# DiTEA: Mixture-of-Experts for Vision-Language-Action Model in Robotic Manipulation

**Chengxuan Li**[12*†], **Xingwan Wang** [3*]

[1]Institute of Automation, Chinese Academy of Science, Beijing, China
[2]School of Advanced Manufacturing and Robotics, Peking University, Beijing, China
[3]University of Science and Technology of China, Hefei, China
lichengxuan0904@gmail.com, wangxingwan20@gmail.com

## Abstract

The current diffusion-based Vision-Language-Action (VLA) models have faster inference speed and the ability to solve the action muti-modality problem in robot manipulation tasks compared to traditional autoregressive models after large-scale pre-training and post-training. However, the diffusion-based VLA models were found to have poor instruction-following ability, and after fine-tuning training on multiple tasks, them often suffer from "skill forgetting" due to conflicting model weights on each task. To address this problem, we propose DiTEA, a Diffusion Transformer-based Mixture-of-Experts (MoE) VLA model. Specifically, it fuses the MoE module into the action head of VLA to form Action MoE, and in addition, we design the Task-Instruction Gate, which uses language instructions to select specific experts for tasks they specialize in, in order to improve the VLA's instruction-following ability. We conducted comprehensive experiments and ablation study to evaluate the efficacy of our model under different designs. Experimental results from simulation and real-world show that our DiTEA has excellent improvement in multi-task compared to baseline and other VLAs.

## Introduction

Recent advancements in Vision-Language-Action (VLA) models have demonstrated remarkable capabilities in robotic manipulation tasks by integrating large-scale pre-trained vision-language models (VLMs) with action generation functions (Brohan et al. 2022; Wu et al. 2023; Chi et al. 2023; Wen et al. 2025a; Team et al. 2024; Li et al. 2023; Stone et al. 2023; Nair et al. 2022). Traditional autoregressive models (Zhao et al. 2023; Brohan et al. 2022; Zitkovich et al. 2023; Bousmalis et al. 2023; Liu et al. 2025), while effective, often struggle with action multi-modality—representing diverse valid trajectories for a given task—and suffer from slow inference speeds. In contrast, diffusion-based VLA models, such as CogACT (Li et al. 2024a) and GR00T N1 (Bjorck et al. 2025), have emerged as a promising alternative, offering faster inference and superior handling of multi-modal action distributions (Chi et al. 2023; Wen et al. 2025b). However, these models exhibit critical limitations, particularly in instruction-following ability, and tend to suffer from "skill forgetting" when fine-tuned on multiple tasks due to conflicting weight updates (Black et al. 2024). Addressing these challenges is crucial for developing robust, generalizable robotic control systems capable of executing diverse real-world tasks. The persistent challenges of skill interference and precise instruction-following in complex multi-task learning find a compelling parallel in the brain's organization. Cognitive neuroscience demonstrates that biological intelligence maintains robustness and flexibility through specialized, modular cortical processing units (Kanwisher 2010). These neural "experts" operate with isolated representations, a mechanism that effectively minimizes catastrophic interference between distinct skills (McClelland, McNaughton, and O'Reilly 1995; O'Reilly and Rudy 2001). Inspired by this fundamental biological principle, we implement a Mixture of Experts (MoE) architecture in our VLA model. Notably, this biologically-inspired approach has been increasingly validated in recent research, with successful MoE applications demonstrated in multi-task reinforcement learning (Zhao et al. 2025; Huang et al. 2025; Wu et al. 2025; Huang et al. 2024), cross-modal reasoning (Mengara and Moon 2025; Suzuki et al. 2023; Zhou et al. 2022; Mustafa et al. 2022; Li et al. 2025), and embodied AI systems (Yuan and Li 2022; Yang et al. 2025; Fan et al. 2022), strongly supporting the effectiveness of our architectural choice.

Furthermore, the prefrontal cortex (PFC) serves as a dynamic gate or router, integrating sensory inputs and task goals to selectively engage specialized neural "expert" circuits. (Schneegans and Bays 2017). This mechanism ensures precise, context-appropriate actions while mitigating interference between skills (Spaak et al. 2017), These neurobiological insights offer concrete design principles for developing artificial multi-task learning systems that balance specialization and generalization (Murphy et al. 2020; Shine et al. 2015). Inspired by the gating mechanism of the PFC, we introduce a Task-Instruction Gate in DiTEA's Action MoE module to resolve the conflict between multi-task learning and precise instruction execution.

