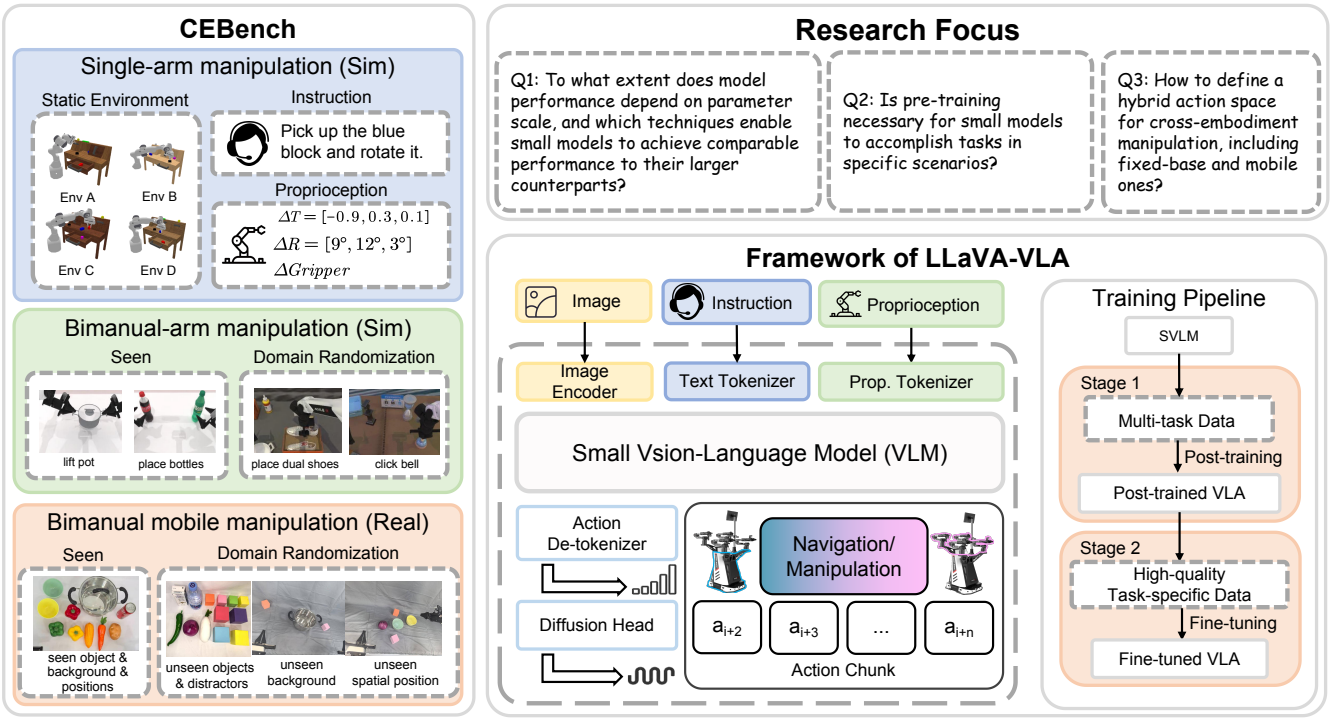

Fig. 1: Overview of this work. We conduct a comprehensive study on the practicality of vision-language-action models. We first construct a cross-embodiment benchmark, CEBench, across simulation and the real world, and offer diverse evaluation settings. Then we explore three critical aspects (Q1-Q3) and offer several key findings. Based on the above findings, we introduce our LLaVA-VLA, a lightweight yet effective baseline capable of mobile manipulation.

***Abstract*— Vision-Language-Action (VLA) models have emerged as a generalist robotic agent. However, existing VLAs are hindered by excessive parameter scales, prohibitive pre-training requirements, and limited applicability to diverse embodiments. To improve the practicality of VLAs, we propose a comprehensive benchmark and an improved baseline. First, we propose CEBench, a new benchmark spanning diverse embodiments in both simulation and the real world with consideration of domain randomization. We collect 14.4k simulated trajectories and 1.6k real-world expert-curated trajectories to support training on CEBench. Second, using CEBench as our testbed, we study three critical aspects of VLAs' practicality and offer several key findings. Informed by these findings, we introduce LLaVA-VLA, a lightweight yet powerful VLA designed for practical deployment on consumer-grade GPUs. Architecturally, it integrates a compact VLM backbone with multi-view perception, proprioceptive tokenization, and action chunking. To eliminate reliance on costly pre-training, LLaVA-VLA adopts a two-stage training paradigm including post-training and fine-tuning. Furthermore, LLaVA-VLA extends the action space to unify navigation and manipulation. Experiments across embodiments demonstrate the capabilities of generalization and versatility of LLaVA-VLA , while real-world mobile manipulation experiments establish it as the first end-to-end VLA model for mobile manipulation. We will open-source all datasets, codes, and checkpoints upon acceptance to foster reproducibility and future research.**

*Wenxuan Song, Jiayi Chen and Xiaoquan Sun contributed equally to this work.
Corresponding author: Haoang Li (haoangli@hkust-gz.edu.cn).
[1]The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China.
[2]Huazhong University of Science and Technology, Wuhan, China.
[3]Westlake University, Hangzhou, China.
[4]Zhejiang University, Hangzhou, China.

