Interleave-VLA: Enhancing Robot Manipulation with Interleaved Image-Text Instructions

Cunxin Fan[1*], Xiaosong Jia[1*], Yihang Sun[1], Yixiao Wang[2], Jianglan Wei[2], Ziyang Gong[1], Xiangyu Zhao[1], Masayoshi Tomizuka[2], Xue Yang[1✉], Junchi Yan[1✉], Mingyu Ding[3✉]

***Abstract*— Vision-Language-Action (VLA) models have shown great promise for generalist robotic manipulation in the physical world. However, existing models are restricted to robot observations and text-only instructions, lacking the flexibility of interleaved multimodal instructions enabled by recent advances in foundation models in the digital world. In this paper, we present Interleave-VLA, the first framework capable of comprehending interleaved image-text instructions and directly generating continuous action sequences in the physical world. It offers a flexible, model-agnostic paradigm that extends state-of-the-art VLA models with minimal modifications and strong zero-shot generalization. A key challenge in realizing Interleave-VLA is the absence of large-scale interleaved embodied datasets. To bridge this gap, we develop an automatic pipeline that converts text-only instructions from real-world datasets in Open X-Embodiment into interleaved image-text instructions, resulting in the first large-scale real-world interleaved embodied dataset with 210k episodes. Through comprehensive evaluation on simulation benchmarks and real-robot experiments, we demonstrate that Interleave-VLA offers significant benefits: 1) it improves out-of-domain generalization to unseen objects by 2-3× compared to state-of-the-art baselines, 2) supports flexible task interfaces, and 3) handles diverse user-provided image instructions in a zero-shot manner, such as hand-drawn sketches. We further analyze the factors behind Interleave-VLA's strong zero-shot performance, showing that the interleaved paradigm effectively leverages heterogeneous datasets and diverse instruction images, including those from the Internet, which demonstrates strong potential for scaling up. More information can be found at the link.**

I. INTRODUCTION

The remarkable success of Large Language Models (LLMs) [1], [2], [3], [4] and Vision-Language Models (VLMs) [5], [6], [7], [8], [9] has established the paradigm of foundation models in the digital world, which are capable of generalizing across a wide range of tasks and domains. Inspired by this progress, the robotic community is actively developing robotic foundation models [10], [11], [12], [13], [14], [15] to bring similar generalizability to unseen tasks and scenarios into the physically embodied world. However, despite the demonstrated effectiveness of interleaved multimodal inputs in digital foundation models, most robotic policies today still accept only observation images and text-based instructions, falling behind VLMs that seamlessly handle mixed-modality sequences and generalize across flexible task interfaces. Relying solely on text instructions can lead to ambiguity or safety issues, as the robot may misinterpret or fail to follow instructions precisely. Text-only interfaces pose serious out-of-distribution (OOD) risks for robots: even a single ambiguous phrase (e.g., "the blue screw on the left") or unfamiliar synonym can lead to critical manipulation errors. This issue is especially severe in industrial settings, where specialized parts are easy to photograph but difficult to describe accurately.

The concept of interleaved instructions for robotic manipulation was first explored in simulation by VIMA [16], which introduced VIMA-Bench to study vision-language planning for 2D object pose estimation. With a high-level 2D action space, VIMA focuses mainly on interface unification without exploring the broader benefits of interleaved instructions, such as improved generalization or real-world applicability with low-level robotic actions. As a result, the practical value of this paradigm remains underexplored due to a lack of real-world datasets and policies capable of handling such input, as shown in Figure 1.

To develop a general and practical robot policy capable of acting on interleaved image-text instructions in the real world, a straightforward solution is to build upon VLA [11], [12], [17], [10], [13], [18] models, which naturally extend VLMs by incorporating action understanding and generation, making them well-suited for robotic tasks. However, existing VLAs [10], [11], [13] are trained primarily with text-only instructions. This limits their ability to benefit from multimodal instruction signals, which have been shown to enhance generalization in vision-language learning [1], [18]. This restriction not only reduces instruction flexibility but also prevents these models from leveraging the richer semantics and improved grounding afforded by interleaved multimodal signals. To address this limitation, we propose a new paradigm called Interleave-VLA, a simple and model-agnostic extension that enables VLA models to process and reason over interleaved image-text instructions.

High-quality image-text interleaved datasets are crucial for training Interleave-VLA. In order to bridge the gap of the lack of image-text interleaved datasets in robotic manipulation, we develop a pipeline to automatically construct interleaved instructions from existing datasets. The proposed pipeline enables automatic and accurate generation of interleaved instructions from real-world dataset Open X-Embodiment [12]. The resulting interleaved dataset contains over 210k episodes and 13 million frames, making it the first large-scale, real-world interleaved embodied dataset. This enables training Interleave-VLA with real-world interaction

[1]Shanghai Jiao Tong University [2]UC Berkeley [3]UNC, Chapel Hill
[*]Equal Contribution
✉ Corresponding authors. Email: yangxue-2019-sjtu@sjtu.edu.cn, yanjunchi@sjtu.edu.cn, md@cs.unc.edu

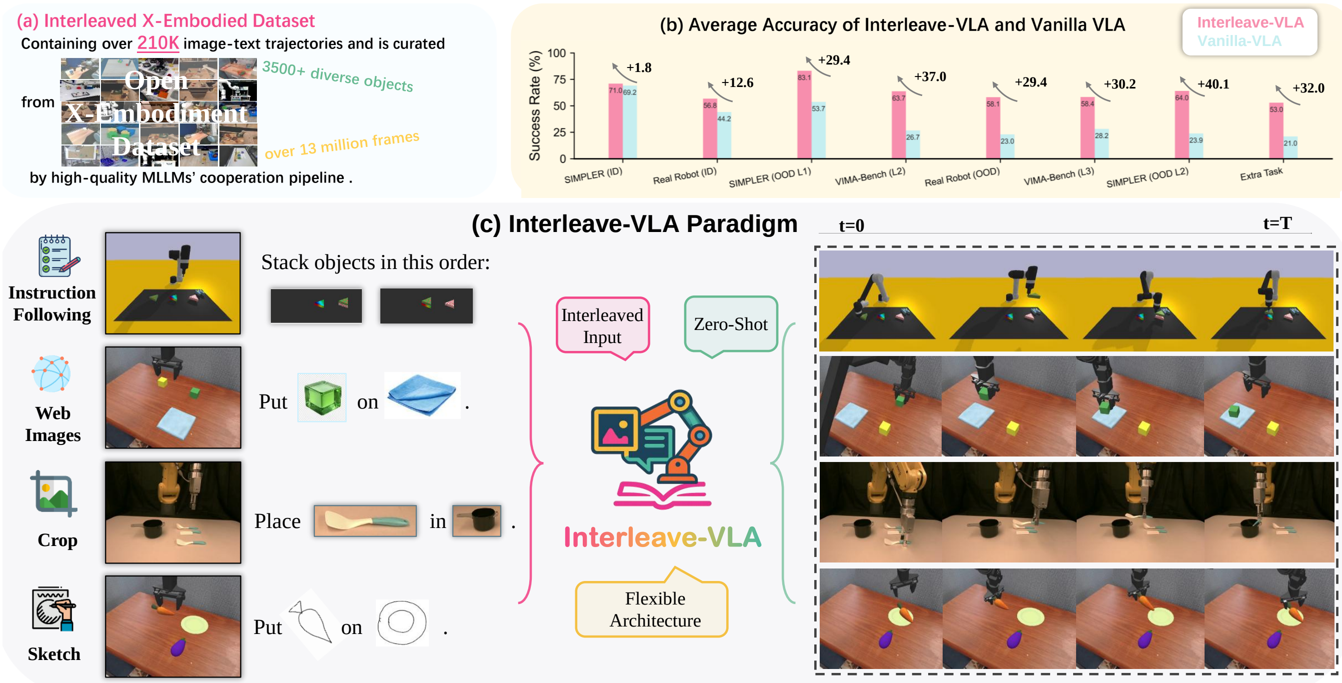

Fig. 1: **a)** Our Interleaved X-Embodiment dataset features diverse, high-quality object-centric images automatically generated from real-world robot demonstrations. **b)** Interleave-VLA achieves **2–3**× stronger out-of-domain generalization compared to text-only VLA models in both simulation and real-robot experiments. **c)** It enables flexible, **zero-shot instruction following** with user-provided, web images, and hand-drawn sketches for practical and intuitive human-robot interaction.

data and diverse visual instruction types.

We demonstrate Interleave-VLA's effectiveness by adapting two leading VLA models, OpenVLA [11] and π_0 [13], with minimal architectural changes, hence to be widely applicable to future generations of VLAs. Experimental results show that Interleave-VLA consistently outperforms its text-only counterparts for both in-domain and out-of-domain tasks. Notably, the interleaved format enables strong zero-shot generalization to novel objects and even user-provided sketches never seen in the training dataset, highlighting the robustness and flexibility of our method, as in Fig. 1.

