ChatVLA: Unified Multimodal Understanding and Robot Control with Vision-Language-Action Model

Zhongyi Zhou[1,2*], Yichen Zhu[2*†], Minjie Zhu[2], Junjie Wen[2], Ning Liu[4], Zhiyuan Xu[4], Weibin Meng[5], Yaxin Peng[3], Chaomin Shen[1†], Feifei Feng[2], Yi Xu[2]

[1]East China Normal University, [2]Midea Group, [3]Shanghai University,
[4]Beijing Innovation Center of Humanoid Robotics, [5]Tsinghua University
chatvla.github.io *

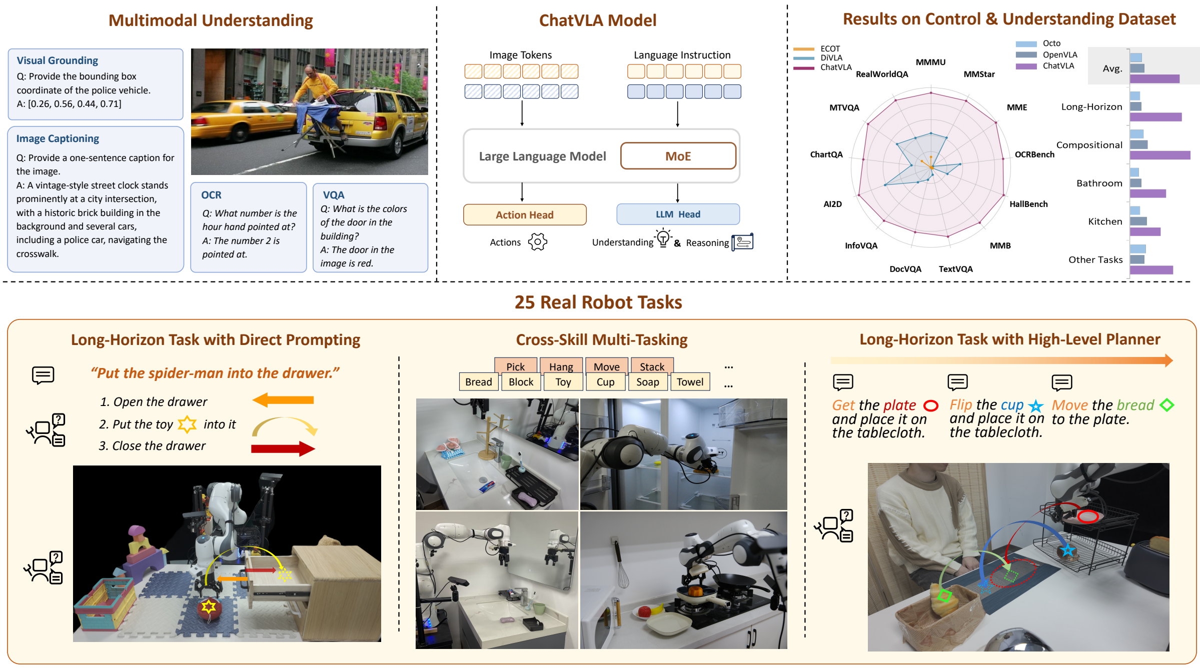

Figure 1: **ChatVLA is the first work to unify multimodal understanding and embodied control.** We conduct extensive evaluations on VQA and multimodal understanding benchmarks to demonstrate that robot foundation models can also engage in chat. Furthermore, we evaluate our approach on diverse real-world robot tasks.

Abstract

Humans possess a unified cognitive ability to perceive, comprehend, and interact with the physical world. Why can't large language models replicate this holistic understanding? Through a systematic analysis of existing training paradigms in vision-language-action models (VLA), we identify two key challenges: *spurious forgetting*, where robot training overwrites crucial visual-text alignments, and *task interference*, where competing control and understanding tasks degrade performance when trained jointly. To overcome these limitations, we propose ChatVLA, a novel framework featuring Phased Alignment Training, which incrementally integrates multimodal data after initial control mastery, and a Mixture-of-Experts architecture to minimize task interference. ChatVLA demonstrates competitive performance on visual question-answering datasets and significantly surpasses state-of-the-art vision-language-action (VLA) methods on multimodal understanding benchmarks. Notably, it achieves a six times higher performance on MMMU and scores 47.2% on MMStar with a more parameter-efficient design than ECoT. Furthermore, ChatVLA demonstrates superior performance on 25 real-world robot manipulation tasks compared to existing VLA methods like OpenVLA. Our findings highlight the potential of our unified framework for achieving both robust multimodal understanding and effective robot control.

1 Introduction

Recent advancements in Vision-Language-Action (VLA) (Black et al., 2024; Kim et al., 2024; Wen et al., 2024c, 2025) models have largely prioritized

* Co-first author. † Corresponding author. This work was done during Zhongyi Zhou's internship at Midea.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 5377–5395
November 4-9, 2025

robotic action mastery. While models trained on robotic control tasks excel at low-level manipulation and physical interaction, they often struggle to interpret and reason about multimodal data like images and text. This is paradoxical, as modern VLA architectures build upon pre-trained vision-language models (VLMs). Conversely, VLMs trained on visual-text pairs demonstrate impressive multimodal scene understanding but lack the ability to physically interact with the environment. This duality highlights a critical challenge: unifying embodied control and multimodal understanding by aligning these disparate data sources (robotic actions and visual-text semantics) without sacrificing performance in either domain.

This work investigates how to unify a single end-to-end neural network capable of multimodal scene understanding, conversational ability, and physical interaction. We first explore existing training paradigms to assess their feasibility for unification. Specifically, we examine three data settings for VLA training: 1) training solely on expert demonstration data containing robot action trajectories (the most common approach, e.g., OpenVLA (Kim et al., 2024), TinyVLA (Wen et al., 2024c), π_0 (Black et al., 2024)); 2) augmenting robot data with reasoning phrases to guide action (similar to ECoT (Zawalski et al., 2024) and DiffusionVLA (Wen et al., 2024a)); and 3) co-training with both visual-text pairs and robot data (as in RT-2 (Brohan et al., 2023a)). We analyze how each configuration impacts the model's ability to balance control and understanding. Our experiments reveal that training solely with robot data erodes conversational ability entirely; adding reasoning data partially preserves multimodal understanding; and introducing visual-text pairs significantly weakens control capabilities. This suggests two key challenges: (1) VLA models suffer from **spurious forgetting** (Zheng et al., 2025; Zhai et al., 2023; Luo et al., 2023), where performance degradation may not reflect complete knowledge loss from pre-trained VLMs, but rather a shift in how the model aligns its internal representations with different tasks. The alignment between robot actions and visual-text data appears fragile and susceptible to being overwritten during fine-tuning. (2) **Task interference** (Wang et al., 2021; Ahn et al., 2025) arises, where the conflicting parameter spaces of control and understanding tasks, sharing overlapping representations, cause mutual performance degradation when trained simultaneously.

To address these challenges, we present ChatVLA, a simple yet effective framework—in terms of both neural architecture and training strategy—for enabling a single neural network to master both understanding and manipulation. We propose Phased Alignment Training, a two-stage strategy inspired by curriculum learning. The model first masters embodied control before incrementally integrating multimodal data to "reactivate" frozen alignment links. Furthermore, we introduce a Mixture-of-Experts (MoE) on the MLP layers. This allows the two tasks to share attention layers (for cross-task knowledge transfer) while isolating task-specific MLPs (to minimize interference). This design is motivated by Dual Coding Theory, which posits that human minds process information through two separate but interconnected systems: one for physical skills and the other for verbal and visual practice. The shared attention layers in ChatVLA facilitate the exchange of mutually beneficial knowledge between understanding and control tasks, while the separate MLP layers process learned knowledge independently.

We evaluate ChatVLA across three dimensions: conversational ability (visual question answering), general multimodal understanding, and general robot control. Specifically, we assess its conversational ability on established datasets like TextVQA (Singh et al., 2019a) and DocVQA (Mathew et al., 2021), where it achieves competitive performance compared to existing VLMs. Furthermore, ChatVLA demonstrates strong multimodal understanding capabilities on general visual and textual benchmarks, including MMMU (Yue et al., 2024), MME (Fu et al., 2023), and MMStar (Chen et al., 2024a). Notably, compared to state-of-the-art VLA methods like ECoT, our method achieves a 6x performance improvement on MMMU and boosts performance on MMStar from 0 to 47.2, using 3.5x fewer parameters in the VLM backbone. Finally, we evaluate ChatVLA on 25 real-world robot tasks encompassing diverse skills like picking, placing, pushing, and hanging, across multiple environments such as bathrooms, kitchens, and tabletops. In this multi-task setting, our method outperforms state-of-the-art VLA methods like OpenVLA. These results validate the effectiveness of our approach, showcasing the potential of a single unified method for both multimodal understanding and robot control.