I. INTRODUCTION

The emergence of Vision-Language-Action (VLA) [1]–[6] models has revolutionized the field of robotics, offering powerful capabilities in visuomotor control and the comprehension of complex instructions through end-to-end learning processes. These models have demonstrated their potential in various robotic tasks, enabling systems to interpret multimodal sensory data and execute actions based on language instructions. However, despite their success, current VLAs face significant hurdles that impede their widespread deployment and practical use in real-world scenarios: 1) **Billions of parameters** make them difficult to deploy in resource-constrained environments, such as mobile platforms and consumer-grade devices. 2) **Extensive pre-training** using large-scale robotic datasets leads to prohibitive training costs and the need for vast computational resources. 3) **Fixed-base manipulation** limits their applicability to cross-embodiment deployment, *i.e.*, mobile manipulation tasks.

Existing work has partly investigated some of the aforementioned issues. TinyVLA [7] introduced a 1B-level model trained from scratch. It employs Low-Rank Adaption [8] and diffusion head for efficient training and inference. MiniVLA [9] is a variant of OpenVLA [10] with a smaller backbone and uses a VQ-VAE tokenizer to quantize actions, achieving fast and precise inference. NORA [11] utilizes FAST [12] to quantize action tokens and diffusion expert. SmolVLA [13] realizes efficient training through skipping layers, pruning visual tokens, and initializing a small VLM backbone. Despite these advances, these approaches have not systematically examined the practicality of their lightweight and pretraining-free designs, and they remain incapable of performing mobile manipulation tasks.

To improve the practicality of VLAs, we propose a comprehensive benchmark and an improved baseline. First, we introduce CEBench, a practical robotic benchmark suite that spans diverse embodiments (single arm [14], bimanual [15], and mobile bimanual) in both simulation and the real world, with explicit consideration of domain randomization. In CEBench, we collect 14.4k simulated demonstrations across 36 tasks and 1.6k high-quality real-world demonstrations across 8 tasks.

Second, using CEBench as our testboard, we study three critical aspects for VLAs: the lightweight designs, the training curriculum, and the unified action space, and empirically offer several valuable findings and insights into their choices. Informed by these findings, we propose LLaVA-VLA, a **lightweight** VLA with strong performance, which is capable of **mobile manipulation**. Figure 1 shows that LLaVA-VLA does **not need pre-training** and can be trained and deployed on consumer-grade GPUs. LLaVA-VLA initializes a compact VLM and integrates multi-view input, proprioception tokenization, and action chunking to balance efficiency and performance while maintaining minimalist model design. To eliminate the reliance on pre-training, we further adopt a two-stage training paradigm, combining post-training on multi-task data with fine-tuning on task-specific data. Finally, we design a unified action space including a specific action space for navigation and a manipulation action space combined through several special tokens.

Extensive evaluation of our LLaVA-VLA on CEBench shows that it matches or surpasses models over 10× larger, especially under domain-randomized settings, which demonstrates its strong visual generalization capabilities. LLaVA-VLA successfully manages tasks like *move to the operation table from outside and place the bottle*, which indicates that LLaVA-VLA is the first end-to-end VLA model capable of handling mobile manipulation tasks. Cross-embodiment experiments demonstrate the versatility of our model and the effectiveness of the proposed hybrid action space design. This work represents a promising step toward democratizing VLAs by making them lightweight, generalized, and practical for deployment in mobile embodied systems.

To summarize, our key contributions are:

- We construct a benchmark to evaluate the practicality of VLAs, which spins diverse embodiments in both simulation and the real world, and considers domain randomization.
- We conduct a comprehensive study on the practicality of VLAs and offer several insightful findings.
- We propose an improved baseline, LLaVA-VLA, which is lightweight, pretraining-free, and capable of mobile manipulation.
- We will release all datasets, codes, and checkpoints upon acceptance to provide a reference for future research in open-source VLAs.

II. RELATED WORKS

A. VLA Models

Vision-Language Models (VLMs) [16]–[18] extend Large Language Models (LLMs) [19], [20] to multimodal understanding by integrating visual inputs, and have shown strong cross-modal reasoning ability with pretrained vision encoders such as CLIP [21] and SigLIP [22] together with large-scale alignment training [23]. These advances further inspire extensions from language and vision to embodied action modeling. Early Vision-Language-Action (VLA) works such as RT-1 and π_0 [1], [2] trained transformer-based policies on web-scale vision-language data and large-scale robot trajectories to improve performance and generalization. OpenVLA [10] released the first open-source 7B VLA trained on public data, while OpenHelix [24] proposed a dual-system architecture. More recently, GR00T N1 [25] advances general-purpose humanoid control with a diffusion-based action generator in a dual-system architecture. PD-VLA [3] and CEED-VLA [4] explored the acceleration of inference. ReconVLA [6] and FlowVLA [26] leveraged extra low-level visual perception. However, these models remain computationally demanding and pose challenges for deployment in resource-constrained settings and mobile manipulation tasks.