Our core contribution can be summarized as follows.

- We introduce a fully automated pipeline that converts text-only instructions into image-text interleaved instructions, creating the first large-scale, real-world interleaved embodied dataset with 210k episodes and 13 million frames based on Open X-Embodiment.
- We propose Interleave-VLA, a simple, generalizable, and model-agnostic adaptation that enables VLA models to process interleaved image-text instructions with minimal architectural changes. To the best of our knowledge, it represents the first end-to-end robotic policy capable of handling interleaved inputs, marking the first extension of this paradigm to physical VLA models.
- Through comprehensive evaluations of Interleave-VLA on SIMPLER, VIMA-Bench, and real-robot settings, we demonstrate consistent in-domain improvements and **2–3× gains in out-of-domain generalization** to novel objects, along with emergent **zero-shot** capabilities for interpreting diverse, user-provided visual instructions, such as hand-drawn sketches.

II. Related Work

Interleaved Vision-Language Models. In the digital domain, recent advances in vision-language models have evolved from handling simple image-text pairs [7], [19], [20], [21] to processing arbitrarily interleaved sequences of images and text [22], [5], [6], [23], [8], [24], [9], [25]. This interleaved format allows models to leverage large-scale multimodal web corpora—such as news articles and blogs—where images and text naturally appear in mixed sequences. Such models have demonstrated improved flexibility and generalization, enabling transfer across diverse tasks and modalities [23]. Despite these successes in the digital world, robotic foundation models in the physical world have yet to fully exploit the benefits of interleaved image-text instructions. Motivated by the progress of interleaved VLMs, we extend this paradigm to the action modality, enabling vision-language-action models to process interleaved instructions. Our results show that multimodal learning with interleaved inputs greatly boosts generalization and displays emergent capabilities in robotic manipulation tasks.

Vision Language Action Models. Vision-language-action (VLA) models have advanced robotic manipulation by enabling policies conditioned on both visual observations and language instructions [11], [12], [17], [10], [13], [18], [26], [27]. Most prior VLA models process single [11] or multiple [10], [13] observation images with text-only instructions, with some exploring additional modalities such as 3D [28] and audio [29]. VIMA [16] pioneers the use of interleaved image-text prompts as a unified interface for robotic manipulation, primarily in simulation. However, its focus is limited to interface design, without systematically exploring the broader advantages of interleaved instructions—such as enhanced generalization and real-world applicability. As a result, most VLA models to date have continued to rely on text-only instructions. In this work, we make the first step to bridge this gap by proposing Interleave-VLA: a simple, model-agnostic paradigm that extends existing VLA models to support interleaved image-text instructions with minimal modifications. Our comprehensive experiments demonstrate that interleaved instructions substantially improve generalization to unseen objects and environments, and unlock strong zero-shot capabilities for diverse user-provided inputs. This highlights the practical value and scalability of interleaved image-text instructions for real-world robotic manipulation.

III. Interleave-VLA and Open Interleaved X-Embodiment Dataset

A. Problem Formulation

Digital foundation models [22], [30] can process multimodal prompts with arbitrarily interleaved images, video frames, and text as input, producing text as output. For robotic foundation models, this paradigm extends naturally: the model receives a multimodal prompt and outputs an action in the robot's action space. For example:

Regular: <obs> Place [the blue spoon near microwave] into [silver pot on towel].

Interleaved: <obs> Place [image1] into [image2].

where <obs> is the observation image(s), and [image1] and [image2] are images representing the target object and the destination, respectively.

B. Interleave-VLA

Our Interleave-VLA framework models the action distribution $P(A_t|o_t)$ based on the observation $o_t = (I_t, \mathscr{I}, \mathbf{q})$. Here, I_t is the observation image(s), $\mathbf{q}$ is the robot's proprioceptive state, and $\mathscr{I}$ is an image-text interleaved instruction. The instruction $\mathscr{I}$ is a sequence mixing text segments l_i and images $\mathbf{I}_i$, i.e., $\mathscr{I} = (l_1, \mathbf{I}_1, l_2, \mathbf{I}_2, \ldots)$. Existing VLA using text instruction is a special case where $\mathscr{I} = (l)$ just contains a single text segment.

Interleave-VLA is a straightforward yet effective adaptation of existing VLA models. It modifies the input format to accept interleaved image and text tokens, without changing the core model architecture. We demonstrate this approach by adapting two state-of-the-art Vision-Language-Action (VLA) models. For OpenVLA [11], we replace the original Prismatic [31] VLM backbone with InternVL2.5 [24], which natively supports image-text interleaved inputs. For π_0 [13], we retain the original architecture and only adjust the input pipeline to handle interleaved tokens. Notably, even though the underlying Paligemma [32] VLM is not trained on interleaved data, Interleave-π_0 can still be trained to effectively process interleaved instructions. This model-agnostic adaptation requires minimal changes in architecture and significantly enhances the zero-shot generalization capabilities of base models, as shown in our experiments.

C. Construction of Open Interleaved X-Embodiment Dataset

A large-scale pretraining dataset is essential for Vision-Language-Action (VLA) Models to learn actions and generalize, as reported in OpenVLA [11] and π_0 [13], this is also the case with Interleave-VLA. However, most current real-world datasets provide only text-based instructions and thus do not support training interleave-VLA models directly. We consequently design a unified pipeline to automatically relabel and generate interleaved data across diverse datasets.

Our overall dataset generation pipeline consists of three main steps: instruction parsing, open-vocabulary detection, and data quality verification, as illustrated in Figure 2. **First**, for instruction parsing, we use Qwen2.5 [33] to extract key objects from language instructions. Compared to rule-based NLP tools like SPaCy [34], LLM prompting is more robust and adaptable to diverse instruction formats. It also enables concise summarization of complex or lengthy instructions, as in datasets such as [35]. **Second**, for open-vocabulary detection, we use the state-of-the-art open-vocabulary detector OWLv2 [36] to locate and crop target objects from trajectory frames based on the parsed instruction keywords, achieving over 99% accuracy in most cases. **Finally**, we introduce data quality verification for harder cases where OWLv2 fails: Qwen2.5-VL [5] verifies the detected objects, and if needed, provides keypoints for more precise segmentation using Segment Anything [37]. This combined approach boosts cropping accuracy for challenging objects (e.g., eggplant) from less than 50% to 95%, ensuring high-quality interleaved data for downstream tasks.

We apply the dataset generation pipeline to 11 datasets from Open X-Embodiment [12]: RT-1 [17], Berkeley Autolab UR5 [38], IAMLab CMU Pickup Insert [39], Stanford Hydra [40], UTAustin Sirius [41], Bridge [42], Jaco Play [43], UCSD Kitchen [44], BC-Z [45], Langugae Table [46], and UTAustin Mutex [35] to form the first large-scale interleaved cross-embodiment dataset in real world. The curated dataset contains 210k episodes and 13 million frames, covering 3,500 unique objects and a wide range of task types.

IV. Experiments

In the experiments, we aim to discuss the following questions: (1) How is the in-domain and out-of-domain performance of Interleave-VLA compared to vanilla VLA? How well does it generalize to unseen objects and environments? (2) What additional emergent generalization capabilities do Interleave-VLA demonstrate? (3) Does Interleave-VLA have the potential for scaling?

A. Experiment Setup and Tasks

Environments. We conduct comprehensive experiments of interleave VLAs against their text-only counterparts in

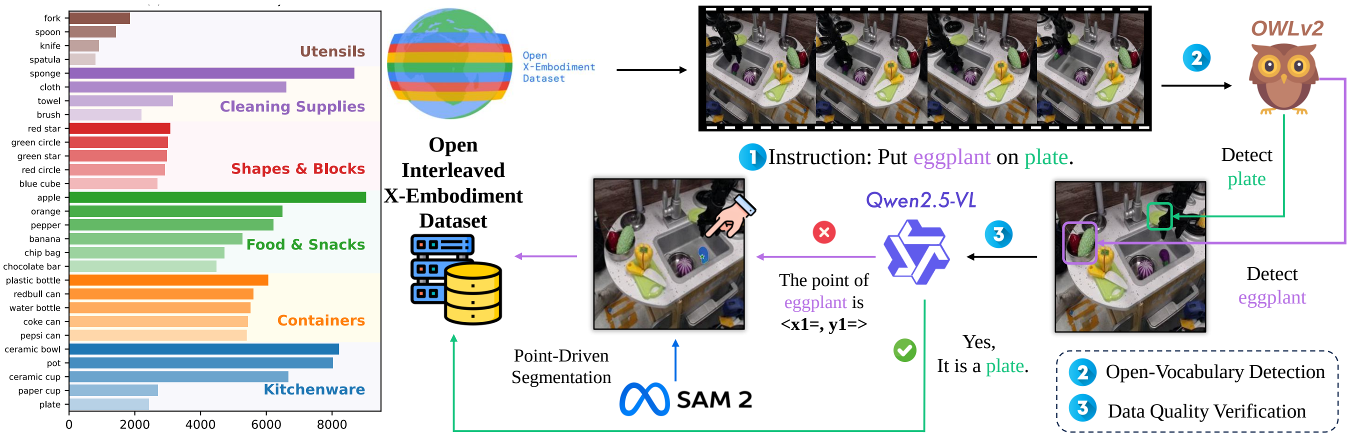

Fig. 2: **Left:** Our open interleaved X-Embodiment dataset features a large number of high-quality cropped images with diversity across objects. **Right:** Interleave dataset generation pipeline: (1) Instruction parsing: use LLM to extract key objects from language instructions. (2) Open-vocabulary detection: use OWLv2 to locate and crop target objects from trajectory frames based on the parsed instruction keywords. (3) Data quality Verification: use QwenVL to verify the detected objects, and if needed, provide keypoints for more precise segmentation using Segment Anything.