In summary, our contributions are the following:

- We provide an in-depth analysis of existing VLA approaches under rigorous settings, demonstrating their limitations in achieving satisfactory performance across both multimodal understanding and robot control.
- We introduce ChatVLA, a simple yet effective framework that unifies conversational ability, multimodal understanding, and robot control within a single neural network.
- We conduct extensive experiments to evaluate ChatVLA's performance on various question-answering and general understanding benchmarks.
- We perform extensive real-world robot experiments, encompassing 25 diverse tasks in realistic home environments (tabletop, kitchen, and bathroom), demonstrating ChatVLA's superior performance in real-world robot control scenarios.

2 Related Work

Multimodal understanding Multimodal Large Language Models (MLLMs) (Lu et al., 2024; Awadalla et al., 2023; Laurençon et al., 2023; Liu et al., 2023b,a; Wang et al., 2024a; Chen et al., 2024c; Zhu et al., 2024c; Ma et al., 2024; Zhou et al., 2024; Zhu et al., 2024a; Luo et al., 2024; Chen et al., 2024c; Li et al., 2023a; Dai et al., 2023; Chen et al., 2024b; Karamcheti et al., 2024) have significantly advanced the field of multimodal understanding by integrating visual and linguistic information to achieve holistic scene comprehension. MLLMs have demonstrated excellent performance on tasks requiring cross-modal alignment, such as visual question answering (VQA), image captioning, and spatial reasoning. This success stems from their ability to map visual features to semantic representations through sophisticated adapter designs. However, current MLLMs lack a connection to the physical world, preventing them from interacting with environments and humans. This work aims to bridge this gap, enabling vision-language models to also act.

Vision-langauge-action models in robot learning. Vision-language-action models (VLAs) form a growing body of research that leverages pre-trained vision-language models (VLMs) as a backbone to enable both language comprehension and observational understanding. These methods typically fine-tune large pre-trained VLMs to predict robot actions (Brohan et al., 2023b; Li et al., 2023b; Huang et al.; Wen et al., 2024c; Pertsch et al., 2025; Black et al., 2024; Kim et al., 2024; Chi et al., 2023; Zhu et al., 2024b; Wang et al., 2024b; Prasad et al., 2024; Black et al., 2023a,b; Dasari et al., 2024; Lin et al., 2024; Reuss et al., 2024; Zhao et al., 2024; Uehara et al., 2024a,b; Ding et al., 2024). These methods have shown strong performance in both simulated and real-world tasks. However, existing VLA models have not demonstrated the ability to perform true multimodal understanding. Based on our experiments, we find that these models lack this capability. In contrast, our work proposes a unified approach that enables a single network to effectively handle both multimodal understanding and robot control.

3 Methodology

This section provides a thorough discussion of our framework. Section 3.1 presents formal definitions. Section 3.2 details our motivation and empirical results demonstrating how existing vision-language-action models (VLAs) suffer from catastrophic forgetting and task interference, thus hindering the unification of multimodal understanding and robot control. Section 3.3 proposes a simple solution to address this problem.

3.1 Formal Definition

Consider two distinct scenarios: robot control and multimodal understanding. In the context of robot control, we typically construct a dataset of demonstrations $D_{robot} = \{\tau_i\}_{i=1}^{N}$, where each demonstration τ_i comprises a sequence of state-action pairs. The state s consists of an observation (image) v and an instruction (text) t, such that $s = (v, t)$. We can represent the sequence of state-action pairs as $\tau_i = \{((v_1, t_1), a_1), ((v_2, t_2), a_2), \ldots, ((v_T, t_T), a_T)\}$, where each tuple $((v_j, t_j), a_j)$ represents the state at timestep j and the corresponding action taken, and T is the length of the demonstration. These demonstrations are typically provided by a human expert.

For multimodal understanding and visual conversation tasks, we have a dataset $D_{v-t} = \{\phi_i\}_{i=1}^{M}$, where each data sample ϕ_i consists of a visual image v_i and a corresponding question (or caption) in textual form t_i, i.e., $\phi_i = \{(v_i, t_i)\}$. Here, M represents the total number of such image-text pairs. The notation $v - t$ denote visual-text data.

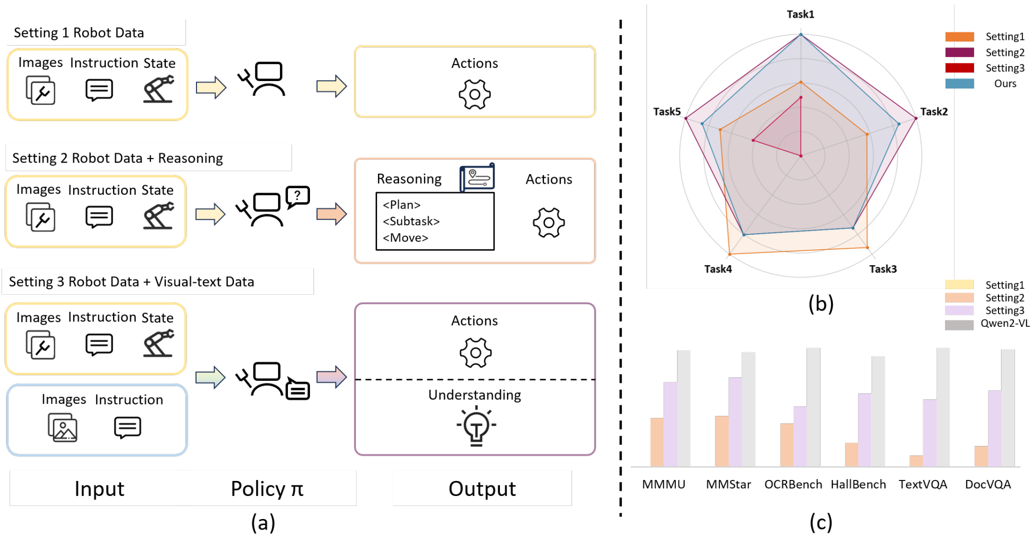

Figure 2: **Analysis of how training data influences VLA performance on control and understanding tasks.** (a) We use three different sets of training data, corresponding to the three main training approaches for VLA models. (b) The experimental results are presented for five real-world robot tasks across three settings. (c) The results on VQA and multimodal understanding benchmarks.

The overarching goal of our work is to develop a general model π capable of addressing both embodied control and multimodal understanding. For embodied control, this involves learning a policy that models the joint distribution of robot actions given the current visual observation and textual instruction: $\pi(a_t|v_t, t_t)$. Simultaneously, for multimodal understanding and visual question answering, the model should capture the distribution of the text (answer or caption) given the visual input: $\pi(t|v)$. Our objective is to create a unified model that can effectively learn both distributions, enabling it to perform well in both robot control tasks and multimodal understanding scenarios.

Current VLA research focuses on developing more robust and generalizable models for learning visuomotor policies (Kim et al., 2024; Black et al., 2024; Wen et al., 2024c). Some approaches explore chain-of-thought-like reasoning to improve policy generation (Zawalski et al., 2024; Wen et al., 2024a; Li et al., 2024), while others investigate co-training VLA models with visual-textual and robot data (Pertsch et al., 2025). In particular, some studies report benefits from co-training with visual-textual data in laboratory settings (Brohan et al., 2023a), while others find it less effective in real-world scenarios (Zawalski et al., 2024). Although a few works suggest that VLA can maintain conversational ability (Wen et al., 2024a; Brohan et al., 2023a), none have thoroughly investigated how this ability, along with general multimodal understanding, is preserved after applying the VLA training paradigm. In the following section, we analyze different training data setups for VLA, focusing specifically on the resulting model's performance in both multimodal understanding and real-world robot control. Our goal is to provide practical guidance for building unified models capable of both.

3.2 Analysis

To understand the capabilities of existing VLA models in terms of multimodal understanding and embodied control, we investigate three distinct training paradigms, each utilizing a different dataset: 1) training solely with robot data, the most prevalent approach in VLA (Black et al., 2024; Awadalla et al., 2023; Kim et al., 2024; Wen et al., 2024c), primarily focused on optimizing robot control performance; 2) augmenting robot data with chain-of-thought-like reasoning, aiming to provide auxiliary information that improves both model generalization and robot task performance (Wen et al., 2024a; Zawalski et al., 2024); and 3) co-training with both visual-textual data and robot data. This latter paradigm was pioneered by RT-2 (Brohan et al., 2023a); however, due to proprietary data and model details, exact replication is challenging. Following RT-2, we used a 3:1 ratio of robot data to visual-text data in this experiment.

In this section, we analyze these three training data setups for VLA models. Specifically, we utilize DiffusionVLA (Wen et al., 2024a), a representative VLA model that supports both language output via autoregression and action generation via a diffusion model. We evaluate performance on six representative benchmarks: four focused on visual question answering and two providing a broader evaluation of multimodal large language models, encompassing tasks like math and OCR. Furthermore, we assess performance on five real-world robot tasks covering diverse skills, including hanging, pulling, picking, and placing. Following the methodology of DiffusionVLA, we generate robot reasoning data. For visual-textual data, we randomly sample 54k image-text pairs from LLaVA. Further details regarding experimental setup and data processing are provided in the Appendix.