III. CEBENCH

Prior benchmarks exhibit a significant gap in practical deployment, lacking comprehensive and unbiased evaluation

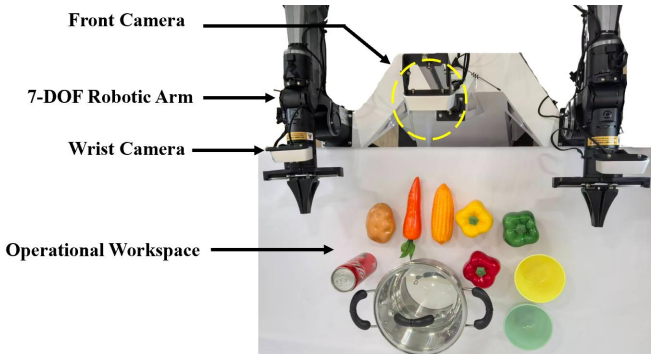

Fig. 2: Real-world setup of the Cobot-Magic system for mobile bimanual manipulation (top view).

across embodiments, potential domain randomization, and mobile manipulation needs. To bridge this gap, we propose Cross-Embodiment Benchmark (**CEBench**), a reliable benchmarking suite designed for systematic evaluation of practicality in VLAs. Notably, we treat CALVIN [14] as a subset of our dataset and mainly introduce our dataset in RoboTwin [15] and the real world.

A. System Setup

We evaluate our LLaVA-VLA in both simulated and real-world environments to comprehensively assess task performance and generalization.

Single-arm manipulation. The CALVIN benchmark [14] includes a Franka Panda single robotic arm and a table environment to study long-horizon language-conditioned manipulation and visual generalization.

Bimanual manipulation. The RoboTwin benchmark [15], built on the Sapien [27] simulator, designed to evaluate positional generalization and visual robustness. It fosters an expert data synthesis pipeline that leverages VLMs and simulation-in-the-loop refinement to automatically generate task-level execution code.

Bimanual mobile manipulation. Figure 2 shows the real-world experimental setup using the Cobot-Magic dual-arm mobile robot, equipped with four Piper robotic arms (two master arms and two puppet arms), which capture RGB images at a resolution of 480 × 640 and 30 Hz.

B. Datasets

For CALVIN, we use the official datasets. On the RoboTwin platform [15], we constructed a large-scale simulation dataset containing 14.4k trajectories and 36 tasks (400 trajectories per task) in an automatic manner. All trajectories were collected under a simplified scenario with a clutter-free tabletop and stable background and lighting. In the real world, we designed 8 tasks and collected 200 trajectories per task. The tasks include:

- **Stack bowls**: Pick the bowl and put it on the other.
- **Restore bottle**: Return a toppled bottle to its upright position.
- **Click bell**: Trigger a desk bell by pressing its small button.
- **Place vegetable**: Grasp a vegetable and correctly place it into the target container.
- **Pack bottles**: Insert two bottles neatly into a designated box.
- **Lift pot**: Grasp a pot with two arms and lift it off the table surface.

These tasks cover different difficulties, from basic pick-and-place operations to more complex bimanual interactions. We also collect 2 mobile manipulation tasks:

- **Move and fetch bottles**: move to the table and fetch the bottle on it.
- **Move and open the drawer**: move to the drawer and open it.

These data provide a solid foundation for systematic training and evaluation.

C. Evaluation and Metrics

For CALVIN, we follow the official evaluation setting and report the success rates of each sub-task as well as the average success length across 5 tasks. For RoboTwin, we select 8 representative tasks to ensure fair and efficient comparison: *click bell*, *click alarmclock*, *lift pot*, *move can pot*, *open laptop*, *pick dual bottles*, *place dual shoes*, *rotate qrcode*. The evaluation on RoboTwin is conducted on **seen** settings as well as unseen settings with domain randomization (**DR**), which includes clutter, random lighting, diverse textures, and variable table heights. For real-world experiments, we evaluate on all tasks in the datasets. To simulate the DR setting in reality, we vary the color and texture of the tabletop and randomly place distractor objects (e.g., *blocks, pens*) in the workspace, creating visual and layout variations that were unseen in training.

D. Baselines

To comprehensively evaluate the performance of policies and VLAs with fewer than 1B parameters, we conduct comparisons with the following baselines in RoboTwin and the real world in Section V.

- **ACT [28]**: A CVAE-based imitation learning approach leverages action chunking for sequence forecasting and temporal ensembling for smooth execution.
- **Difussion policy [29]**: A visuomotor policy learning framework that models action prediction through a conditional denoising diffusion process.
- **TinyVLA [7]**: A compact VLA model that leverages a lightweight multimodal backbone to efficiently generate robot actions from vision-language inputs, enabling fast inference and strong generalization across diverse tasks.
- **RDT [30]**: A diffusion-based Transformer for bimanual manipulation leverages multimodal inputs and unified action spaces for efficient few-shot learning.

For the CALVIN benchmark, we compare with representative methods on the official leaderboard.

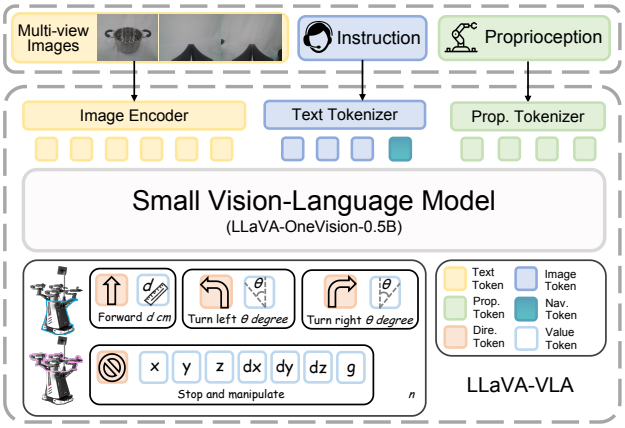

Fig. 3: Model architecture of our LLaVA-VLA.