Motivated by these critical limitations identified in diffusion-based VLAs and the powerful organizing principles of biological neural systems outlined above, we design DiTEA, a Diffusion Transformer-based MoE VLA model. DiTEA fundamentally re-architects the action gen-

*These authors contributed equally.
†Chengxuan Li is the corresponding author.

eration paradigm by integrating a MoE model into the action head, moving beyond singular optimization of multi-task learning to tackle the inherent conflict between acquiring diverse skills and maintaining precise instruction adherence. The highlighting innovation of our design is the Action MoE module governed by a biologically-inspired Task-Instruction Gate. This gate, adaptively activates the most suitable task experts based on a joint understanding of the visual and textual instruction, ensuring robust performance across diverse tasks while mitigating skill interference. Through experimental validation spanning virtual simulations and physical robotic platforms, we confirm DiTEA's fundamental ability to harmonize task execution with instruction adherence, resolving a critical tension in multi-task VLA systems.

The contributions of this work are summarized below:

- We propose DiTEA, a novel diffusion-based VLA model that integrates MoE architecture into the action head to enable task-specific action generation.
- We introduce a Task-Instruction Gate mechanism that dynamically selects specialized experts based on task requirements, significantly improving the model's instruction-following capability.
- We conduct comprehensive experiments in both simulated and real-world environments, demonstrating DiTEA's superior performance in multi-task settings compared to existing VLA approaches.

## Related Work

### Vision Language Action Models

Since pre-trained Vision Language Models (VLMs) (Chen et al. 2024b; Karamcheti et al. 2024; Liu et al. 2023; Team et al. 2023) have strong language and vision comprehension capabilities, numerous Vision Language Action (VLA) models have been designed by extending them to integrate action generation functions (Li et al. 2023; Zitkovich et al. 2023; Cai et al. 2023; Li et al. 2024b; Ding et al. 2024; Yu et al. 2024; Shentu et al. 2024; Kim et al. 2024). For instance, RT-2 (Zitkovich et al. 2023) tokenizes 7D actions into discrete tokens and autoregressively predicts them using the VLM PaLI-X (Chen et al. 2024a), just like the language tokens. OpenVLA (Kim et al. 2024) uses a fusion approach based on the Llama 2 language model architecture (Touvron et al. 2023) that combines DINOv2 (Oquab et al. 2023) and SigLIP (Zhai et al. 2023) dual vision coders. Recent studies, such as CogACT (Li et al. 2024a), GR00T N1 (Bjorck et al. 2025) and DiffusionVLA (Wen et al. 2025b), have introduced diffusion modeling as an innovative approach to robot action modeling. These diffusion-based methods excel at representing the variety of effective trajectories a robot may execute to accomplish a task, and have shown particular strengths in modeling complex, non-unique action distributions (Li et al. 2024a). Despite showing promise via fine-tuning across multiple robotic datasets, these methods exhibit degraded instruction-following capabilities (Black et al. 2024).

### Mixture-of-Experts Model

The Mixture-of-Experts (MoE) architecture improves computational efficiency and generalization by dividing a large network into specialized smaller sub-networks with sparse activation, particularly effective for multi-task, multi-modal learning with complex data distributions (Riquelme et al. 2021; Fedus, Zoph, and Shazeer 2022; Zoph et al. 2022; Xue et al. 2024). DeepSeekMoE (Dai et al. 2024) is an innovative MoE architecture that achieves expert specialization through fine-grained segmentation and shared expert isolation. DiT-MoE (Fei et al. 2024) applies MoE architecture to diffusion Transformers for conditional image generation, achieving efficient performance through shared expert routing and expert-level balance loss. MoRE (Zhao et al. 2025) is a novel VLA model that incorporates the MoE architecture into its backbone, introducing reinforcement learning (RL) for fine-tuning large-scale VLA models using substantial amounts of mixed-quality data. DriveMoE (Yang et al. 2025) applies MoE architecture to end-to-end autonomous driving through a Scene-Specialized Vision MoE and a Skill-Specialized Action MoE. BIG-MoE (Ma et al. 2024) concatenates prompt and feature tokens for gating network scoring, isolating expert network input to enhance noise resilience and processing precision. The Isolated Gating Mechanism in BIG-MoE inspires our approach to enhance the instruction-following capability of VLA models through Task-Instruction Gate.