Results on multimodal understanding and question-answering benchmark. The experimental results are presented in Figure 2. The bottom-right portion of the figure displays performance on six benchmarks, encompassing both visual question answering (VQA) and general understanding tasks. The top-right portion of Figure 2 shows the average success rate across a total of 112 trials conducted on five real-world robot tasks.

The bottom-right table includes results for the base model, Qwen2-VL (Wang et al., 2024a). Some results are intuitive. For example, training the model solely on robot data yields a performance of 0 across all benchmarks. This model completely loses its conversational ability, exhibiting only murmuring when asked a question. As expected, the smallest performance drop compared to the base model occurs when training uses both visual-text pairs and robot data. Interestingly, training with robot data including reasoning also boosts performance from 0 to a non-negligible level, despite the highly structured, template-driven nature of the reasoning phrases within that data. Even though the reasoning phrases are similar and structured, explicitly allowing the model to “speak out" significantly improves performance on question answering and even general understanding.

Conclusion 1. Our observations suggest that the pre-trained VLM component suffers from what appears to be catastrophic forgetting. Training solely with robot data causes the model to lose previously acquired conversational and understanding abilities. However, our experiments indicate that this isn’t necessarily a complete loss of knowledge, but rather a misalignment caused by the robot data. Training with a fixed reasoning template seems to “reactivate" the visual-text alignment, enabling the model to engage in conversation and demonstrate understanding. In Section 6, we will delve into the specific knowledge that is reactivated and discuss how future work can further bridge the gap between the base VLM and the VLA model. We term this phenomenon “spurious forgetting."

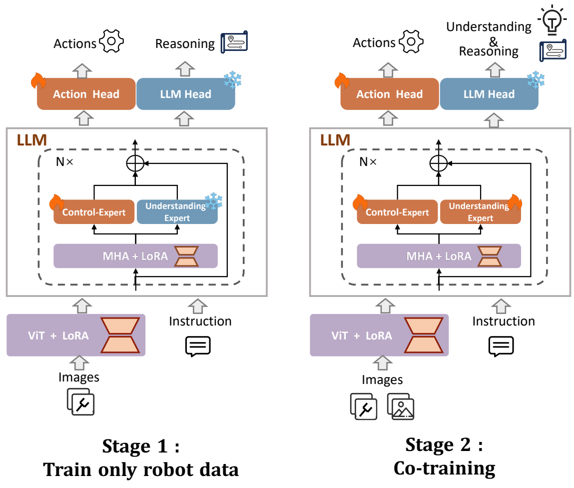

Figure 3: **Training strategy.** Our framework is initially trained on robot data with action trajectories, then co-trained with visual-text and robot data to maintain performance in both domains.

Results on real robot multi-task settings. We further evaluated different approaches to our real robot setup. All methods were trained on 25 real robot tasks, and we selected five diverse tasks, covering skills like pushing, picking, and hanging, for comparison. Details, including the number of trials for each experiment, can be found in the Appendix. Surprisingly, training with only robot data yielded worse performance than incorporating reasoning. This confirms previous findings that leveraging either visual or textual chain-of-thought enhances the generalization of robot models. Intriguingly, co-training robot data with visual-text data resulted in a significant performance drop in real-world task success rates.

Conclusion 2. The initial observation that incorporating reasoning into robot data improves performance aligns with Dual Coding Theory. This theory posits that physical motor skills and visual-linguistic understanding are not mutually exclusive but rather interconnected, offering overlapping benefits. However, the performance of robot control dramatically decreased when visual-text pairs were added to the training data. This suggests that the distinct representations required for action generation and understanding may compete within the

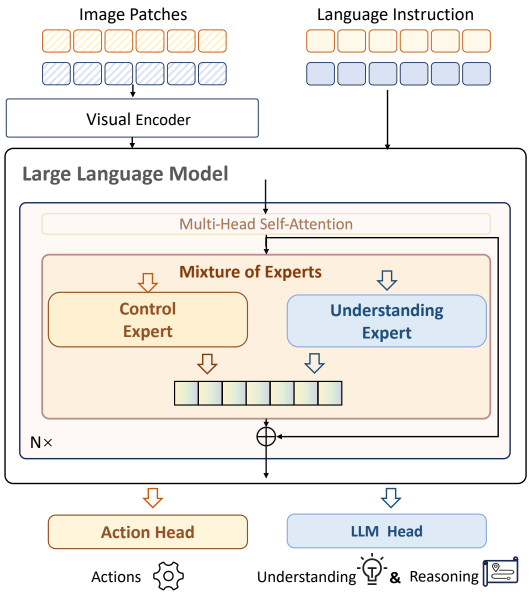

Figure 4: **Illustration of the Mixture-of-Experts component of ChatVLA.** Two distinct expert types process robot data and visual-text data separately, while shared self-attention layers facilitate knowledge transfer between the two domains.

shared parameter space. This phenomenon, we named as **partial task interference**, requires careful resolution. A unified system should connect the two data types while simultaneously enabling separable representation learning for each task.

3.3 Method: ChatVLA

As discussed above, training on robot policy data can interfere with learning of visual-text relationships. Furthermore, training exclusively on robot data can diminish visual-text alignment, leading to a degradation of the model's conversational abilities. Therefore, addressing these two challenges is crucial for successfully unifying both perspectives within a single VLA model. We will first describe the training strategy used to address spurious forgetting, and then outline the general architecture of our method to tackle the second challenge.

Phased alignment training. Previously, we identified that spurious forgetting is a key factor in causing VLA to lose its ability to chat and understand scenes. Since the pre-trained VLM is well-trained and excels at visual-related tasks, it is intuitive that the ability to chat and understand scenes can be reactivated with a small amount of visual-text pair data. In contrast, robot control tasks are much more complex to train, so the priority should be to develop an excellent model that excels at embodied control tasks. Our training strategy is straightforward yet effective. We first train the VLA model on robot data. During this training, we also include reasoning data to ensure continuous alignment between the visual and text components. Once the robot data is trained, we co-train both visual-text and robot data to help the model retain proficiency in both tasks.

Mixture of experts. The previous section demonstrated the use of phased alignment training to address the spurious forgetting problem, enabling the model to retain knowledge from the previously trained VLM. However, this approach does not fully resolve task interference issues, as the model still requires co-training on both visual-text and robot data. We introduce the mixture-of-expert to resolve the problem, which is in Figure 4. Specifically, given x^l be the input of the l-th block. The input can either belong to the D_{robot} or D_{v-l}. Notably, we design a dual router, the one to deal with tasks regarding multimodal understanding and conversational ($f(\text{FFN}_{v-l})$), and the other learn representation on robot control ($f(\text{FFN}_{robot})$). The input is first coming through a multi-head self-attention $x^{l'} = \text{MHA}(x^{l-1}) + x^{l-1}$, where $\text{MHA}(\cdot)$ represents multi-head self attention. It is then fed into the mixture-of-expert layer, which can be represented as:

$$MoE(x^{l'}) = \begin{cases} f(\text{FFN}_{v-l})(x^{l'}), & m = 0 \\ f(\text{FFN}_{robot})(x^{l'}), & 1 \leq m \leq M_r \end{cases}$$

This is then added with input from skip connection $x^l = x^{l'} + MoE(x^{l'})$. Notice that in stage 1 training, only the control expert is activated.

To differentiate task outputs, we employ distinct system prompts, such as "Answer based on question" for understanding and conversation tasks, and "Predict robot action" for control tasks. Intuitively, a static MoE architecture applied to the MLP layers can be viewed as a high-dimensional feature extractor that partitions the shared parameter space. This allows each task (e.g., understanding and control) to utilize a substantial portion of dedicated neurons, enabling the model to excel at both. A key advantage of this MoE-like architecture is that during inference, only one path is activated, preserving the model parameters of the base model. Our results demonstrate that this straightforward approach leads to simultaneous improvements in understanding, conversation, and control performance.

Why sharing self-attention layers? A prevailing solution is a use mixture of attention to learn

Table 1: **Understanding task:** Evaluation of MLLMs and VLAs on 6 Multimodal Understanding benchmarks and 7 VQA benchmarks. Boldface denotes top-ranked methods, underlined entries signify secondary performers.