IV. LLaVA-VLA and Its Associated Findings

To build a lightweight, pretraining-free, and cross-embodiment VLA for practical use, we systematically explore design choices of VLAs. Specifically, we formulate three research questions to guide our explorations:

Q1: To what extent does model performance depend on parameter scale, and which techniques enable small models to achieve comparable performance to their larger counterparts?

Q2: Is pre-training necessary for small models to accomplish tasks in specific scenarios?

Q3: How to define a unified action space for manipulation across fixed-base and mobile embodiments?

A. LLaVA-VLA

We conduct our study in a top-down manner. We first present our **Light**weight **VLA** (LLaVA-VLA). As shown in Figure 3, LLaVA-VLA is built upon a pre-trained LLaVA-OneVision-0.5B [31] backbone, taking concatenated multi-view images as input, augmented with proprioceptive signals, and producing action chunks through an action tokenizer. To enable mobile manipulation, we construct a hybrid action space consisting of direction tokens and their corresponding value tokens, which allows the model to seamlessly switch between navigation and manipulation. In the following part, we discuss the design choices made during its development step by step to provide the key findings (F1-F8).

B. Study on Lightweight Designs (Q1)

Toward lightweight VLAs, we first investigate whether compact models can rival large-scale counterparts (F1), then identify the key design choices that make such performance attainable (F2-4).

Key Findings 1: Model performance is not strictly proportional to parameter scale. Small VLAs can achieve performance comparable to their large-scale counterparts. Table I, LLaVA-VLA-0.5B achieves performance on par with 7B models despite less than 10% parameters. On the first sub-task, it reaches a success rate of 96.2%, comparable to 97.4% of its 7B counterpart. On the last task, it achieves 50.6%, significantly outperforming 23.5% of the 3B RoboFlamingo and 43.5% of the 7B OpenVLA. With an average action length of 3.65, it fully surpasses these larger models. These results suggest that performance gains are not solely determined by scale. With the proposed techniques, small VLAs can match or surpass larger models.

TABLE I: Comparison among different versions of LLaVA-VLA in terms of success rates and average length on RoboTwin. Here, B denotes billions.

Model	Param.	Success Rate (%) 1/5	2/5	3/5	4/5	5/5	Avg. Len. ABC→D
LLaVA-VLA (ours)	0.5B	96.2	84.8	72.6	60.8	50.6	3.65
LLaVA-VLA (ours)	7B	97.4	86.2	73.4	64.6	53.4	3.75
RoboFlamingo [32]	3B	82.4	61.9	46.6	33.1	23.5	2.47
OpenVLA [10]	7B	91.3	77.8	62.0	52.1	43.5	3.27

Key Findings 2: Multi-view images are critical as they enable stereoscopic perception of 3D space, and containing multi-view information in 1 image is an effective way. In manipulation tasks, third-person view images provide global contextual information, while first-person view images offer precise object-to-gripper positional cues, which are crucial for achieving precise manipulation. While inheriting the above information, multi-view images capture disparity information, which is essential for constructing a three-dimensional understanding of the scene. Therefore, incorporating both perspectives is essential.

Several strategies exist for handling multi-view inputs [33], [34]. Encoding each image separately and then concatenating its image tokens typically leads to an excessive number of image tokens and introduces considerable redundancy, resulting in suboptimal performance. One potential remedy is to apply token compression methods to reduce visual token count. However, this approach may incur information loss, which may result in slight performance degradation. Consequently, we adopt a simpler yet effective strategy: vertically concatenating the first- and third-person view images into a single composite image. This approach not only reduces the number of tokens while preserving complete multi-view visual information, but also aligns with the training paradigm of our VLM backbone, thereby avoiding potential performance degradation.

TABLE II: Comparison of different methods for integrating multi-view images on CALVIN.

Method	Success Rate (%) 1/5	2/5	3/5	4/5	5/5	Avg. Len. ABC→D
Concate Image Tokens	65.3	37.1	23.5	14.7	9.2	1.50
Merged Image	**94.8**	**84.5**	**71.3**	**62.5**	**53.8**	**3.68**

Key Findings 3: Proprioception is critical as it improves understanding of physical states, and tokenizing proprioception works better than encoding them by linear layers. Proprioceptive information is critical for enabling robots to infer their current state and maintain action continuity. A common approach is to encode this information using an MLP. In our design, we translate proprioception values into a sequence of proprioception tokens via a proprioception

tokenizer, which can be regarded as an inverse form of the action de-tokenizer. As shown in Table IV, this integration facilitates better exploitation of the VLM's language modeling capabilities for understanding and generating coherent actions.

TABLE III: Comparison of different methods for integrating proprioceptive information on CALVIN.

Method	Success Rate (%)					Avg. Len.
	1/5	2/5	3/5	4/5	5/5	ABC→D
MLP Projector	90.4	76.0	58.0	48.0	37.2	3.09
Prop. Tokenizer	**94.8**	**84.5**	**71.3**	**62.5**	**53.8**	**3.68**

Key Findings 4: Action chunking is critical as it strengthens the model's planning capability and action stability. Action chunking plays a pivotal role in manipulation tasks [28]. Training VLAs to predict action chunks implicitly endows them with planning capabilities and improves the temporal coherence of the generated actions. We employ this design and set the action chunking size to 5.

TABLE IV: Comparison of different chunk sizes in terms of average length on CALVIN.