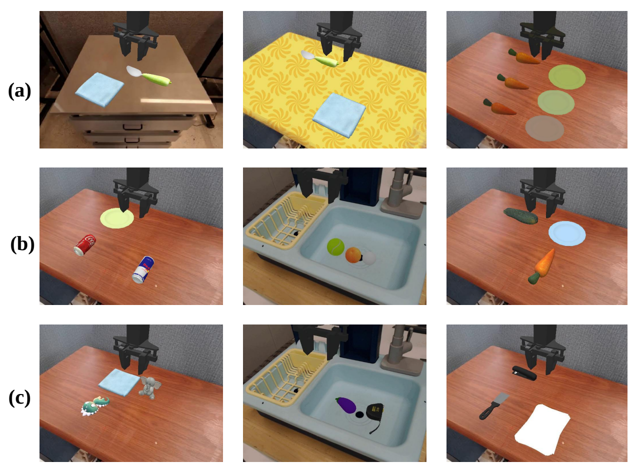

Fig. 3: Illustration of generalization settings in SIMPLER. (a) Visual generalization: unseen environments, tablecloths, and lighting conditions. (b) Semantic generalization with novel objects from known categories. (c) Semantic generalization with objects from entirely new categories not seen during training.

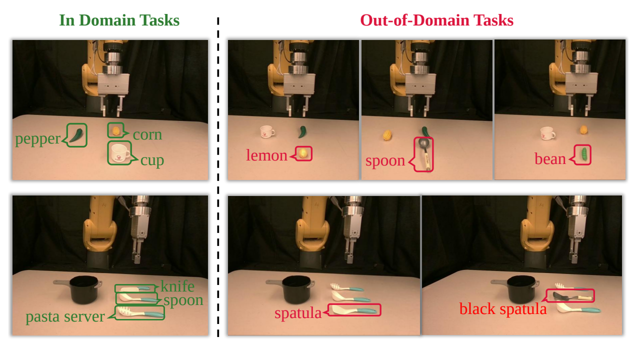

Fig. 4: Real-world generalization experiments. In-Domain and Out-of-Domain settings in the real world on a FANUC LRMate 200iD/7L robotic arm.

both simulator-based evaluation and real robot evaluation. We use SIMPLER [47] and VIMA-Bench [16] as our simulation environments. **SIMPLER** is designed to closely match real-world tasks and bridge the real-to-sim gap. We adapted SIMPLER to support interleaved image-text instructions, allowing us to evaluate the performance of Interleave-VLA models in a realistic setting. The interleaved instruction is generated automatically by our pipeline in Section III-C. **VIMA-Bench** is designed to experiment with interleaved instruction following abilities that natively focus on evaluation of planner-based tasks, where models are evaluated on object recognition and multi-task understanding. We also conduct **real robot** experiments on FANUC LRMate 200iD/7L robotic arm outfitted with an SMC gripper.

Tasks. For **SIMPLER**, we evaluate on the Visual Matching setup on the WidowX robot. This setup is designed to test the model's in-domain capability by closely matching the real-world training and simulated evaluation distributions. To comprehensively evaluate generalization, we design two main categories of tasks following [48]: *visual generalization* and *semantic generalization*. *Visual generalization* assesses robustness to novel tablecloth, lighting, and environments. *Semantic generalization* assesses the model's ability to correctly identify and manipulate target objects in the presence of diverse distractors. This evaluation is further divided into two categories: (1) novel objects from previously seen categories, and (2) objects from entirely new, unseen categories. See Figure 3 for an overview. For **VIMA-Bench**, in addition to the original tasks, we introduce three new tasks to demonstrate that the Interleave model can effectively interpret *sketch-based instructions*—a user-friendly approach for human-robot interaction [49]. For **real robot** experiments, we evaluate two different manipulation tasks: (1) "Pick up pepper/corn/cup" with generalization to "bean/lemon/cup", and (2) "Put pasta server/spoon/knife into pot" with generalization to "spatula/black spatula". Refer to Figure 4 for the experimental setup.

B. Simulation Performance

For **SIMPLER**, we adapt the state-of-the-art VLA model π_0 into Interleave-VLA to support interleaved instructions. Interleave-VLA and other baselines are trained on the full Bridge Data V2 [42] for fair comparison, with Interleave-VLA using the interleaved version. Our results demonstrate that interleaved instructions not only enhance performance on standard in-domain tasks, but more importantly, enable 2-3$\times$ stronger generalization to semantically out-of-domain tasks. To explore the benefits of interleaved cross-embodiment dataset, we present a co-trained version of Interleave-VLA using our Open Interleaved X-Embodiment Dataset. Although Bridge Dataset V2 is already large and diverse, making significant improvements challenging, additional gains are observed in semantic generalization, confirming that cross-embodiment skill transfer emerges with interleaved training. Detailed results are provided in Table I.

In **VIMA-Bench**, we adapt another SOTA VLA model OpenVLA into Interleave-VLA to support interleaved instructions, demonstrating the broad applicability of our approach. We benchmark Interleave-VLA against end-to-end VLA models (Gato, Flamingo, GPT) adapted for interleaved instruction inputs. Our results show that Interleave-VLA consistently outperforms the original OpenVLA across all levels of generalization, achieving over **2$\times$ higher performance on average**. Beyond the standard VIMA-Bench tasks, we introduce three new tasks utilizing sketches for both training and evaluation, further highlighting the flexibility of Interleave-VLA in handling diverse instruction modalities. Note that VIMA is not included in comparison, as it relies on a separately trained detector to provide bounding boxes, which are unavailable to end-to-end VLA models.

C. Real robot Performance

For **real robot** experiments, we evaluate two object sets, collecting 20 teleoperated demonstrations per object using a space mouse. As shown in Table II, our adapted Interleave-VLA from π_0 achieves **2-3$\times$** higher out-of-domain performance compared to the text-only π_0. Unlike the SIMPLER experiments, where training on large-scale Bridge Data V2 enables strong performance out-of-the-box, the FANUC robot experiments are limited to a much smaller dataset. In this low-data regime, directly training π_0 yields poor results. However, pretraining on our Open Interleaved X-Embodiment Dataset enables strong cross-embodiment transfer, significantly boosting performance. This emergent transfer ability with interleaved image-text instructions is consistent with previous findings for text-only instructions [12]. Such strong cross-embodiment transfer is important, as it reduces the need for costly and time-consuming large-scale demonstration collection.

D. Analysis of Interleave-VLA's Generalization and Emergent Capabilities

1) Task Flexibility and Emergent Generalization of Interleave-VLA: In diverse manipulation tasks, interleaved format introduced by VIMA [16] offers a unified sequence-based interface. As shown in Figure 5, our Interleave-VLA effectively handles VIMA-Bench tasks including goal image matching and multi-step instruction following (e.g., Task 4 and Task 11), where multiple goal images must be processed in order. These results confirm the flexibility and effectiveness of image-text interleaved instructions for general robotic manipulation.

Next, we evaluate the generalization capabilities of the interleaved format in real-world scenarios, moving beyond the clean simulation environment and high-level SE(2) action space of VIMA-Bench to SIMPLER and real-robot experiments. Our results (Table I and II) consistently show that Interleave-VLA delivers substantially stronger generalization than text-only baselines in diverse tasks, especially in challenging out-of-domain scenarios with unseen objects and distractors.

Notably, Interleave-VLA exhibits a remarkable **emergent capability**: it enables users to flexibly specify instructions in a completely **zero-shot manner**, without requiring any additional finetuning on unseen input modalities. Table III demonstrates the examples of image instruction types and their corresponding high performance. Instructions can be in diverse formats, including: (1) **Cropped Image Instructions:** Users can directly crop a region from the screen to indicate the target object. (2) **Internet Image Instructions:** Users may supply any image—such as a photo retrieved from the Internet—to represent the desired object. (3) **Hand-Drawn Sketch Instructions:** Users can draw sketches or cartoons about the objects.