Method	Params	Multimodal Understanding Benchmarks						VQA Benchmarks						
		MMMU	MMStar	MME	OCRBench	HallBench	MMB	TextVQA	DocVQA	InfoVQA	AI2D	ChartQA	MTVQA	RealWorldQA
Multimodal Large Language Models														
Janus	1.3B	30.5	37.6	1338.0	482	30.3	69.4	—	—	—	52.8	—	—	—
DeepSeek-VL	1.3B	32.2	39.9	—	409	27.6	64.6	—	—	—	51.5	—	—	—
Qwen2-VL	2B	**41.1**	**48.0**	**1872.0**	**809**	**41.7**	<u>74.9</u>	**79.7**	**88.57**	**61.37**	<u>74.7</u>	<u>73.5</u>	**18.1**	**62.9**
SmolVLM	2.3B	38.8	41.7	—	656	39.5	—	<u>72.7</u>	81.6	—	64.2	—	—	—
LLaVA-Phi	2.7B	—	—	1335.1	—	—	59.8	48.6	—	—	—	—	—	—
MobileVLM-V2	3B	—	—	1440.5	—	—	63.2	57.5	—	—	—	—	—	—
MoE-LLaVA	3.6B	—	—	1431.3	—	—	68	57	—	—	—	—	—	—
Phi-3-Vision	4.2B	<u>40.4</u>	—	—	—	—	80.5	70.9	—	—	**76.7**	**81.4**	—	—
LLaVA-1.5	7B	34.2	—	<u>1510.7</u>	—	—	64.3	58.2	—	—	63.1	55.0	—	—
DeepSeekVL	7B	36.6	—	—	456	—	73.2	—	—	—	—	—	—	—
LLaVA-Next	8B	36.4	—	—	—	—	**79.7**	55.7	—	—	66.9	65.8	—	—
Vision-Language-Action Models														
OpenVLA	7B	0	0	0	0	0	0	0	0	0	0	0	0	0
ECoT	7B	5.4	0	0	12	0.9	—	0	0	0	0	0	1.7	0
DiVLA	2B	17.2	21.1	186.5	294	9.0	—	7.5	15.2	14.7	43.1	17.2	6.2	25.2
ChatVLA(Ours)	2B	**37.4**	<u>47.2</u>	**1435.2**	<u>729</u>	<u>39.9</u>	**69.0**	**71.2**	<u>83.3</u>	<u>53.3</u>	**67.6**	**59.9**	<u>11.5</u>	<u>57.0</u>

Table 2: **Long-horizon real robot tasks with direct prompting.** *The task is completed in a sequence.* The Avg. Len. denotes the average success length of the model. Task 1: Sort toys. Task 2: Stack building blocks. Task 3: Place the toy in the drawer. Task 4: Clean building blocks to the box.

Method	Task 1					Task 2			Task 3				Task 4		
	1	2	3	4	Avg. Len.	1	2	Avg. Len.	1	2	3	Avg. Len.	1	2	Avg. Len.
Octo	0.23	0.08	0.00	0.00	0.08	0.29	0.14	0.21	0.11	0.11	0.11	0.11	0.50	0.17	0.33
OpenVLA	0.15	0.08	0.00	0.00	0.06	0.43	0.14	0.29	0.22	0.11	0.11	0.15	0.50	0.33	0.42
ChatVLA(Ours)	**0.92**	**0.69**	**0.31**	**0.23**	**0.54**	**0.86**	**0.43**	**0.64**	**1.00**	**1.00**	**1.00**	**1.00**	**0.83**	**0.67**	**0.75**

task-specific representation. However, based on our experiments (detailed in Section 4), we believe that understanding and robot control tasks share representations that are beneficial to both. For example, a typical robot control scenario requires the model to understand the scene, recognize objects, determine their locations, and then translate this information into actions. These high-dimensional representations share similar semantic concepts. Therefore, the interconnected nature of these two tasks is crucial for simultaneously improving performance on both understanding and control.

4 Experiment

In this section, we conduct a series of experiments to evaluate the performance of ChatVLA across a range of embodied control and multimodal understanding tasks.

4.1 Results on Multimodal Understanding and Visual-Question Answering

We evaluate the visual question answering abilities of ChatVLA using Vlmevalkit (Duan et al., 2024) on TextVQA (Singh et al., 2019b), DocVQA (Mathew et al., 2021), InfoVQA (Mathew et al., 2022), AI2D (Kembhavi et al., 2016), ChartQA (Masry et al., 2022), MTVQA (Tang et al., 2024), and RealworldQA (RealWorld Team, 2024). We also tested against more challenging benchmarks designed for MLLMs, i.e., MMMU (Yue et al., 2024), MMStar (Chen et al., 2024a), MME (Fu et al., 2023), OCRBench (Liu et al., 2024), HallBench (Guan et al., 2024) and MMBench (Liu et al., 2023c). As delineated in Table 1, ChatVLA demonstrates competitive performance relative to existing VLMs across multiple benchmarks. Notably, in VQA tasks, our framework achieves a notable performance of 71.2 on TextVQA, surpassing current SOTA VLAs by substantial margins. Specifically, it outperforms ECoT and DiVLA by relative improvements of 9.2x and 9.5x over these baseline models. The model exhibits particularly strong capabilities in multimodal reasoning tasks requiring complex cross-modal integration. On the MMStar benchmark, ChatVLA attains a score of 37.4, demonstrating 2.2x and 6.9x performance enhancements over DiVLA and ECoT respectively.

4.2 Results on Real Robot Tasks

The embodied control performance of ChatVLA is evaluated on 25 realworld manipulation tasks. All these evaluated tasks can be divided into three categories according to the granularity of the language instructions. A more detailed description of these tasks can be found in the Appendix (Section 6). We

Table 3: **Long-horizon real robot tasks with high-level policy model.** *The task is completed in a sequence.* The Avg. Len. denotes the average success length of the model. Task 5-8: Move the block to the basket then put the toy into the drawer. Task 9-10: Move two blocks to the basket sequentially. Task 11-13: Prepare the breakfast for me.

Method	Task 5-8					Task 9-10			Task 11-13			
	1	2	3	4	Avg	1	2	Avg	1	2	3	Avg
Octo	0.42	0.25	0.17	0.08	0.23	0.33	0.22	0.28	0.15	0.08	0.00	0.08
OpenVLA	0.42	0.33	0.33	0.17	0.31	0.44	0.22	0.33	0.23	0.08	0.00	0.10
ChatVLA(Ours)	**1.00**	**0.92**	**0.92**	**0.92**	**0.94**	**0.89**	**0.78**	**0.83**	**0.69**	**0.54**	**0.54**	**0.59**

Table 4: **Real robot multi-tasking.** We evaluated our model in a multi-task setting across diverse scenes, including bathrooms, kitchens, and tabletops. These tasks also encompassed a range of skills.

Method	Bathroom				Kitchen		Tabletop						Avg
	Task 14	Task 15	Task 16	Task 17	Task 18	Task 19	Task 20	Task 21	Task 22	Task 23	Task 24	Task 25	
Octo	3/11	0/6	1/9	0/7	0/11	3/11	1/7	2/9	1/7	2/13	2/9	3/7	18/107
OpenVLA	2/11	0/6	2/9	1/7	1/11	4/11	2/7	1/9	1/7	4/13	0/9	2/7	20/107
ChatVLA(Ours)	**6/11**	**2/6**	**5/9**	**3/7**	**3/11**	**6/11**	**4/7**	**5/9**	**4/7**	**6/13**	**4/9**	**7/7**	**55/107**

conducted 528 trials on a real robot to evaluate the model's ability.

Long-horizon tasks with direct prompting. The model is asked to executing tasks directly from language instruction(e.g., "Sort toys"). The four tasks we evaluated were all completed within a toy scenario constructed on a desktop setup. Challenging tasks of this category include Task 1, where all toys are randomly positioned in varying poses, and Task 3, which demands the integration of three distinct skills: opening, picking, and closing. Our method demonstrates substantial advantages in executing tasks directly from high-level descriptions across all evaluated scenarios. The approach maintains consistent performance in multi-step sequences, achieving a 0.54 average success length in Task 1 (6.75x higher than Octo) and perfect success rates throughout Task 3's three-step sequence.

Long-horizon tasks with high-level planner. The model receives intermediate commands that specify the current sub-task objectives (e.g., "pick object and place to target location"). The primary challenge in this evaluation stems from the substantial variations between sub-tasks, which involve: (1) diverse object types (e.g., plates, cups, bread), (2) multiple required skills (e.g., pick-place,flip), (3) varying location heights (e.g. top/bottom shelf positions) as visually demonstrated in the bottom-right panel of Fig.1. These variations collectively create a testbed for evaluating the model's compositional reasoning capability - specifically, its capacity to integrate object manipulation, spatial reasoning, and interference adaptation. This requirement is clearly reflected in the experimental results shown in Table 3, where our method outperforms OpenVLA and Octo across all task configurations.