Chunk Size	1	5	12	20
Avg. Len.	2.25	**3.68**	3.35	0.70

C. Study on Training Curriculum (Q2)

Key Findings 5: Cross-embodiment large-scale pre-training is not essential. Post-training on in-domain multi-task data is sufficient to establish the mapping from vision and language to action. Large-scale cross-embodiment pre-training often suffers from low-quality samples and discrepancies in action spaces, which limit its effectiveness. In contrast, domain-specific datasets are typically composed of high-quality demonstrations collected either in simulation or from human experts in real-world settings. Consequently, conducting post-training across multiple tasks within downstream datasets, followed by fine-tuning on a single task, emerges as a promising approach. In both real-world and simulated experiments with dual-arm configurations, we conduct the **two-stage** training paradigm, including post-training and fine-tuning, and make LLaVA-VLA achieves the best overall performance. This suggests that the vision-to-action mapping can be effectively learned from domain-specific data when training tasks provide sufficient diversity in goals and scenes.

TABLE V: Comparison of different training curricula on seen tasks in RoboTwin.

Training Curriculum	Open Laptop	Lift Pot
Pre-training	20.0%	18.0%
Post-training	**38.0%**	**39.0%**

D. Study on Action Space (Q3)

To design a unified action space, we first investigate whether it should be discrete or continuous (F6-7), and then explore how navigation and manipulation can be integrated within the same framework (F8).

Key Findings 6: Continuous action space in the diffusion head is not indispensable. With action chunking, discrete actions can achieve comparable performance. While many VLAs adopt a diffusion head to generate precise continuous actions, this design compromises the autoregressive nature of the model, thereby limiting its scalability when integrated with advanced techniques on VLMs and LLMs. In contrast, our approach combines action tokenization with action chunking, achieving competitive performance while preserving the autoregressive property.

TABLE VIII: Comparison among different action spaces of LLaVA-VLA on CALVIN.

Decoder Type	Success Rate (%)					Avg. Len.
	1/5	2/5	3/5	4/5	5/5	ABC→D
Discrete	94.8	84.5	71.3	62.5	53.8	3.68
Continuous	93.5	84.4	73.5	63.3	54.3	3.70

Key Findings 7: Fine-grained action tokenization does not lead to high performance. While a finer granularity in action representation might intuitively seem to improve the model's ability to capture subtle differences, it introduces more training complexity to fit a larger action space. As shown in Table IX, using a larger number of action bins leads to lower success rates compared with coarser discretization, indicating that overly fine-grained action tokenization results in a loss of efficiency and generalization.

TABLE IX: Comparison of different numbers of bins on CALVIN.

Bin Numbers	Success Rate (%)					Avg. Len.
	1/5	2/5	3/5	4/5	5/5	ABC→D
256	**94.8**	**84.5**	**71.3**	**62.5**	**53.8**	**3.68**
1024	96.4	86.5	72.1	60.2	48.6	3.65

Key Findings 8: The unified action space can be realized through a combination of direction token and value token. For mobile manipulation tasks, we aim for the VLA to simultaneously output both navigation data and manipulation data. An intuitive approach is to tokenize the navigation data in the same manner as the manipulation data. However, experiments revealed that this method is unstable and often causes the robot to abruptly resume movement during the manipulation phase after coming to a stop. We designed the navigation output as a direction token (*forward, turn left, turn right, stop*) followed by a value token representing the distance to advance or the angle to rotate. Specifically, to ensure stability during manipulation, when the direction token is *stop*, we append the value tokens for manipulation after it. This design allows the model to flexibly switch between navigation and manipulation. Furthermore, we introduce a

TABLE VI: Comparison with various manipulation baselines on CALVIN.

Category	Method	Params	w/o Pre-training	Success Rate (%) 1/5	2/5	3/5	4/5	5/5	Avg. Len. ABC→D
Generative Methods	3D-VLA [35] (*ICML'24*)	1B	×	44.7	16.3	8.1	1.6	0	0.70
	GR-1 [36] (*ICLR'24*)	195M	×	85.4	71.2	59.6	49.7	40.1	3.06
	Vidman [37] (*NIPS'24*)	1B	×	91.5	76.4	68.2	59.2	46.7	3.42
Diffusion Policy	3D Diffuser Actor [38] (*CoRL'24*)	70M	✓	93.8	80.3	66.2	53.3	41.2	3.35
Large VLA Models	RoboFlamingo [32] (*ICLR'24*)	3B	✓	82.4	61.9	46.6	33.1	23.5	2.47
	OpenVLA [10] (*CoRL'24*)	7B	×	91.3	77.8	62.0	52.1	43.5	3.27
	LLaVA-VLA (Ours)	500M	✓	**94.8**	**84.5**	**71.3**	**62.5**	**53.8**	**3.68**

TABLE VII: **Evaluation on RoboTwin benchmark.** Success rates for 8 tasks on the Seen and DR settings. Best result in each row highlighted in **Bold**.