The interleaved instruction format naturally accommodates these diverse inputs, thereby enhancing the intuitiveness of human-robot interaction and removing the need to explicitly name, categorize or describe objects with precise texts. The strong performance gains in both in-domain and out-of-domain tasks underscore the importance of interleaved image-text instructions for building more adaptable and practical robotic systems.

2) Interleave-VLA Training: Importance of Interleave Diversity: Interleave-VLA achieves stronger generalization than standard VLA models thanks to multimodal learning from image-text interleaved format. This is directly reflected by our experimental results in both simulation (Section IV-B) and real world (Section IV-C). We identify two key factors driving this zero-shot generalization: (1) training dataset scale and diversity (2) prompt image diversity.

Our experiments demonstrate that both the scale and diversity of the training dataset are critical for strong Interleave-VLA performance, particularly in out-of-domain generalization. When the in-domain dataset is limited (e.g., real-robot experiments; see Table II), pretraining on a large-scale dataset is essential—models without such pretraining exhibit significantly worse performance. When the in-domain dataset is large and diverse (e.g., SIMPLER; see Table I) where further improvement is expected to be more challenging, incorporating cross-embodiment data can still further improve semantic generalization and enhance out-of-domain

TABLE I: Benchmark results on **SimplerEnv**. Tasks T1–T4 are **In-Domain** Visual Matching setup. We add 3 **Out-of-Domain** evaluation suites, namely: Visual, Semantic L1, and Semantic L2 corresponding to (a), (b), and (c) respectively in Figure 3. Interleave-VLA performs better than its text counterpart by over 2.5x in Out-of-Domain tasks. Co-training with other datasets in our Open Interleaved X-Embodiment Dataset further boosts performance in semantic generalization tasks. We use **bold** and underline to represent the 1^{st} and 2^{nd} highest numbers.

Model Name	In Domain				Out-of-Domain			
	T1: Carrot	T2: Eggplant	T3: Spoon	T4: Stack	Visual	Semantic L1	Semantic L2	AVG
RT-1-X [12]	4.2	0.0	0.0	0.0	0.0	4.0	6.1	3.4
Octo [50]	12.5	41.7	15.8	0.0	12.6	10.8	8.4	10.6
π_0 [51]	52.5	87.9	**83.8**	**52.5**	71.4	26.7	21.0	39.7
Interleave-VLA	**57.5**	94.2	80.8	51.6	**73.4**	63.7	53.0	63.4
Interleave-VLA co-trained	57.1	**95.8**	80.5	42.1	71.5	**70.7**	**57.3**	**66.5**

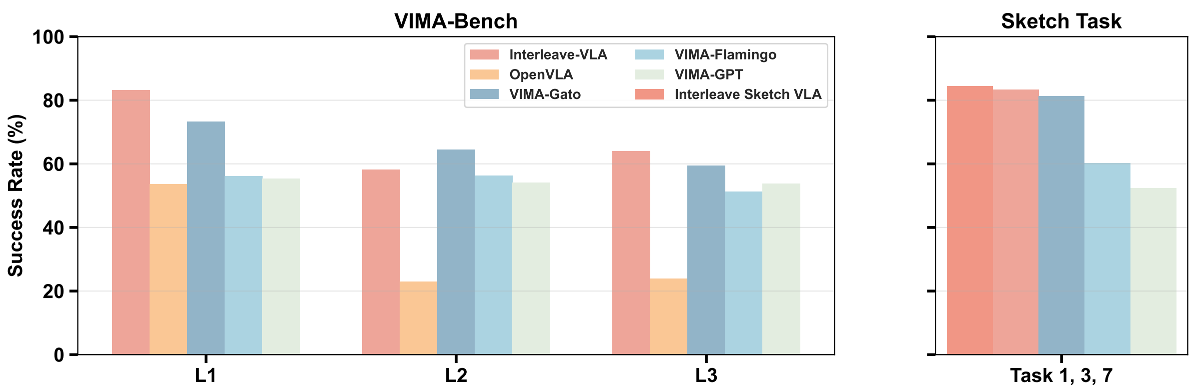

Fig. 5: VIMA-Bench results across three levels of task generalization: L1 (placement), L2 (combinatorial), and L3 (novel object). Interleave-VLA consistently outperforms OpenVLA at all levels, demonstrating stronger generalization. We also introduce sketch-based tasks to highlight the flexibility of image-text interleaved instructions.

TABLE II: Comparison of success rates (Succ) and correct object picking rates (Acc) in real-robot experiments. Interleave-VLA adapted from π_0 achieves **2-3× higher out-of-domain performance** compared to π_0. "PT" indicates pretraining on our interleaved dataset built in Section III-C. Notably, although the pretraining dataset does not include FANUC robot arm data, it still enables strong **cross-embodiment transfer** to FANUC.

Model Name	In-Domain						Out-of-Domain					
	pepper		corn		cup		bean		lemon		spoon	
	Succ.	*Acc.*	*Succ.*	*Acc.*	*Succ.*	*Acc.*	*Succ.*	*Acc.*	*Succ.*	*Acc.*	*Succ.*	*Acc.*
Interleave-VLA w/o PT	17	33	0	33	0	33	0	40	0	33	0	17
π_0 w/ PT	58	83	33	**100**	25	**100**	8	8	17	42	**75**	**92**
Interleave-VLA w/ PT	**58**	**100**	**75**	**100**	**67**	**100**	**75**	**100**	**67**	**100**	**75**	**92**

Model Name	pasta server		spoon		knife		spatula		black spatula	
	Succ.	*Acc.*	*Succ.*	*Acc.*	*Succ.*	*Acc.*	*Succ.*	*Acc.*	*Succ.*	*Acc.*
Interleave-VLA w/o PT	33	67	8	58	17	58	0	67	0	50
π_0 w/ PT	**58**	**83**	**58**	75	**33**	58	8	8	33	42
Interleave-VLA w/ PT	50	67	**58**	**83**	**33**	**58**	**25**	**100**	**50**	**67**

robustness. It suggests that cross-embodiment co-training benefits Interleave-VLA, aligning with results from Open X-Embodiment. Overall, our findings highlight the critical role of our large-scale Open Interleaved X-Embodiment Dataset in enabling robust and generalizable Interleave-VLA models across varying scale in-domain data regimes.

For prompt image diversity, Table IV demonstrates that combining Internet images with task-specific images cropped from robot observations yields the best overall performance. Using only Internet images leads to lower in-domain accuracy due to limited task relevance, while relying solely on cropped images improves in-domain results but lacks diversity. Mixing both sources provides complementary advantages, resulting in enhanced accuracy and stronger gen-

TABLE III: Interleave-VLA unlocks powerful **zero-shot** generalization to diverse instruction modalities, including hand-drawn sketches, user-cropped images, and Internet photos, **without ever seeing them in training dataset**. The consistently high accuracy demonstrates that Interleave-VLA can robustly interpret and execute visually grounded instructions, showing strong potential for flexible and practical human-robot interaction.

Task	Prompt A	A Succ. (%)	A Acc. (%)	Prompt B	B Succ. (%)	B Acc. (%)
		58.3	90.0		48.8	86.0
		75.8	100		58.8	100
		71.7	100		80.8	100
		70.0	96.0		73.3	100
		69.6	100		76.3	100
		75.5	100		71.7	100

TABLE IV: Ablation study on prompt image diversity for Interleave-VLA on SIMPLER. "In-Domain" reports the average performance on SIMPLER Visual Matching; "Out-of-Domain" averages results on one unseen instruction from Table III and one unseen object from Figure 3. Combining both task-specific and Internet images as prompts achieves the best overall performance.

Prompt Type	In-Domain	Out-of-Domain
Internet Only	59.2	69.1
Task-specific Only	67.5	67.1
Mixed	**71.0**	**71.7**

eralization.

V. Conclusion

We present Interleave-VLA, a simple and effective paradigm for adapting existing VLA models to handle image-text interleaved instructions. To overcome the lack of real-world interleaved datasets, we develop an automatic pipeline that generates a large-scale dataset with 210k episodes and 13 million frames from Open X-Embodiment. With minimal modifications to current VLA models, Interleave-VLA achieves 2–3x improvement in generalization across both simulation and real-world experiments, substantially reducing the OOD safety risks of purely textual commands. Furthermore, our approach demonstrates strong emergent zero-shot generalization to diverse user instructions never seen during training—including hand-drawn sketches, cropped images, and Internet photos—making it both practical and flexible for real-world robotic applications.