Table 5: **Global batch size and learning rate scheduler**

Stage	Global Batch Size	Learning Rate	Scheduler
Stage 1	128	2e-5	constant
Stage 2	128	2e-5	cosine

Cross-skill multi-tasking. These tasks require the integration of multiple manipulation skills (e.g., picking, placing, pushing, and hanging) across various real-world environments, specifically categorized into three test domains: bathroom scenarios (Tasks 14-17), kitchen environments (Tasks 18-19), and tabletop configurations (Tasks 20-25). As demonstrated in Table 4, ChatVLA achieves superior performance compared to both Octo and OpenVLA across all task categories. The model exhibits particularly strong performance in challenging bathroom and kitchen tasks, where robotic arm operations are constrained to a severely limited spatial range. This experimental setup inherently introduces substantial safety considerations during model evaluation, consequently establishing rigorous requirements for the operational precision and system robustness of the assessed models.

4.3 Ablation Study

How important is mixture-of-experts in VLA? This section investigates whether the MoE mechanism in VLA is crucial for enabling VLA models with both multi-modal understanding ability and robotic controlling ability. Specifically, using the exact same training configuration, we compare the baseline model with that removing the MoE module. The experimental results are presented in Table 7 and Table 8 in Appendix. As is shown in the two tables, robot control performance decreased to 14% of the original model's capability, while multi-modal understanding retained only 70% of its original performance. This stark contrast highlights the critical role of MoE in mitigating task interference, as proposed in Section 3.2.

Is a two-stage training paradigm necessary? This section investigates whether a two-stage training paradigm is necessary for achieving both effective robot control and robust multimodal understanding. Specifically, under identical training settings, we evaluate two ablated variants of our method: (1) a model trained only in the first stage using robot data, and (2) a model trained solely in the second stage using both robot data and visual-text data. The results are summarized in Table 9 and Table 10 in Appendix.

As shown, removing the second stage leads to a slight decrease in robot control performance but causes a dramatic collapse in multi-modal understanding, retaining only 25% of the full model's capability. Conversely, skipping the first stage and training the model only in the second stage results in a more pronounced degradation in robot control (dropping to 57%), while multi-modal understanding remains relatively preserved at 70%.

4.4 Implement Details

Robot setup. We utilize a 7-Degree-of-Freedom Franka Emika robot equipped with a Robotiq gripper. The robot system includes two ZED 2 cameras positioned on the left and right sides, along with a ZED Mini wrist-mounted camera. Data collection is performed using teleoperation equipment at a frequency of 15 Hz.

Data details. For visual-text data, we use LLaVA-1.5 (Liu et al., 2023a) dataset for co-training. Following the data ratio mentioned in ECoT, we use set the ratio of visual-text data to robot data as 1:3. Using robot data, we evaluated our method on 25 real-world robot tasks, including long-horizon tasks with direct prompting. The data was randomly sampled from the LLaVA fine-tuning dataset. We hypothesize that carefully curated data is crucial for mitigating spurious forgetting, a topic we plan to explore in future work. We use the image resolution of 320 × 240, with three camera views for robotic data.

Training Details. We use Qwen2-VL-2B as our VLM backbone and the set of action head follows DiVLA (Wen et al., 2024b). We train our ChatVLA using a phased alignment training, as is discussed in Section 3.3. In the first stage, we train our model on robot data of 25 tasks, only activating the control expert and its corresponding action head. In the second stage, we co-train both visual-text data and robot data. The total training cost is 320 GPU hours. Global batch size and learning rate scheduler is shown in Table 5.

5 Conclusion

Integrating embodied control and multimodal understanding in Vision-Language-Action (VLA) models is challenging, as current methods often compromise one for the other. We identified key limitations: robot-only training degrades conversational ability, while visual-text co-training diminishes control performance due to spurious forgetting and task interference. To address this, we introduce ChatVLA, a unified framework combining Phased Alignment Training and a Mixture-of-Experts architecture. ChatVLA achieves competitive VQA and general understanding performance while excelling at real-world robot control (25 tasks across diverse scenes), outperforming OpenVLA and ECoT with 3.5x fewer parameters. These results demonstrate that a single network can effectively harmonize multimodal reasoning, conversation, and physical interaction.

Acknowledgments

This work is supported by the Sci-Tech Innovation Initiative by the Science and Technology Commission of Shanghai Municipality (24ZR1419000), and the National Science Foundation of China (12471501).

Author Contributions

Project Leader & Advisor: Yichen Zhu
Paper Writing: Yichen Zhu, Zhongyi Zhou
Algorithm Development: Zhongyi Zhou, Yichen Zhu, Minjie Zhu, Junjie Wen
Policies Training & Evaluation: Zhongyi Zhou, Yichen Zhu
Data Collection and Processing: Zhongyi Zhou
Leadership: Yichen Zhu, Chaomin Shen

Limitations

Our work explores the unification of multimodal understanding and robot control. This is the first study on this topic, aiming to spark discussion and advance the field. However, there are several limitations. First, while we identified that spurious forgetting can be mitigated with visual-text data, it is crucial to select a representative dataset that can reactivate all misaligned visual-text links in the model. In our work, the data was randomly selected, but we believe that curating a more targeted dataset could significantly enhance model performance. Additionally, our work does not include tasks of extended duration, like those presented in Pi0 (e.g., laundry folding). Increasing the complexity of robotic tasks may complicate optimization, requiring careful refinement of both the training strategy and neural architectures.

References

Hongjoon Ahn, Jinu Hyeon, Youngmin Oh, Bosun Hwang, and Taesup Moon. 2025. Prevalence of negative transfer in continual reinforcement learning: Analyses and a simple baseline. In *The Thirteenth International Conference on Learning Representations*.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. 2024. π_0: A vision-language-action flow model for general robot control. *Preprint*, arXiv:2410.24164.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. 2023a. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*.

Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. 2023b. Zero-shot robotic manipulation with pretrained image-editing diffusion models. *arXiv preprint arXiv:2310.10639*.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. 2023a. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. 2023b. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024a. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. 2023. Diffusion policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*.

W Dai et al. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*.

Sudeep Dasari, Oier Mees, Sebastian Zhao, Mohan Kumar Srirama, and Sergey Levine. 2024. The ingredients for robotic diffusion transformers. *arXiv preprint arXiv:2410.10088*.

Pengxiang Ding, Han Zhao, Wenjie Zhang, Wenxuan Song, Min Zhang, Siteng Huang, Ningxi Yang, and Donglin Wang. 2024. Quar-vla: Vision-language-action model for quadruped robots. In *European Conference on Computer Vision*, pages 352–367. Springer.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the*

32nd ACM International Conference on Multimedia, pages 11198–11201.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. In *ICLR 2024 Workshop: How Far Are We From AGI*.

Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. 2024. Prismatic vlms: Investigating the design space of visually-conditioned language models. *arXiv preprint arXiv:2402.07865*.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. 2024. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*.

H Laurençon, L Saulnier, L Tronchon, S Bekman, A Singh, A Lozhkov, T Wang, S Karamcheti, A Rush, and D Kiela. 2023. Obelisc: An open web-scale filtered dataset of interleaved image-text documents. *arXiv preprint arXiv:2306.16527*.

Jinming Li, Yichen Zhu, Zhibin Tang, Junjie Wen, Minjie Zhu, Xiaoyu Liu, Chengmeng Li, Ran Cheng, Yaxin Peng, and Feifei Feng. 2024. Improving vision-language-action models via chain-of-affordance. *arXiv preprint arXiv:2412.20451*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. 2023b. Vision-language foundation models as effective robot imitators. *arXiv preprint arXiv:2311.01378*.

Fanqi Lin, Yingdong Hu, Pingyue Sheng, Chuan Wen, Jiacheng You, and Yang Gao. 2024. Data scaling laws in imitation learning for robotic manipulation. *Preprint*, arXiv:2410.18647.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023c. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12).

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Gen Luo, Xue Yang, Wenhan Dou, Zhaokai Wang, Jifeng Dai, Yu Qiao, and Xizhou Zhu. 2024. Mono-internvl: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training. *arXiv preprint arXiv:2410.08202*.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Liang Zhao, et al. 2024. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. *arXiv preprint arXiv:2411.07975*.

Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. 2022. Infographicvqa. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2582–2591.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. 2025. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*.

Aaditya Prasad, Kevin Lin, Jimmy Wu, Linqi Zhou, and Jeannette Bohg. 2024. Consistency policy: Accelerated visuomotor policies via consistency distillation. *arXiv preprint arXiv:2405.07503*.

RealWorld Team. 2024. RealWorldQA: A Comprehensive Real-World Question Answering Dataset.

Moritz Reuss, Ömer Erdinç Yağmurlu, Fabian Wenzel, and Rudolf Lioutikov. 2024. Multimodal diffusion transformer: Learning versatile behavior from multimodal goals. *Robotics: Science and Systems*.

Amanpreet Singh, Vivek Natarajan, et al. 2019a. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019b. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, et al. 2024. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985*.

Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. 2025. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*.

Masatoshi Uehara, Yulai Zhao, Kevin Black, Ehsan Hajiramezanali, Gabriele Scalia, Nathaniel Lee Diamant, Alex M Tseng, Tommaso Biancalani, and Sergey Levine. 2024a. Fine-tuning of continuous-time diffusion models as entropy-regularized control. *arXiv preprint arXiv:2402.15194*.