Simulation Task	Small Model w/o Pre-training						Large Model	
	ACT		DP		LLaVA-VLA (Ours)		RDT	
	Seen	DR	Seen	DR	Seen	DR	Seen	DR
Click Bell	4.0%	2.0%	54.0%	0.0%	**81.0%**	**72.0%**	80.0%	9.0%
Click Alarmclock	11.0%	0.0%	61.0%	5.0%	**73.0%**	**65.0%**	61.0%	12.0%
Lift Pot	7.0%	2.0%	31.0%	0.0%	39.0%	**21.0%**	**72.0%**	9.0%
Move Can Pot	0.0%	0.0%	**39.0%**	0.0%	28.0%	**16.0%**	25.0%	12.0%
Open Laptop	31.0%	0.0%	49.0%	0.0%	38.0%	**31.0%**	**59.0%**	31.0%
Pick Dual Bottles	4.0%	0.0%	22.0%	0.0%	26.0%	7.0%	**41.0%**	**10.0%**
Place Dual Shoes	0.0%	0.0%	7.0%	0.0%	**8.0%**	**5.0%**	4.0%	3.0%
Rotate Qrcode	0.0%	0.0%	13.0%	0.0%	29.0%	**12.0%**	**49.0%**	5.0%
Average success	7.1%	0.5%	34.5%	0.6%	40.3%	**28.6%**	**48.9%**	11.4%

special <Navigation> token following task instructions to prompt the model to perform mobile manipulation tasks.

TABLE X: Comparison of unified action space on the mobile manipulation tasks in the real world.

Unified Action Space	Move and fetch bottles	Move and open the drawer
Action Value Token	2/10	1/10
Direction + Vaule Token	4/10	4/10

V. Evaluation of LLaVA-VLA

We comprehensively evaluate our final architecture on CEBench to evaluate its manipulation performance, capabilities of visual generalization, cross-embodiment versatility, and abilities of mobile manipulation.

A. Training Setup

All post-training is conducted on 8 NVIDIA H100 GPUs unless otherwise noted, and fine-tuning is conducted on 1 NVIDIA 4090 GPU. For the CALVIN ABC→D task split, we post-train on multiple tasks for a single epoch without fine-tuning, which costs approximately six hours. For bimanual manipulation on RoboTwin and in the real world, we perform 2 epochs of post-training followed by 8 epochs of fine-tuning.

B. Single-arm Manipulation on CALVIN

Evaluation detail. We report the average completed trajectory length (Avg. Len.) across all five subtasks as well as success rates on each subtask. Following the official ABC→D settings [14], the evaluation is conducted in an unseen scene. To ensure reliable evaluation, we test each method 1000 times.

Evaluation results. Table VI shows that our LLaVA-VLA, with a lightweight 0.5B LLM and no large robot-dataset pre-training, tops all sub-tasks in success rate and attains the best average completed length of 3.68 against other Large VLA Models [10], [32]. Compared with 3D-aware baselines [38], our LLaVA-VLA outperforms them with a 0.33 increase in the average length of the completed trajectory, demonstrating its strong spatial reasoning ability. Despite no pre-training, our model outperforms methods [35], [37] that rely on large-scale video pre-training and future image prediction. These results show that combining our techniques enables a light model to outperform architecturally complex, large-parameter, and training-expensive models.

C. Bimanual Manipulation on RoboTwin

Evaluation details. We evaluate our LLaVA-VLA 100 times in both the seen and DR settings and report success rates per task. For additional details, please refer to official settings [15].

Evaluation results. Table VII shows that as a small model without pre-training, our LLaVA-VLA achieves a success rate of 40.3% on seen tasks and 28.6% on domain-randomization tasks, outperforming diffusion-based [29] and VAE-based [28] baselines. Compared to larger VLAs that require extensive pretraining, our LLaVA-VLA achieves higher success rates in the DR environment, which indicates that our LLaVA-VLA owns visual generalization as well as spatial comprehension capabilities.

TABLE XI: Comparison of success rates on real-world bimanual tasks.

Embodiments	Basic Single-arm Tasks								Dexterous Bimanual Tasks					
Task	Stack Bowls		Restore Bottle		Click Bell		Place Vagetable		Pack Bottles		Lift Pot		Avg.	Success
Settings	Seen	DR	Seen	DR	Seen	DR	Seen	DR	Seen	DR	Seen	DR	Seen	DR
ACT [28]	30%	10%	15%	0%	30%	10%	20%	0%	10%	0%	7%	0%	18.6%	3.0%
TinyVLA [7]	25%	10%	10%	0%	25%	5%	15%	0%	20%	5%	10%	5%	17.5%	4.2%
LLaVA-VLA (Ours)	**58%**	**40%**	**38%**	**27%**	**66%**	**54%**	**50%**	**32%**	**28%**	**16%**	**25%**	**15%**	**44.2%**	**30.7%**