Limitations. While Interleave-VLA achieves strong generalization, training with interleaved inputs is more computationally demanding due to the increased length of image tokens and often requires more training steps to converge. Future work could focus on compressing image tokens to improve efficiency. Additionally, building a true robotic foundation model may require supporting interleaved outputs as well as inputs. Recent studies [14], [52] indicate that generating text or future images alongside actions can further enhance VLA performance. Therefore, developing unified VLA models with interleaved input and output is a promising direction.

References

[1] 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.

[2] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[3] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[4] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, *et al.*, "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*, 2024.

[5] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[6] C. Team, "Chameleon: Mixed-modal early-fusion foundation models," *arXiv preprint arXiv:2405.09818*, 2024.

[7] 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.

[8] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu, *et al.*, "Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling," *arXiv preprint arXiv:2412.05271*, 2024.

[9] G. Luo, X. Yang, W. Dou, Z. Wang, J. Liu, J. Dai, Y. Qiao, and X. Zhu, "Mono-internvl: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2025.

[10] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," *arXiv preprint arXiv:2307.15818*, 2023.

[11] 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*.

[12] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain, *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 6892–6903.

[13] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, *et al.*, "π0: A vision-language-action flow model for general robot control, 2024," *URL https://arxiv.org/abs/2410.24164*.

[14] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, *et al.*, "$\pi_{0.5}$: a vision-language-action model with open-world generalization," *arXiv preprint arXiv:2504.16054*, 2025.

[15] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via ac-

tion diffusion," *The International Journal of Robotics Research*, p. 02783649241273668, 2023.

[16] Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, and L. Fan, "Vima: Robot manipulation with multimodal prompts," PMLR, pp. 14 975–15 022, 2023.

[17] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv e-prints*, pp. arXiv–2212, 2022.

[18] G. R. Team, S. Abeyruwan, J. Ainslie, J.-B. Alayrac, M. G. Arenas, T. Armstrong, A. Balakrishna, R. Baruch, M. Bauza, M. Blokzijl, *et al.*, "Gemini robotics: Bringing ai into the physical world," *arXiv preprint arXiv:2503.20020*, 2025.

[19] 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.

[20] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[21] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao, "Eva: Exploring the limits of masked visual representation learning at scale," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 19 358–19 369.

[22] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.

[23] F. Li, R. Zhang, H. Zhang, Y. Zhang, B. Li, W. Li, Z. Ma, and C. Li, "Llava-interleave: Tackling multi-image, video, and 3d in large multimodal models," in *The Thirteenth International Conference on Learning Representations*.

[24] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu, *et al.*, "Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling," *arXiv preprint arXiv:2412.05271*, 2024.

[25] D. Jiang, X. He, H. Zeng, C. Wei, M. Ku, Q. Liu, and W. Chen, "Mantis: Interleaved multi-image instruction tuning," *Transactions on Machine Learning Research*.

[26] Y. Fang, Y. Yang, X. Zhu, K. Zheng, G. Bertasius, D. Szafir, and M. Ding, "Rebot: Scaling robot learning with real-to-sim-to-real robotic video synthesis," *arXiv preprint arXiv:2503.14526*, 2025.

[27] 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.

[28] 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 *International Conference on Machine Learning*. PMLR, 2024, pp. 61 229–61 245.

[29] W. Zhao, P. Ding, Z. Min, Z. Gong, S. Bai, H. Zhao, and D. Wang, "Vlas: Vision-language-action model with speech instructions for customized robot manipulation," in *The Thirteenth International Conference on Learning Representations*.

[30] Y. Jin, K. Xu, L. Chen, C. Liao, J. Tan, Q. Huang, C. Bin, C. Song, D. ZHANG, W. Ou, *et al.*, "Unified language-vision pretraining in llm with dynamic discrete visual tokenization," in *The Twelfth International Conference on Learning Representations*.

[31] S. Karamcheti, S. Nair, A. Balakrishna, P. Liang, T. Kollar, and D. Sadigh, "Prismatic vlms: Investigating the design space of visually-conditioned language models," in *Forty-first International Conference on Machine Learning*, 2024.

[32] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello, *et al.*, "Paligemma: A versatile 3b vlm for transfer," *CoRR*, 2024.

[33] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, *et al.*, "Qwen2. 5 technical report," *arXiv preprint arXiv:2412.15115*, 2024.

[34] M. Honnibal, "spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing," *(No Title)*, 2017.

[35] R. Shah, R. Martín-Martín, and Y. Zhu, "Mutex: Learning unified policies from multimodal task specifications," in *7th Annual Conference on Robot Learning*, 2023. [Online]. Available: https://openreview.net/forum?id=PwqiqaaEzJ

[36] M. Minderer, A. Gritsenko, and N. Houlsby, "Scaling open-vocabulary object detection," *Advances in Neural Information Processing Systems*, vol. 36, pp. 72 983–73 007, 2023.

[37] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[38] L. Y. Chen, S. Adebola, and K. Goldberg, "Berkeley UR5 demonstration dataset," https://sites.google.com/view/berkeley-ur5/home.

[39] S. Saxena, M. Sharma, and O. Kroemer, "Multi-resolution sensing for real-time control with vision-language models," in *Conference on Robot Learning*. PMLR, 2023, pp. 2210–2228.

[40] S. Belkhale, Y. Cui, and D. Sadigh, "Hydra: Hybrid robot actions for imitation learning," in *Proceedings of the 7th Conference on Robot Learning (CoRL)*, 2023.

[41] H. Liu, S. Nasiriany, L. Zhang, Z. Bao, and Y. Zhu, "Robot learning on the job: Human-in-the-loop autonomy and learning during deployment," in *Robotics: Science and Systems (RSS)*, 2023.

[42] H. Walke, K. Black, A. Lee, M. J. Kim, M. Du, C. Zheng, T. Zhao, P. Hansen-Estruch, Q. Vuong, A. He, V. Myers, K. Fang, C. Finn, and S. Levine, "Bridgedata v2: A dataset for robot learning at scale," in *Conference on Robot Learning (CoRL)*, 2023.

[43] S. Dass, J. Yapeter, J. Zhang, J. Zhang, K. Pertsch, S. Nikolaidis, and J. J. Lim, "Clvr jaco play dataset," 2023. [Online]. Available: https://github.com/clvrai/clvr_jaco_play_dataset

[44] G. Yan, K. Wu, and X. Wang, "ucsd kitchens Dataset," August 2023.

[45] 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 *Conference on Robot Learning*. PMLR, 2022, pp. 991–1002.

[46] C. Lynch, A. Wahid, J. Tompson, T. Ding, J. Betker, R. Baruch, T. Armstrong, and P. Florence, "Interactive language: Talking to robots in real time," *IEEE Robotics and Automation Letters*, 2023.

[47] X. Li, K. Hsu, J. Gu, O. Mees, K. Pertsch, H. R. Walke, C. Fu, I. Lunawat, I. Sieh, S. Kirmani, *et al.*, "Evaluating real-world robot manipulation policies in simulation," in *8th Annual Conference on Robot Learning*.

[48] A. Stone, T. Xiao, Y. Lu, K. Gopalakrishnan, K.-H. Lee, Q. Vuong, P. Wohlhart, S. Kirmani, B. Zitkovich, F. Xia, *et al.*, "Open-world object manipulation using pre-trained vision-language models," in *7th Annual Conference on Robot Learning*.

[49] P. Sundaresan, Q. Vuong, J. Gu, P. Xu, T. Xiao, S. Kirmani, T. Yu, M. Stark, A. Jain, K. Hausman, *et al.*, "Rt-sketch: Goal-conditioned imitation learning from hand-drawn sketches," in *8th Annual Conference on Robot Learning*, 2024.

[50] O. M. Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu, *et al.*, "Octo: An open-source generalist robot policy," *arXiv preprint arXiv:2405.12213*, 2024.

[51] A. Zren, "open-pi-zero," https://github.com/allenzren/open-pi-zero, 2025.

[52] Q. Zhao, Y. Lu, M. J. Kim, Z. Fu, Z. Zhang, Y. Wu, Z. Li, Q. Ma, S. Han, C. Finn, *et al.*, "Cot-vla: Visual chain-of-thought reasoning for vision-language-action models," *arXiv preprint arXiv:2503.22020*, 2025.

[53] 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.

[54] 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," *arXiv preprint arXiv:2410.07864*, 2024.

[55] L. X. Shi, Z. Hu, T. Z. Zhao, A. Sharma, K. Pertsch, J. Luo, S. Levine, and C. Finn, "Yell at your robot: Improving on-the-fly from language corrections," *CoRR*, 2024.