Masatoshi Uehara, Yulai Zhao, Kevin Black, Ehsan Hajiramezanali, Gabriele Scalia, Nathaniel Lee Diamant, Alex M Tseng, Sergey Levine, and Tommaso Biancalani. 2024b. Feedback efficient online fine-tuning of diffusion models. *arXiv preprint arXiv:2402.16359*.

Liyuan Wang, Mingtian Zhang, Zhongfan Jia, Qian Li, Chenglong Bao, Kaisheng Ma, Jun Zhu, and Yi Zhong. 2021. Afec: Active forgetting of negative transfer in continual learning. *Advances in Neural Information Processing Systems*, 34:22379–22391.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Yixiao Wang, Yifei Zhang, Mingxiao Huo, Ran Tian, Xiang Zhang, Yichen Xie, Chenfeng Xu, Pengliang Ji, Wei Zhan, Mingyu Ding, et al. 2024b. Sparse diffusion policy: A sparse, reusable, and flexible policy for robot learning. *arXiv preprint arXiv:2407.01531*.

Junjie Wen, Minjie Zhu, Yichen Zhu, Zhibin Tang, Jinming Li, Chengmeng Li, Zhongyi Zhou, Xiaoyu Liu, Chaomin Shen, Yaxin Peng, and Feifei Feng. 2024a. Diffusionvla: Scaling robot foundation models via unified diffusion and autoregression.

Junjie Wen, Minjie Zhu, Yichen Zhu, Zhibin Tang, Jinming Li, Zhongyi Zhou, Chengmeng Li, Xiaoyu Liu, Yaxin Peng, Chaomin Shen, et al. 2024b. Diffusion-vla: Scaling robot foundation models via unified diffusion and autoregression. *arXiv preprint arXiv:2412.03293*.

Junjie Wen, Yichen Zhu, Jinming Li, Zhibin Tang, Chaomin Shen, and Feifei Feng. 2025. Dexvla: Vision-language model with plug-in diffusion expert for general robot control. *arXiv preprint arXiv:2502.05855*.

Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Kun Wu, Zhiyuan Xu, Ran Cheng, Chaomin Shen, Yaxin Peng, Feifei Feng, et al. 2024c. Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation. *arXiv preprint arXiv:2409.12514*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. 2024. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2023. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*.

Tony Z Zhao, Jonathan Tompson, Danny Driess, Pete Florence, Seyed Kamyar Seyed Ghasemipour, Chelsea Finn, and Ayzaan Wahid. 2024. Aloha unleashed: A simple recipe for robot dexterity. In *8th Annual Conference on Robot Learning*.

Junhao Zheng, Xidi Cai, Shengjie Qiu, and Qianli Ma. 2025. Spurious forgetting in continual learning of language models. In *The Thirteenth International Conference on Learning Representations*.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. 2024. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024a. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*.

Minjie Zhu, Yichen Zhu, Jinming Li, Junjie Wen, Zhiyuan Xu, Ning Liu, Ran Cheng, Chaomin Shen, Yaxin Peng, Feifei Feng, et al. 2024b. Scaling diffusion policy in transformer to 1 billion parameters for robotic manipulation. *arXiv preprint arXiv:2409.14411*.

Minjie Zhu, Yichen Zhu, Xin Liu, Ning Liu, Zhiyuan Xu, Chaomin Shen, Yaxin Peng, Zhicai Ou, Feifei Feng, and Jian Tang. 2024c. A comprehensive overhaul of multimodal assistant with small language models. *arXiv preprint arXiv:2403.06199*.

6 Appendix

6.1 Ablation Study

Could the same architecture scale to larger LLM backbones seamlessly? We have conducted experiments on the larger Qwen2-VL-7B model. The experimental results are presented in Table 11 and Table 12. The results demonstrate that the architecture can seamlessly scale to larger model sizes, showing performance improvements in both multi-modal understanding and robot control. However, due to the limited robotic data, the performance improvement in robot control tasks is limited. We believe that with an increase in the data scale, using a larger backbone will lead to more significant performance gains.

Will this recipe suitable for different base models? We trained a new version of ChatVLA using PaliGemma-3B (Beyer et al., 2024) as our VLM backbone. The results are shown in Table 13 The results indicate that while the performance of our model is influenced to some extent by the capabilities of the backbone, it demonstrates competitive performance compared with existing VLMs across multiple benchmarks.

6.2 Discussion

Can robot data effectively enhances the model's ability in multimodal understanding? We evaluate our method on a recent robotic multimodal understanding benchmark: Embodied Reasoning QA Evaluation Dataset (ERQA), which is presented by Gemini Robotics (Team et al., 2025).

We specifically chose this dataset because 28% of the questions involve multiple images, requiring the integration of concepts across them, which makes these questions significantly challenging.

Table 6: **Performance on Embodied Reasoning Task**

Method	Embodied Reasoning QA (ERQA)
Qwen2-VL-2B	27.5
ChatVLA(2B)	**33.5**

The results show that our model improves by 6% on ERQA compared to our VLM backbone Qwen2-VL-2B, proving that the inclusion of robot data effectively enhances the model's ability in multimodal understanding. Specifically, we further analyze the results in 8 subcategories. The results show that our model achieved 37.83% accuracy in the multi-view reasoning category, far surpassing Qwen2-VL-2B's accuracy of 13.51%. This demonstrates that robot data primarily contributes to the model's multimodal understanding by improving cross-scene adaptability (e.g., varying lighting/object layouts) and multi-perspective analysis. Therefore, the model approximates the level of human-like visual understanding necessary for navigation and interaction within physical environments.

What vision-language data are preferred? In stage 2, we employed the llava-1.5 (Liu et al., 2023a) dataset for co-training, which allowed the model to achieve compatible results on both VQA and MLLM benchmarks compared to Qwen2-VL. However, we argue that the remaining performance gap is attributed to the limitations of the visual-textual data used. To explore this further, we conducted an in-depth analysis of the results between ChatVLA and Qwen2-VL on the MMMU dataset, as illustrated in Fig. 5.

The MMMU dataset is divided into six categories, and ChatVLA's performance is slightly lower than Qwen2-VL in three of them: art, medicine, and social science. A closer inspection of the results for the corresponding subcategories reveals that the performance discrepancies primarily occur in five specific domains: art theory, lab medicine, pharmacy, literature, and psychology. These fields are relatively narrow in scope and involve specialized knowledge that is difficult to obtain. Upon reviewing the composition of the llava dataset, we were surprised to find that its subdatasets, including COCO, GQA, OCR-VQA, TextVQA, and VisualGenome, lack the expert knowledge required for these domains, which likely contributed to the observed performance drop.

This finding also highlights the considerable potential of our model: with more appropriate expert data for training, we believe that we can achieve significantly better performance in multimodal understanding.

What is the appropriate ratio of visual-text data to robot data? While co-training with visual-text data, we followed the settings discussed in ECOT (Zawalski et al., 2024) and set the overall visual-text data to robot data ratio at 1:3. However, whether other data ratios are beneficial or detrimental to multimodal understanding and robot tasks still requires attention. Therefore, under the same number of steps, we modified the ratio of visual-text data to robot data in co-training to 1:1 and 3:1, respectively. The results are shown in Table 14.

Table 7: **Ablation Study of Mixture of Experts on Understanding tasks:** Evaluation on 6 Multimodal Understanding benchmarks and 7 VQA benchmarks. We use bold to denote top-ranked methods.

Method	Multimodal Understanding Benchmarks						VQA Benchmarks						
	MMMU	MMStar	MME	OCRBench	HallBench	MMB	TextVQA	DocVQA	InfoVQA	AI2D	ChartQA	MTVQA	RealWorldQA
Static MoE	**37.4**	**47.2**	**1435.2**	**729**	**39.9**	**69.0**	**71.2**	**83.3**	**53.3**	**67.6**	**59.9**	**11.5**	**57.0**
3B Dense Model	26.8	37.3	1165.6	407	27.7	37.4	44.9	57.2	35.6	49.5	46.1	1.7	55.4

Table 8: **Ablation Study of Mixture of Experts on Real Robot tasks:** The embodied control performance is evaluated on 25 real-world manipulation tasks. Task 1-4 are long-horizon real robot tasks with direct prompting. Task 5-13 are long-horizon real robot tasks with high-level planner. Task 14-25 are under multi-task setting in real family scenes, including bathrooms, kitchens, and tabletops.