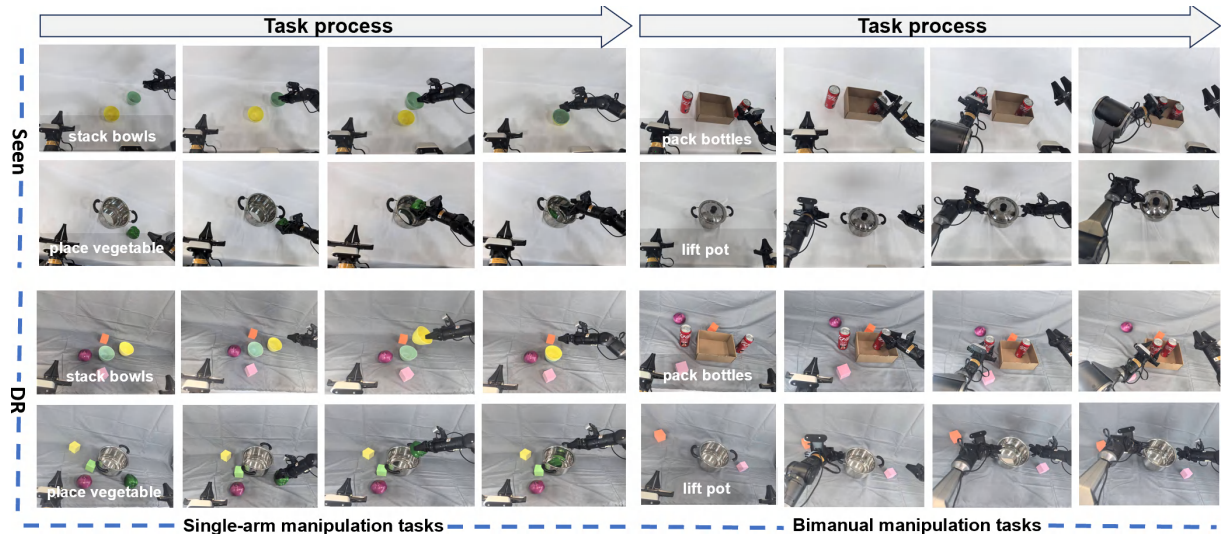

Fig. 4: Visualization of real-world tasks. The top two rows illustrate the seen tasks, while the bottom two rows correspond to settings with domain randomization. Out of the eight real-world tasks, we select two representative examples of single-arm manipulation (left) as well as two examples of bimanual collaboration (right).

D. Bimanual Manipulation in the Real World

Evaluation details. For fixed-base bimanual manipulation, each method is evaluated over 100 episodes per task under both easy and hard settings, and for mobile manipulation, we report results over 10 episodes per task.

Evaluation results on fixed-base tasks. Table XI shows that our LLaVA-VLA consistently outperforms the ACT [28] and TinyVLA [7] across 6 tasks, demonstrating its effectiveness in real-world experiments. Notably, when evaluated in unseen scenarios with background variations and distractor objects, both ACT and TinyVLA experience a dramatic drop in success rate, approaching zero. In contrast, Figure 4 shows that our LLaVA-VLA is much less affected, demonstrating strong robustness and visual generalization capabilities.

Evaluation results on mobile manipulation tasks. Since existing VLA models lack mobile manipulation capabilities, we adopt the implementation of ACT in the mobile ALOHA [39] setting as the baseline. However, Figure 5 shows that ACT suffers from low navigation accuracy, which prevents it from reliably executing manipulation tasks, resulting in only a 10% success rate. Moreover, ACT is not equipped with multi-task learning capabilities. In contrast, our approach leverages the strengths of small VLMs to learn from diverse multi-task trajectories, thereby enabling effective instruction following and precise mobile manipulation.

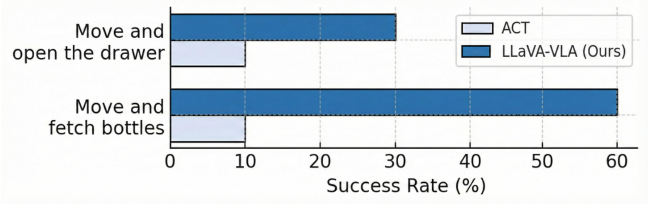

Fig. 5: Evaluation in real-world mobile manipulation tasks.

VI. Conclusion

In this work, we introduced CEBench, a practical benchmark spanning diverse embodiments in both simulation and real-world with explicit consideration of domain randomization. By exploring the design space of lightweight VLAs through extensive experiments, we developed LLaVA-VLA—a lightweight yet powerful VLA that avoids costly pre-training while maintaining strong performance. Experiments demonstrate its cross-embodiment versatility and visual generalization, while real-world evaluations establish LLaVA-VLAas the first end-to-end VLA model capable of mobile manipulation and provide a roadmap toward more practical and accessible VLA research.

VII. Acknowledgments

This work was supported in part by the Natural Science Foundation of China under Grant 62403401, in part by the National Science and Technology Innovation 2030 - Major

Project under Grant 2022ZD0208800, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515011992 and Grant 2026A1515012323, in part by the Guangdong Provincial Project under Grant 2024QN11X127, and in part by the AI Research and Learning Base of Urban Culture under Grant 2023WZJD008.

REFERENCES

[1] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, , *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *Proceedings of Robotics: Science and Systems*, 2023.

[2] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, *et al.*, "π_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

[3] W. Song, J. Chen, P. Ding, H. Zhao, W. Zhao, Z. Zhong, Z. Ge, J. Ma, and H. Li, "Accelerating vision-language-action model integrated with action chunking via parallel decoding," *arXiv preprint arXiv:2503.02310*, 2025.

[4] W. Song, J. Chen, P. Ding, Y. Huang, H. Zhao, D. Wang, and H. Li, "Ceed-vla: Consistency vision-language-action model with early-exit decoding," *arXiv preprint arXiv:2506.13725*, 2025.

[5] W. Song, J. Chen, W. Li, X. He, H. Zhao, C. Cui, P. D. S. Su, F. Tang, X. Cheng, D. Wang, *et al.*, "Rationalvla: A rational vision-language-action model with dual system," *arXiv preprint arXiv:2506.10826*, 2025.

[6] W. Song, Z. Zhou, H. Zhao, J. Chen, P. Ding, H. Yan, Y. Huang, F. Tang, D. Wang, and H. Li, "Reconvla: Reconstructive vision-language-action model as effective robot perceiver," *arXiv preprint arXiv:2508.10333*, 2025.