APPENDIX

A. Interleave-VLA Implementation Details

We extend two state-of-the-art VLA models, π_0 [13] and OpenVLA [11], to develop Interleave-VLA. While VLA models encompass a wide range of architectures [14], [18], [53], [54], [55], [17], [10], [50], [15], we focus on those based on VLM backbones due to their inherent ability to process image-text pairs. However, our approach is not restricted to VLM-based methods and can be extended to other sequence modeling approaches for action prediction [15], [50], [54], [17]. The key modification involves interleaving image and text embeddings within the input sequence. Investigating the feasibility of this modification for other sequence modeling VLAs is an exciting direction for future research. In this work, we focus on and provide adaptations of Interleave-VLA from π_0 and OpenVLA in the following sections in more detail.

1) Interleave-VLA from π_0: We make minimal architectural changes to the π_0 [13] model: only the input processor. Specifically, to enable interleaved image-text instructions, we extend its tokenizer vocabulary by introducing special tokens `<BOI>` (beginning of image) and `<EOI>` (end of image). These newly added tokens are used to delineate image embeddings within the instruction sequence. Specifically, the input tokens are constructed as follows:

<BOI> <image>$_1$ ... <image>$_{256}$ <EOI> <text> <BOI> <image>$_{257}$ ... <image>$_{512}$ <EOI> <text> <BOI> <image>$_{513}$ ... <image>$_{768}$ <EOI> <text> ...

Here, each `<image>` token represents a patch embedding from the visual encoder, and the `<BOI>` and `<EOI>` tokens mark the boundaries of each interleaved image segment. This design allows the model to flexibly process multimodal instructions by alternating between image and text tokens within a unified sequence.

Our Interleave-VLA approach is both *effective* and *model-agnostic*, requiring only *minimal modifications*. Its *effectiveness* is evidenced by substantial improvements in generalization performance over π_0, achieving 2–3× gains as shown in Table I and Table II. Interleave-VLA is *model-agnostic*, seamlessly integrating into existing VLA models without requiring assumptions about the VLM. In Interleave-VLA based on π_0, the VLM backbone Paligemma [32] demonstrates compatibility despite not being pre-trained on Internet-scale interleaved image-text data. Moreover, our approach introduces only *minimal modifications*, with no architectural changes needed for the underlying VLM backbone. These facts highlight the practicality and broad applicability of Interleave-VLA for advancing multimodal robot learning.

2) Interleave-VLA from OpenVLA: While architectural changes are not required to the VLM backbone—as demonstrated in our adaptation from π_0—we further investigate whether modifying the backbone architecture affects its effectiveness. Specifically, we replace OpenVLA's original Prismatic VLM [31] backbone with InternVL2.5 [24], which inherently supports the interleaved image-text format. As shown in Figure 5, our Interleave-VLA adaptation based on OpenVLA continues to function effectively, achieving more than double the performance of the original OpenVLA. This result further highlights the model-agnostic nature of Interleave-VLA and its compatibility with diverse VLA architectures.

B. Evaluation Details

1) Evaluation on SIMPLER:

a) SIMPLER Evaluation Tasks: Our evaluation on SIMPLER [47] includes both In-Domain and Out-of-Domain tasks. The In-Domain tasks follow the original SIMPLER WidowX BridgeData V2 Visual Matching setup. Since SIMPLER tasks use text-based instructions, we adapt them into interleaved image-text instructions using the method described in Section III-C, based on the first frame of the rollout before the robot arm begins moving.

In the WidowX BridgeData V2 setup, SIMPLER does not support generalization tasks (referred to as the Variant Aggregation setup). To overcome this limitation, we introduce a set of challenging Out-of-Domain tasks inspired by the Open Vocabulary manipulation evaluations [48]. Unlike prior methods that rely on separate VLMs to detect target objects in the scene and inject this information into the robot policy, our Interleave-VLA directly leverages interleaved image-text instruction to perform these tasks without requiring additional modules. These tasks are deliberately designed to be more challenging than the original SIMPLER tasks, requiring the robot to generalize to novel objects and environments unseen during training on BridgeData V2 [42].

We describe the 13 tasks (4 In-Domain and 9 Out-of-Domain, as illustrated on the left of Figure 4) used in the SIMPLER evaluation. The Out-of-Domain tasks are introduced in the order they appear from top left to bottom right, in Figure 4.

1) **widowx spoon on towel** (In-Domain): This task is part of the original SIMPLER Visual Matching setting and is included in the BridgeData V2.
2) **widowx carrot on plate** (In-Domain): Also from the original SIMPLER Visual Matching setting, this scenario is present in the training data.
3) **widowx stack cube** (In-Domain): This stacking task is included in the original SIMPLER Visual Matching setting and present in the training data.
4) **widowx put eggplant in basket** (In-Domain): This task is part of the original SIMPLER Visual Matching setting and is present in the training data.
5) **widowx spoon on towel, unseen environment** (Out-of-Domain, Visual Generalization): The environment overlay is sourced from the RT-1 Dataset [17] and is not seen during Bridge V2 training. The robot must generalize to a novel environment.
6) **widowx spoon on towel, unseen tablecloth** (Out-of-Domain, Visual Generalization): The tablecloth overlay is a random image from the internet, unseen in Bridge V2 training data, requiring the robot to generalize to new visual backgrounds.

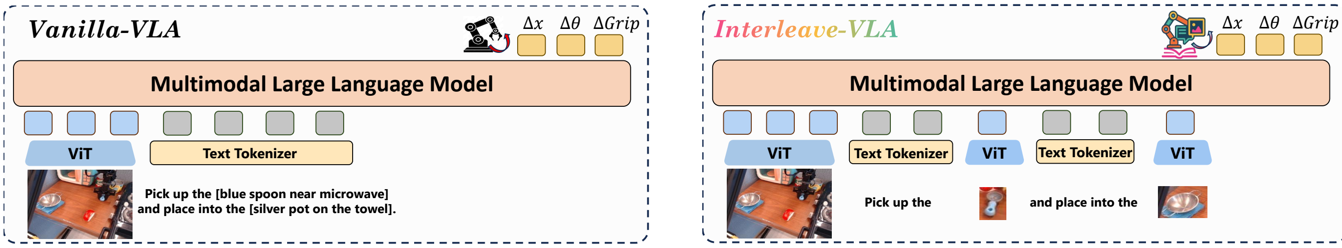

Fig. 6: Comparison of Interleave-VLA and Vanilla VLA architectures. Interleave-VLA is model-agnostic and requires minimal modifications to existing VLA architectures. The only change is the input format, which allows for interleaved image-text instructions.

7) **widowx spoon on towel, unseen lighting** (Out-of-Domain, Visual Generalization): The scene lighting changes dynamically with different colors (RGB) at 5Hz. The robot must generalize to novel and rapidly changing lighting conditions.
8) **widowx redbull on plate** (Out-of-Domain, Semantic Generalization): This is an unseen object from a known category. While similar cans (e.g., tomato can) appear in training, the Redbull can is new. The robot must use language grounding to identify and manipulate the correct object among distractors (e.g., a Coca-Cola can).
9) **widowx tennis ball in basket** (Out-of-Domain, Semantic Generalization): This is an unseen object from a known category. While similar balls (e.g., white ball, blue ball) appear in training, the tennis ball is new. The robot must use language grounding to select and manipulate the correct object among distractors (an orange and a ping pong ball).
10) **widowx zucchini on plate** (Out-of-Domain, Semantic Generalization): This task involves an unseen object from a known category. While a similar zucchini appears only once among 40,000 training episodes, this specific zucchini is entirely novel. The robot must leverage language grounding to accurately identify and manipulate the correct object, distinguishing it from distractors such as a carrot.
11) **widowx toy dinosaur on towel** (Out-of-Domain, Semantic Generalization): This is a completely unseen category. The robot must use language grounding to identify and manipulate the correct object among distractors (a toy elephant).
12) **widowx tape measure in basket** (Out-of-Domain, Semantic Generalization): This is a completely unseen category. The robot must use language grounding to identify and manipulate the correct object among distractors (a purple eggplant).
13) **widowx stapler on paper pile** (Out-of-Domain, Semantic Generalization): This task involves a completely unseen category for both the object and the destination. The robot must leverage language grounding to accurately identify and manipulate the correct object (a stapler) among distractors (e.g., a spatula) and place it onto the unseen destination, the paper pile.

b) SIMPLER Baselines: Our experiment in Table I compares Interleave-VLA (adapted from π_0) with π_0 [13], RT-1-X [17], and Octo-Base [50]. RT-1-X and Octo models are evaluated using their official checkpoints and code, following the evaluation protocol in the SIMPLER [47] repository. For π_0, we use the reimplementation from the GitHub repository [51], which is specifically trained on BridgeData V2 [42] and supports direct evaluation on SIMPLER. Interleave-VLA is built upon this reimplemented π_0 codebase, with modifications to the input tokens and training on the interleaved BridgeData V2, using the interleaved dataset construction pipeline described in Section III-C. To further highlight the benefits of large-scale, diverse, cross-embodiment data, we also co-train Interleave-VLA with our curated Open Interleaved X-Embodiment Dataset, as detailed in Section III-C.