Method	Long-horizon Tasks								Multi Tasks in Real Family Scenes										
	T1	T2	T3	T4	T5-8	T9-10	T11-13	T14	T15	T16	T17	T18	T19	T20	T21	T22	T23	T24	T25
Static MoE	**28/52**	**9/14**	**15/15**	**9/12**	**45/48**	**15/18**	**23/39**	**6/11**	**2/6**	**5/9**	**3/7**	**3/11**	**6/11**	**4/7**	**5/9**	**4/7**	**6/13**	**4/9**	**7/7**
3B Dense Model	4/52	2/14	2/15	3/12	11/48	4/18	10/39	1/11	0/6	1/9	0/7	0/11	2/11	1/7	0/9	0/7	2/13	0/9	3/7

Table 9: **Ablation Study of training stages on Understanding task:** Evaluation on 6 Multimodal Understanding benchmarks and 7 VQA benchmarks. We use bold to denote top-ranked methods.

Stage 1	Stage 2	Multimodal Understanding Benchmarks						VQA Benchmarks						
		MMMU	MMStar	MME	OCRBench	HallBench	MMB	TextVQA	DocVQA	InfoVQA	AI2D	ChartQA	MTVQA	RealWorldQA
✓		10.0	11.7	426	187	19.6	11.3	21.4	18.9	15.3	24.8	13.0	0.3	32.2
	✓	27.2	32.2	1032.6	265	31.2	30.7	39.3	32.3	16.6	45.2	24.1	1.7	49.8
✓	✓	**37.4**	**47.2**	**1435.2**	**729**	**39.9**	**69.0**	**71.2**	**83.3**	**53.3**	**67.6**	**59.9**	**11.5**	**57.0**

Table 10: **Ablation Study of training stages on Real Robot tasks:** The embodied control performance is evaluated on 25 real-world manipulation tasks. Task 1-4 are long-horizon real robot tasks with direct prompting. Task 5-13 are long-horizon real robot tasks with high-level planner. Task 14-25 are under multi-task setting in real family scenes, including bathrooms, kitchens, and tabletops.

Stage 1	Stage 2	Long-horizon Tasks								Multi Tasks in Real Family Scenes										
		T1	T2	T3	T4	T5-8	T9-10	T11-13	T14	T15	T16	T17	T18	T19	T20	T21	T22	T23	T24	T25
✓		21/52	6/14	12/15	4/12	46/48	13/18	**24/39**	5/11	**2/6**	3/9	**3/7**	2/11	5/11	3/7	3/9	3/7	3/13	**4/9**	**7/7**
	✓	13/52	5/14	10/15	3/12	27/48	10/18	17/39	5/11	0/6	1/9	2/7	1/11	3/11	3/7	2/9	2/7	4/13	2/9	4/7
✓	✓	**28/52**	**9/14**	**15/15**	**9/12**	**45/48**	**15/18**	23/39	**6/11**	**2/6**	**5/9**	**3/7**	**3/11**	**6/11**	**4/7**	**5/9**	**4/7**	**6/13**	**4/9**	**7/7**

Table 11: **Performance of Scaling up to 7B on Understanding Tasks:** Evaluation on 6 Multimodal Understanding benchmarks and 7 VQA benchmarks.

Method	Multimodal Understanding Benchmarks						VQA Benchmarks						
	MMMU	MMStar	MME	OCRBench	HallBench	MMB	TextVQA	DocVQA	InfoVQA	AI2D	ChartQA	MTVQA	RealWorldQA
ChatVLA (2B)	37.4	47.2	1435.2	729	39.9	69.0	71.2	83.3	53.3	67.6	59.9	11.5	57.0
ChatVLA (7B)	50.7	60.5	1877.3	807	46.5	80.7	79.8	86.2	67.9	78.4	67.0	18.6	66.1

Table 12: **Performance of Scaling up to 7B on Real Robot tasks:** The embodied control performance is evaluated on 25 real-world manipulation tasks. Task 1-4 are long-horizon real robot tasks with direct prompting. Task 5-13 are long-horizon real robot tasks with high-level planner. Task 14-25 are under multi-task setting in real family scenes, including bathrooms, kitchens, and tabletops.

Method	Long-horizon Tasks								Multi Tasks in Real Family Scenes										
	T1	T2	T3	T4	T5-8	T9-10	T11-13	T14	T15	T16	T17	T18	T19	T20	T21	T22	T23	T24	T25
ChatVLA (2B)	28/52	9/14	15/15	9/12	45/48	15/18	23/39	6/11	2/6	5/9	3/7	3/11	6/11	4/7	5/9	4/7	6/13	4/9	7/7
ChatVLA (7B)	33/52	10/14	15/15	11/12	48/48	17/18	28/39	8/11	4/6	7/9	4/7	4/11	8/11	5/7	7/9	5/7	7/13	6/9	7/7

Table 13: **Performance of different VLM backbone on Understanding Tasks:** Evaluation on 6 Multimodal Understanding benchmarks and 7 VQA benchmarks.

Method	Multimodal Understanding Benchmarks						VQA Benchmarks						
	MMMU	MMStar	MME	OCRBench	HallBench	MMB	TextVQA	DocVQA	InfoVQA	AI2D	ChartQA	MTVQA	RealWorldQA
PaliGemma-3B	34.9	**48.3**	**1686.1**	614	32.2	65.6	**73.0**	78.0	40.5	**68.3**	54.2	**13.7**	54.2
ChatVLA (PaliGemma-3B)	35.3	48.0	1679.4	653	33.4	64.8	72.4	76.6	41.9	68.1	56.5	12.8	55.3
ChatVLA (Qwen2-VL-2B)	**37.4**	47.2	1435.2	**729**	**39.9**	**69.0**	71.2	**83.3**	**53.3**	67.6	**59.9**	11.5	**57.0**

Surprisingly, a smaller amount of visual-text data resulted in better performance. This aligns with the discussion in the previous subsection and the broader discussion in the paper, which suggests that even a limited amount of visual-text data is sufficient to reactivate visual-text alignment and bridge the gap between the base VLM and the VLA model.

6.3 Evaluation Metrics

The calculation method for long-horizon tasks is as follows: One point is awarded for each successfully completed step. After all steps of the task are executed, the total score is calculated. Additionally, "Avg. Len." represents the average success length of the model. This means that for multiple executions of the long-sequence tasks, the lengths of the sequences in which the model achieved success are recorded. Then, the average value of these lengths is calculated to obtain the "Avg. Len.", which serves as an important indicator to evaluate the performance of the model in handling long-sequence tasks in terms of the length of successful operation sequences.

6.4 Robot task

The embodied control performance of ChatVLA is evaluated on 25 real world manipulation tasks. **Long-horizon tasks with direct prompting.** As is shown in 6, all the tasks of this category are set under a real world toy scene.

- Task 1: Sort toys. On the desktop, there are two toy animals with random positions and postures, as well as two building blocks. The robotic arm needs to place all the animals on the desktop in the box on the left and all the building blocks in the basket on the right.
- Task 2: Stack cubes. The robotic arm first needs to pick up the orange building block from the right side and stack it on the yellow building block in the middle. Then, it needs to pick up the smallest pink square and stack it on the orange building block that was just stacked.
- Task 3: Place the toy in the drawer. The drawer is closed. Therefore, the robotic arm first needs to rotate and pull open the drawer. Then, it should pick up the toy on the table and place it into the drawer. Finally, close the gripper to shut the drawer.
- Task 4: Clean building blocks to the box. The robotic arm needs to put the building blocks on the table into the box on the right side one by one until there are no more building blocks on the table.

Long-horizon tasks with high-level planner. The settings are shown in 7.

- Task 5: Move the orange block to the basket. The robotic arm needs to pick up the building block next to the doll on the table and place it into the box on the right side.
- Task 6: Open the drawer. The robotic arm needs to rotate and grip the drawer handle, and then move parallel to the right to open the drawer.
- Task 7: Put the toy into it. The robotic arm needs to pick up the toy in the middle and place it into the open drawer.
- Task 8: Close the drawer. The robotic arm needs to close the gripper and gently push the open drawer to the left until the drawer is closed.
- Task 9: Move semi-circle building-block to basket. The robotic arm needs to pick up the semi-circular building block and place it into the basket on the right side.
- Task 10:Move rectangle building-block to basket. The robotic arm needs to pick up the rectangle building block and place it into the basket on the right side.
- Task 11: Get the plate and place it on the tablecloth. The robotic arm needs to pick up the pink plate from the upper part of the shelf on the right side and then place it on the tablecloth at the center of the table.
- Task 12: Flip the cup and place it on the tablecloth. The robotic arm needs to go to the bottom layer of the shelf on the right side, grip the mug, then turn it over and place it on the tablecloth in the middle of the table.
- Task 13: Move the bread to the plate. The robotic arm needs to grip the bread from the bread basket on the left side and place it on the plate that was just taken down.

Table 14: **Understanding task:** Evaluation on 6 Multimodal Understanding benchmarks and 7 VQA benchmarks. We use bold to denote top-ranked methods.