[7] J. Wen, Y. Zhu, J. Li, M. Zhu, Z. Tang, K. Wu, Z. Xu, N. Liu, R. Cheng, C. Shen, *et al.*, "Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation," *IEEE Robotics and Automation Letters*, 2025.

[8] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *International Conference on Learning Representations (ICLR)*, 2021.

[9] S. Belkhale and D. Sadigh, "Minivla: A better vla with a smaller footprint," 2024.

[10] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. P. Foster, P. R. Sanketi, Q. Vuong, *et al.*, "Openvla: An open-source vision-language-action model," in *8th Annual Conference on Robot Learning*.

[11] C.-Y. Hung, Q. Sun, P. Hong, A. Zadeh, C. Li, U. Tan, N. Majumder, S. Poria, *et al.*, "Nora: A small open-sourced generalist vision language action model for embodied tasks," *arXiv preprint arXiv:2504.19854*, 2025.

[12] K. Pertsch, K. Stachowicz, B. Ichter, D. Driess, S. Nair, Q. Vuong, O. Mees, C. Finn, and S. Levine, "Fast: Efficient action tokenization for vision-language-action models," *arXiv preprint arXiv:2501.09747*, 2025.

[13] M. Shukor, D. Aubakirova, F. Capuano, P. Kooijmans, S. Palma, A. Zouitine, M. Aractingi, C. Pascal, M. Russi, A. Marafioti, *et al.*, "Smolvla: A vision-language-action model for affordable and efficient robotics," *arXiv preprint arXiv:2506.01844*, 2025.

[14] O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard, "Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks," *IEEE Robotics and Automation Letters*, 2021.

[15] T. Chen, Z. Chen, B. Chen, Z. Cai, Y. Liu, Q. Liang, Z. Li, X. Lin, Y. Ge, Z. Gu, *et al.*, "Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation," *arXiv preprint arXiv:2506.18088*, 2025.

[16] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *12th International Conference on Learning Representations, ICLR 2024*, 2024.

[17] G. Team, R. Anil, S. Borgeaud, Y. Wu, J. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. Dai, A. Hauth, *et al.*, "Gemini: A family of highly capable multimodal models, 2024," *arXiv preprint arXiv:2312.11805*, vol. 10, 2024.

[18] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[19] H. Touvron, T. Lavril, G. Izacard, *et al.*, "Llama: Open and efficient foundation language models," 2023. [Online]. Available: https://arxiv.org/abs/2302.13971

[20] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A frontier large vision-language model with versatile abilities," *arXiv preprint arXiv:2308.12966*, 2023.

[21] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[22] 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.

[23] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

[24] C. Cui, P. Ding, W. Song, S. Bai, X. Tong, Z. Ge, R. Suo, W. Zhou, Y. Liu, B. Jia, *et al.*, "Openhelix: A short survey, empirical analysis, and open-source dual-system vla model for robotic manipulation," *arXiv preprint arXiv:2505.03912*, 2025.

[25] J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, S. Huang, *et al.*, "Gr00t n1: An open foundation model for generalist humanoid robots," *arXiv preprint arXiv:2503.14734*, 2025.

[26] Z. Zhong, H. Yan, J. Li, X. Liu, X. Gong, W. Song, J. Chen, and H. Li, "Flowvla: Thinking in motion with a visual chain of thought," *arXiv preprint arXiv:2508.18269*, 2025.

[27] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang, *et al.*, "Sapien: A simulated part-based interactive environment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 097–11 107.

[28] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," 2023.

[29] 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*, p. 02783649241273668, 2023.

[30] S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu, H. Su, and J. Zhu, "Rdt-1b: a diffusion foundation model for bimanual manipulation," in *The Thirteenth International Conference on Learning Representations*, 2024.

[31] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, *et al.*, "Llava-onevision: Easy visual task transfer," *arXiv preprint arXiv:2408.03326*, 2024.

[32] X. Li, M. Liu, H. Zhang, C. Yu, J. Xu, H. Wu, C. Cheang, Y. Jing, W. Zhang, H. Liu, *et al.*, "Vision-language foundation models as effective robot imitators," in *The Twelfth International Conference on Learning Representations*.

[33] 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.

[34] M. J. Kim, C. Finn, and P. Liang, "Fine-tuning vision-language-action models: Optimizing speed and success," 2025.

[35] 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 41st International Conference on Machine Learning*, 2024, pp. 61 229–61 245.

[36] H. Wu, Y. Jing, C. Cheang, G. Chen, J. Xu, X. Li, M. Liu, H. Li, and T. Kong, "Unleashing large-scale video generative pre-training for visual robot manipulation," *ICLR*, 2024.

[37] Y. Wen, J. Lin, Y. Zhu, J. Han, H. Xu, S. Zhao, and X. Liang, "Vidman: Exploiting implicit dynamics from video diffusion model for effective robot manipulation," *Advances in Neural Information Processing Systems*, vol. 37, pp. 41 051–41 075, 2024.

[38] T.-W. Ke, N. Gkanatsios, and K. Fragkiadaki, "3d diffuser actor: Policy diffusion with 3d scene representations," in *Conference on Robot Learning*. PMLR, 2025, pp. 1949–1974.

[39] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," in *Conference on Robot Learning (CoRL)*, 2024.