Both Interleave-VLA (including the co-trained variant) and π_0 models were trained with a learning rate of 5e-5, a global batch size of 1024, for approximately 30 epochs. The model input consists of a single observation image (no history), interleaved image-text instruction tokens, one proprioceptive token (no history), and four action tokens. Training takes roughly 2 days on 4×H100 GPUs with a per device batch size of 16. Actions and proprioception across the diverse datasets are normalized to the 7D format: xyz position, Euler orientation, and gripper state, with all values scaled to the range $[-1,1]$.

The results presented in Table I reflect the best performance across checkpoints. Notably, performance can vary significantly between checkpoints, even among those that appear mostly converged. This variability is particularly pronounced for challenging tasks requiring precise manipulation, such as ”widowx stack cube”. These observations align with findings reported in the π_0 reimplementation GitHub repository [51].

c) SIMPLER Evaluation Results: Table V provides detailed generalization results for the top-performing models: π_0, Interleave-VLA (adapted from π_0), and Interleave-VLA co-trained, as reported in Table I. Interleave-VLA consistently surpasses π_0 across all Out-of-Domain generalization tasks, demonstrating the effectiveness of multimodal learning from interleaved image-text data for both visual and semantic generalization. The co-trained Interleave-VLA model achieves further improvements, especially on semantic generalization tasks such as "RedBull on Plate," where similar RedBull cans are present in the RT-1 dataset for the Google robot. This highlights positive cross-embodiment task transfer to the WidowX robot. Overall, these results show that training with large-scale, diverse robot data enhances model generalization to novel tasks and robot embodiments, supporting our approach of curating the Open Interleaved X-Embodiment Dataset.

Note that the Unseen Environment setting is omitted for the Interleave-VLA co-trained model because the scene overlay is sourced from the RT-1 Google Robot dataset, which is included in the co-train data. As a result, the model tends to generate actions intended for the Google Robot. During evaluation, however, the robot used is WidowX, leading to a mismatch in embodiment and causing the model to produce incorrect actions.

2) Evaluation on VIMA-Bench:

a) VIMA-Bench Evaluation Tasks: We evaluate performance on the majority of VIMA-Bench tasks, but excluding those requiring historical memory. Memory-dependent tasks are omitted because Interleave-VLA, like common VLA models [11], [12], [17], [10], [13], [18], [26], [27], is designed for memory-independent, first-order Markov settings. In general, common VLA models characterize the conditional distribution $p(\mathbf{A}_t|\mathbf{o}_t)$, where $\mathbf{A}_t = [\mathbf{a}_t, \mathbf{a}_{t+1}, \ldots, \mathbf{a}_{t+H-1}]$ represents a sequence of future actions, and $\mathbf{o}_t$ denotes the current observation (comprising multiple RGB images, a language command, and the robot's proprioceptive state). Extending VLAs to handle historical memory in interleaved instruction scenarios remains an interesting direction for future work.

VIMA-Bench employs interleaved image-text instructions for task specification. To evaluate text-instructed VLA models, we transform these interleaved instructions into text-only instructions by utilizing the shape and texture names provided in the VIMA-Bench codebase. For example:

`VIMA-Bench Instruction: Put the`  `into the`  `.`

`Transformed Instruction: Put the rainbow triangle into the blue square.`

b) VIMA-Bench Baselines: We evaluate Interleave-VLA (adapted from OpenVLA) against several baselines: OpenVLA [11], VIMA-Gato [16], VIMA-Flamingo [16], and VIMA-GPT [16]. All models are trained on the same dataset generated using an oracle model, which has access to the exact 2D poses of all objects in the scene. This dataset generation process is provided by VIMA. For OpenVLA, the training data consists of text-instructed samples. Both Interleave-VLA and OpenVLA are trained on an equivalent amount of the generated VIMA dataset using the following training hyperparameters: a constant learning rate of 2e-5 and a global batch size of 128. This comparison demonstrates the effectiveness of Interleave-VLA in improving generalization performance over existing VLA models. The results for VIMA-Gato, VIMA-Flamingo, and VIMA-GPT are taken from the original VIMA paper [16] and serve as additional benchmarks. These models, adapted by the VIMA team, serve as benchmarks to assess the progression of VLA models from earlier architectures like Gato, Flamingo, and GPT to the more advanced OpenVLA.

c) VIMA-Bench Evaluation Results: The detailed results for the memory-independent VIMA-Bench tasks are presented in Table VI. The results demonstrate that Interleave-VLA benefits significantly from interleaved image-text instructions, which enhance its ability to identify and manipulate the correct object by 2×. This approach proves more effective than relying solely on text descriptions to distinguish objects with the desired texture and shape among distractors.

3) Evaluation on real robot:

a) Real robot Evaluation Tasks: We evaluate on two distinct manipulation tasks: Lift and Pick&Place, corresponding to the first and second rows of results shown in Table II. Visual illustrations of these tasks are shown on the right side of Figure 4. The tasks are designed to be challenging, requiring the robot to generalize to novel objects not seen during training. We describe these tasks in more detail.

The Lift task includes:

1) **Lift pepper** (In-Domain): 20 demonstrations collected with varied object arrangements and positions.
2) **Lift cup** (In-Domain): 20 demonstrations collected with varied object arrangements and positions.
3) **Lift corn** (In-Domain): 20 demonstrations collected with varied object arrangements and positions.
4) **Lift lemon** (Out-of-Domain, Semantic Generalization): The target is an unseen object, as lemons are not included in the collected demonstrations. Although the lemon category appears in the pretraining data, it appears with different textures, robots, and environments. VLA models must utilize language grounding to accurately identify and lift the target lemon among two distractor items.
5) **Lift bean** (Out-of-Domain, Semantic Generalization): The target belongs to a completely unseen category, as beans are absent from both the collected demonstrations and the pretraining dataset. VLA models must rely on language grounding to correctly identify and lift the target bean among two distractor items.
6) **Lift spoon** (Out-of-Domain, Semantic Generalization): The target is an unseen object from a known category, as the demonstrations do not include this specific spoon. While the spoon category appears in the pretraining data, it is represented with different textures, robots, and environments. VLA models must leverage language grounding to accurately identify and lift the target spoon among two distractor items.

The Pick&Place task includes:

1) **Pick up kitchen cutter and place into the pot** (In-Domain): 20 demonstrations collected with varied object arrangements and positions.

TABLE V: Detailed evaluation results on 9 Out-of-Domain generalization tasks based on SIMPLER. Success rates (%) are reported for π_0, Interleave-VLA (adapted from π_0), and Interleave-VLA co-trained with our Open Interleaved X-Embodiment Dataset, covering both visual and semantic generalization. Generalization results confirm that Interleave-VLA outperforms π_0 across all tasks, with further cross-embodiment improvements from co-training.

Model	Visual Generalization			Semantic Generalization						
	Unseen Tablecloth	Unseen Environment	Unseen Lighting	Redbull on Plate	Tennis Ball in Basket	Zucchini on Plate	Toy Dinosaur on Towel	Tape Measure in Basket	Stapler on Paper Pile	Average
π_0	78.0	77.0	59.2	0.0	30.0	50.0	24.0	1.0	38.0	39.7
Interleave-VLA	**80.0**	**79.0**	**61.3**	35.0	**73.0**	**83.0**	**39.0**	53.0	**70.0**	63.4
Interleave-VLA co-trained	74.6	–	63.3	**82.5**	48.0	82.1	38.3	**64.0**	**70.0**	**66.5**

TABLE VI: Detailed VIMA-Bench results for L1, L2, and L3 level generalization evaluations. Interleave-VLA generally outperforms other VLA models and improves the generalization capacity of OpenVLA [11] by over 2×.