Method	Multimodal Understanding Benchmarks					VQA Benchmarks						
	MMMU	MMStar	MME	OCRBench	HallBench	TextVQA	DocVQA	InfoVQA	AI2D	ChartQA	MTVQA	RealWorldQA
1:1	36.1	44.7	1426.9	691	36.2	72.6	82.9	54.0	65.382	62.6	10.0	57.9
3:1	35.3	45.3	1399.5	726	36.4	**72.7**	**83.6**	**54.3**	67.0	**63.2**	10.3	**58.8**
1:3	**37.4**	**47.2**	**1435.2**	**729**	**39.9**	71.2	83.3	53.3	**67.6**	59.9	**11.5**	57.0

Figure 5: Comparison with Qwen2-VL on $MMMU_{val}$.

Cross-skill multi-tasking. The settings are shown in 8.

- Task 14:Put the soap to the soap box. This is a bathroom task. The robotic arm needs to pick up the soap from the left side of the washbasin and place it into the soap dish on the right side of the washbasin.
- Task 15:Pick up the cup and hang it on the shelf. This is a bathroom task. The robotic arm needs to pick up the cup from the sink and hang it on the shelf in front of the mirror.
- Task 16:Pick up the tooth-paste and put it on the table. This is a bathroom task. The robotic arm needs to pick up the toothpaste from the sink and place it on the table.
- Task 17:Remove the towel from the shelf. This is a bathroom task. The robotic arm needs to take down the towel hanging on the shelf and place it on another towel.
- Task 18:Move the bread from the pot to the plate. This is a kitchen task. The robotic arm needs to pick up the bread from the pot and place it on the plate.
- Task 19:Pick up the bread from the refrigerator. This is a kitchen task. The robotic arm needs to find the bread in the refrigerator and pick it up.
- Task 20:Move the banana onto the plate. The robotic arm needs to pick up the banana at a random position and place it on the plate in the middle.
- Task 21: Move the bread to the empty plate. The robotic arm needs to ignore the distractions, grip the bread, and then find the empty one among the two plates in front of it, and put the bread into that plate.
- Task 22:Hang on the cup. The robotic arm needs to pick up the mug and hang it on the shelf on the left side.
- Task 23:Move the tennis ball to the tennis can.

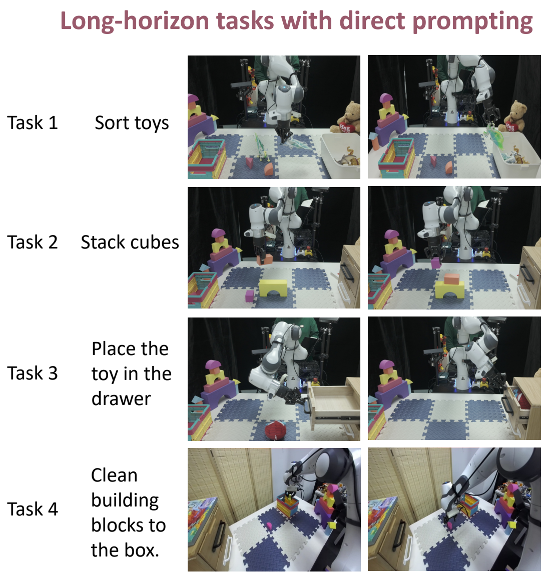

Figure 6: Settings of Long-horizon tasks with direct prompting

The robotic arm needs to pick up the tennis ball and lift it up to place it into the tennis ball can.

- Task 24:Stack the green cube onto the pink cube. The robotic arm needs to pick up the green cube on the right and stack it on top of the square on the left side.
- Task 25:Take away the lid of the box and put it on the table. The robotic arm needs to pick up the lid that is covering the box on the left side of the table and place the lid on the tabletop in the middle.

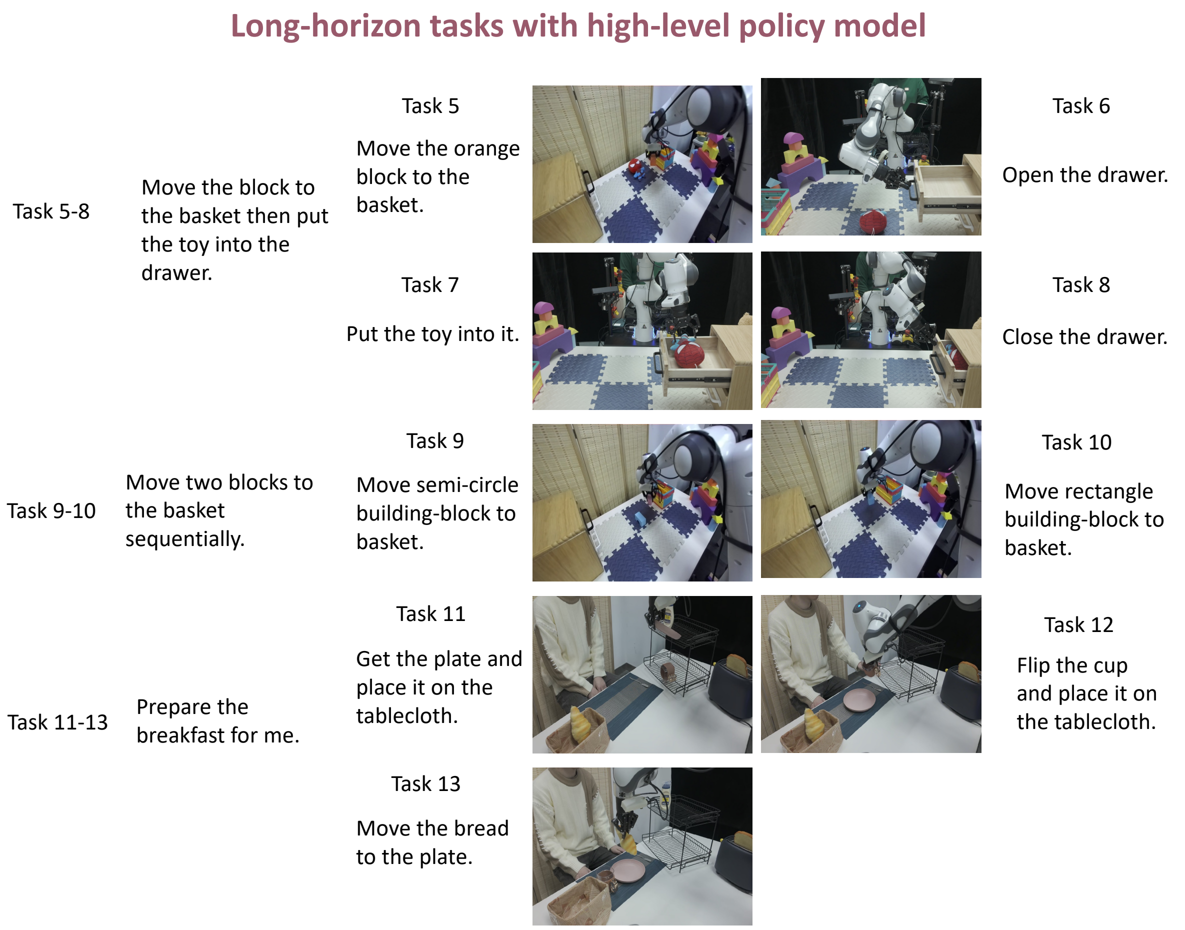

Figure 7: Settings of Long-horizon tasks with high-level planner

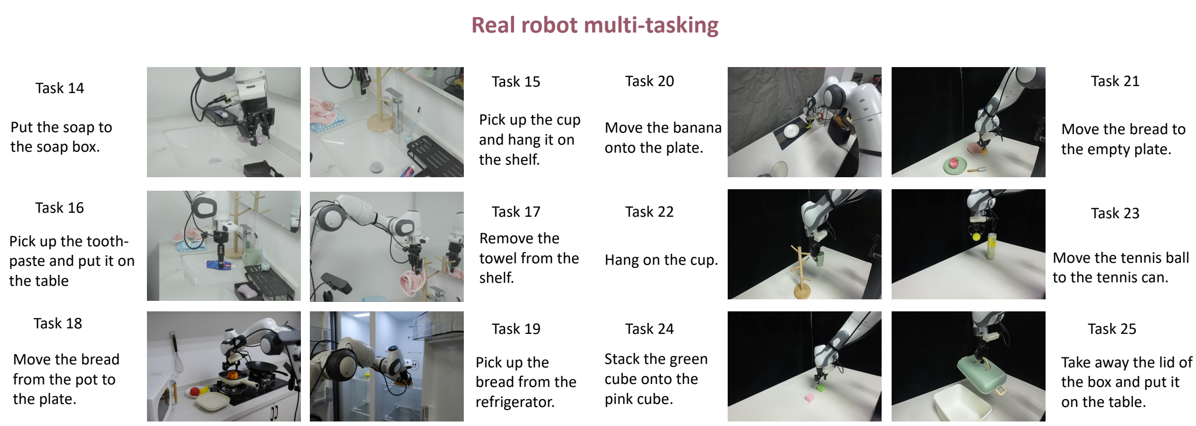

Figure 8: Settings of Cross-skill multi-tasking.