## Preliminaries

### Mixture-of-Experts

The Mixture-of-Experts (MoE) model is a machine learning architecture that enhances efficiency and scalability by combining multiple specialized sub-models, or "experts," each responsible for processing a subset of the input space. The MoE framework leverages a gating mechanism to dynamically route inputs to the most suitable experts, enabling efficient computation and improved performance on complex tasks.

In an MoE model, given an input $x$, the output is computed as a weighted combination of the outputs from $K$ experts. Each expert is typically a neural network, denoted as $E_i(x)$, where $i = 1, 2, \ldots, K$. The gating network, $G(x)$, assigns weights to each expert based on the input, producing a probability distribution over the experts. The final output $y$ is expressed as:

$$y = \sum_{i=1}^{K} G(x)_i \cdot E_i(x) \tag{1}$$

Here, $G(x)_i$ represents the gating probability for the $i$-th expert, satisfying $\sum_{i=1}^{K} G(x)_i = 1$ and $G(x)_i \geq 0$. The gating function is often implemented as a softmax over a learned function of the input:

$$G(x)_i = \frac{\exp(h_i(x))}{\sum_{j=1}^{K} \exp(h_j(x))} \tag{2}$$

where $h_i(x)$ is a learned function (e.g., a linear transformation followed by a non-linearity) that determines the relevance of the $i$-th expert for input $x$.

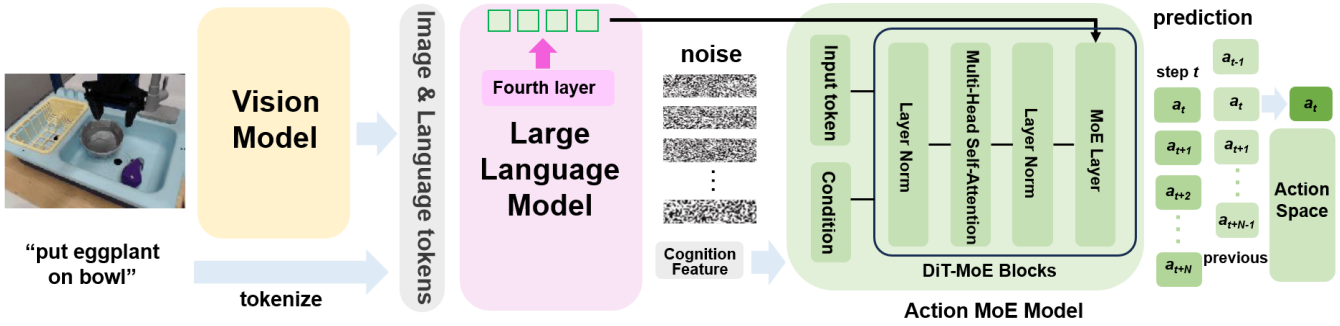


Figure 1: **Overview:** Our model consists of three parts: 1) a vision module: it consists of pre-trained DINO v2 and SigLIP, which encode images into visual tokens; 2) a language module, the main body of which is the LLAMA 2 7B model, which fuses the visual tokens and commands to generate a cognition feature for the action module to generate the robot's actions; 3) an action module: it consists of a DiT-MoE architecture, which we call Action MoE, and again has stronger multi-task learning capability and instruction following ability.

To enhance efficiency, MoE models often employ sparse gating, where only a subset of experts (e.g., top-$k$) is activated for each input, reducing computational cost. This sparsity is critical for scaling to large models while maintaining manageable resource requirements. Additionally, load balancing is often introduced to ensure even distribution of inputs across experts, typically through an auxiliary loss term:

$$L_{\text{balance}} = \alpha \sum_{i=1}^{K} \left( \frac{1}{N} \sum_{n=1}^{N} G(x_n)_i - \frac{1}{K} \right)^2 \tag{3}$$

where $N$ is the number of inputs, and $\alpha$ is a weighting factor. This loss encourages balanced utilization of experts.