Model Name	task1	task2	task3	task4	task7	task11	task15	AVG
VIMA-Bench L1								
OpenVLA [11]	83	70	78	4	**92**	0	49	53.71
Interleave-VLA	**87**	**82**	81	54	82	**100**	**96**	**83.14**
VIMA-Gato	79	68	**91**	**57**	74	61	83	73.29
VIMA-Flamingo	56	58	63	48	62	66	40	56.14
VIMA-GPT	62	57	41	55	54	77	41	55.29
VIMA-Bench L2								
OpenVLA [11]	18	20	68	2	31	0	22	23.00
Interleave-VLA	36	32	75	44	26	**100**	**94**	58.14
VIMA-Gato	**56.5**	**53.5**	**88**	**55.5**	53	63	81.5	**64.43**
VIMA-Flamingo	51	52.5	61.5	49.5	**55.5**	82	42	56.29
VIMA-GPT	52	52	49.5	54.5	51	76.5	43	54.07
VIMA-Bench L3								
OpenVLA [11]	27	36	61	3	26	0	14	23.86
Interleave-VLA	**52**	55	81	53	46	**98**	**63**	**64.00**
VIMA-Gato [16]	51	**58**	**84.5**	**56.5**	49	65	52	59.43
VIMA-Flamingo [16]	49	50	66.5	47	50	66	30.5	51.29
VIMA-GPT [16]	**52**	51	55	49.5	**50.5**	82	37	53.86

2) **Pick up ladle and place into the pot** (In-Domain): 20 demonstrations collected with varied object arrangements and positions.
3) **Pick up pasta server and place into the pot** (In-Domain): 20 demonstrations collected with varied object arrangements and positions.
4) **Pick up the white and blue spatula and place it into the pot** (Out-of-Domain, Semantic Generalization): The target is an unseen object from a known category. The demonstrations do not include any spatula. While the spatula category appears in the pretraining data, it is shown with different textures, robots, and environments. VLA models must utilize language grounding to accurately identify and manipulate the target spatula among two distractor kitchenware items.
5) **Pick up the black and white spatula and place it into the pot** (Out-of-Domain, Semantic Generalization): Similar to the previous task, but the target spatula is black and white. The robot must leverage language grounding to correctly identify and manipulate the target spatula among two distractor kitchenware items.

b) Real robot Baselines: We compare Interleave-VLA (adapted from π_0) with pretraining against the following baselines: π_0 with pretraining and Interleave-VLA without pretraining. The pretraining dataset is a subset of our curated Open Interleaved X-Embodiment Dataset, as described in Section III-C. Interleave-VLA w/ PT is pretrained on this dataset and subsequently fine-tuned on the collected demonstrations from the FANUC robot arm before evaluation. For π_0 w/ PT, the same pretraining and fine-tuning protocol is applied, except the dataset is not interleaved. This setup allows for a direct comparison to evaluate the benefits of interleaved image-text instructions for generalization. The Interleave-VLA w/o PT is trained exclusively on the collected FANUC demonstrations, without exposure to the Open Interleaved X-Embodiment Dataset, enabling us to assess the impact of large-scale, diverse pretraining on performance. All models are fine-tuned with a learning rate of 5e-5, a global batch size of 128, and evaluated across several checkpoints to mitigate the performance variability noted in Appendix B.1.b.

c) Real robot Evaluation Results: Tables VII and VIII present the detailed evaluation results for the Lift and Pick&Place tasks, respectively. Interleave-VLA, adapted from π_0, is compared against π_0 and Interleave-VLA without pretraining (w/o PT). In generalization tasks, Interleave-VLA consistently outperforms π_0 in semantic generalization by 2×, highlighting the effectiveness of multimodal learning from interleaved image-text data. The results further demonstrate that pretraining on the Open Interleaved X-Embodiment Dataset significantly enhances performance across all tasks. For small-scale datasets (60 demonstrations in total per task), pretraining on the Open Interleaved X-Embodiment Dataset proves essential for achieving strong performance, as cross-embodiment pretraining enables the model to learn more robust representations and generalize effectively, even to the FANUC robot, which is not included in the pretraining data.

C. Task Flexibility and Emergent Generalization Details

To highlight the task flexibility and emergent generalization capabilities of Interleave-VLA when faced with unseen instructions, we leverage the interleaved image-text interface to evaluate its performance across diverse user input styles during deployment. The Interleave-VLA model used in this evaluation is directly taken from the SIMPLER evaluation suite (Table I and Table V) without any additional fine-tuning. A summary of Interleave-VLA's performance statistics is presented in Table III.

Below, we describe the three tasks and their corresponding prompts in the order they appear in Table III:

1) **Place {eggplant, carrot} on the plate**. Two types of instructions are provided. The first row includes a hand-drawn sketch of an eggplant and a carrot, created by a human on-the-fly. The second row features a sketch-style image of an eggplant and a carrot sourced from the Internet.
2) **Place {green, yellow} block on the towel**. Two types of instructions are included. The first row contains a hand-drawn sketch of a green and yellow block, created by a human on-the-fly. The second row features random images representing a green and yellow block, sourced from the Internet.
3) **Place {block, spoon} on the towel**. Two types of instructions are used. The first row includes a hand-drawn sketch of a block and a spoon, created by a human on-the-fly. The second row features cropped images of the desired target objects, captured from a screen by a human on-the-fly.

Interleave-VLA demonstrates remarkable emergent generalization capabilities, even when faced with diverse instruction styles such as Internet images, object crops (from a familiar input style but with unseen images), and sketches (a completely novel input style not encountered during training). These emergent capabilities go beyond the typical generalization to novel objects and environments evaluated in prior VLA models [13], [11]. They highlight Interleave-VLA's adaptability to new tasks and instruction formats, showcasing its practical flexibility in processing diverse multimodal inputs.

D. Open Interleaved X-Embodiment Dataset Details

The Open Interleaved X-Embodiment Dataset, curated as described in Section III-C for training Interleave-VLA, integrates data from 11 sources within the Open X-Embodiment Dataset. To ensure coherent training and facilitate cross-embodiment transfer, the action space across all datasets is standardized to a unified 7D pose format: xyz position, Euler orientation, and gripper state. This normalization adheres to practices established in recent VLA research [11], [13], [50]. Our dataset features an extensive variety of over 3500 diverse object categories, as depicted on the left of Figure 2. Additionally, Figure 7 highlights the wide range of skills encompassed within the dataset and Figure 8 provides a detailed breakdown of its composition and partitioning.

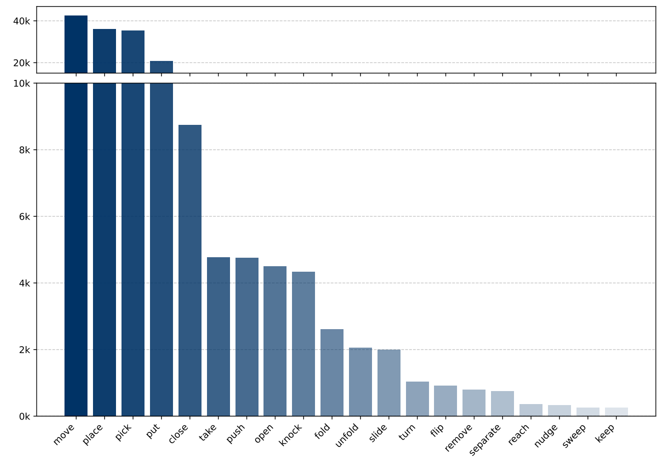

Fig. 7: Our Open Interleaved X-Embodiment Dataset is diverse in skills.

Interleaved X-Embodiment Dataset Composition	
RT-1 [17]	41.01%
Bridge [42]	28.25%
BC-Z [45]	20.34%
Language Table [46]	7.81%
UTAustin Mutex [35]	0.71%
Jaco Play [43]	0.51%
Berkeley Autolab UR5 [38]	0.47%
IAMLab CMU Pickup Insert [39]	0.30%
Stanford Hydra [40]	0.27%
UTAustin Sirius [41]	0.26%
UCSD Kitchen [44]	0.07%

Fig. 8: Composition of open data sources in our curated Open Interleaved X-Embodiment Dataset.

TABLE VII: Detailed evaluation of the ”Lift task”. We conduct 12 trials for each object and report both the number of successful trials (# Succ) and the number of trials where the correct object is manipulated (# Acc).

Category	Task	# Trials	Interleave-VLA w/ PT # Succ / # Acc	Interleave-VLA w/o PT # Succ / # Acc	π_0 w/ PT # Succ / # Acc
In-Domain	pepper	12	7/12	2/4	7/10
In-Domain	corn	12	9/12	0/4	4/12
In-Domain	cup	12	8/12	0/4	3/12
Out-of-Domain	spoon	12	9/11	0/2	9/11
Out-of-Domain	bean	12	9/12	0/4	1/1
Out-of-Domain	lemon	12	8/12	0/4	2/5
		Mean Success / Accuracy Rate	69.4 % / 98.6 %	2.8 % / 30.6 %	36.1 % / 70.8 %

TABLE VIII: Detailed evaluation on ”Pick&Place task”. We conduct 12 trials for each object and report both the number of successful trials (# Succ) and the number of trials where the correct object is manipulated (# Acc).

Category	Task	# Trials	Interleave-VLA w/ PT # Succ / # Acc	Interleave-VLA w/o PT # Succ / # Acc	π_0 w/ PT # Succ / # Acc
In-Domain	pasta server	12	6/8	4/8	7/10
In-Domain	spoon	12	7/10	1/7	7/9
In-Domain	knife	12	4/7	2/7	4/12
Out-of-Domain	spatula	12	3/8	0/8	1/1
Out-of-Domain	black spatula	12	6/8	0/6	4/5
		Mean Success / Accuracy Rate	43.3 % / 68.3 %	11.7 % / 60 %	38.3 % / 61.7 %