### Diffusion-based VLA

Diffusion models, initially developed for generative tasks, have been adapted for robotics through diffusion policy, modeling action prediction as a generative process. For a state $s_t$ (e.g., visual and language inputs), the policy generates action sequences by iteratively denoising a noisy action distribution, modeling the conditional distribution:

$$p(a_t|s_t) = \int p(a_t^{(0)}|a_t^{(T)}, s_t) p(a_t^{(T)})\, da_t^{(T)}, \tag{4}$$

where $a_t^{(T)}$ is a noisy action sampled from a Gaussian prior.

Diffusion-based VLA models, such as CogACT, leverage a Diffusion Transformer (DiT) to integrate vision $v$, language $l$, and actions $a$, which models the joint distribution $p(a|v,l)$ using a denoising process:

$$a_t^{(k-1)} = \sqrt{\alpha_k} a_t^{(k)} + \sqrt{1-\alpha_k} \epsilon_\theta(v, l, a_t^{(k)}, k), \tag{5}$$

where $\epsilon_\theta$ is a transformer predicting noise, and $\alpha_k$ controls the noise schedule. The training objective minimizes:

$$L = \mathbb{E}\left[ \|\epsilon - \epsilon_\theta(v, l, a_t^{(k)}, k)\|^2 \right]. \tag{6}$$

In Diffsion-based VLA, actions are organized into *action chunks*, where a sequence of actions $a = [a_t, a_{t+1}, \ldots, a_{t+C-1}]$ over $C$ timesteps is predicted as a single unit. This chunking reduces temporal dependencies and enhances efficiency, allowing the transformer to model long-horizon tasks via attention over multimodal inputs. The action chunk output is conditioned on $v$ and $l$, enabling tasks like language-guided robotic manipulation with improved coherence and robustness.

## Methods

In this section, we will introduce **DiTEA**, the MoE-based VLA model, in detail. Specifically, we will first introduce its main architecture, followed by a detailed description of its **Action MoE** module and **Task-Instruction Gate** mechanism.

### Model Architecture

Our visual and language modules are adapted from the existing VLMs in (Karamcheti et al. 2024) and have a total of about 7B parameters. We briefly describe them below.

**Vision Module** The vision module converts raw image inputs into a streamlined set of perceptual tokens to facilitate efficient visual data handling. It utilizes two state-of-the-art vision transformer architectures, DINOv2 (Oquab et al. 2023) and SigLIP (Zhai et al. 2023), pretrained on vast Internet-scale image corpora to capture intricate visual details and robust semantic insights. For each time step $t$, the input image $o_t$ is processed to produce two downsampled feature maps, $f_t^{\text{DINO}}$ and $f_t^{\text{Sig}}$, which are merged along their channel axes. These combined features pass through a linear projection and are serialized into a sequence of visual tokens, $V = \{v_1, v_2, \ldots, v_{N_V}\}$, with a standard length of $N_V = 256$, optimizing computational efficiency.

**Language Module** The language module fuses visual and language inputs to enable advanced cognitive reasoning. Built on the LLAMA2 architecture (Touvron et al. 2023), it processes a language instruction $l$ through LLAMA2's tokenizer to generate a sequence of linguistic tokens, $T = \{l_1, l_2, \ldots, l_{N_T}\}$. These are combined with the visual tokens $V$ and a learnable cognition token $c$, forming a unified sequence. A causal attention mechanism processes this

sequence, yielding an output feature $f_t^c$ tied to the cognition token, which encodes the integrated multimodal information. This feature drives the action module to determine precise, task-oriented actions.

**Action Module** Inspired by CogACT (Li et al. 2024a), we adopt DiT as our action module, which serves as a powerful backbone for the action decoding process, capable of modelling complex and time-dependent actions. Incorporating MoE structure on top of this is called **Action MoE**.

## Action MoE

The main body of Action MoE is a DiT-MoE structure, (as shown in Figure 1), whereas traditional DiT is mainly composed of an MLP layer and a self-attention layer, DiT-MoE replaces the MLP layer with a MoE layer, where each expert is an MLP with the same parameter settings. Each expert is responsible for its own different task, so that the parameters of different task experts are isolated from each other, preventing weight conflicts in multi-task learning. Meanwhile, in order to capture the commonalities in each task and prevent redundancy among the experts' parameters, we specifically set up a shared expert, which is different from the other experts in that it is activated by gating every time. In addition to this, since directly learned routing policies often encounter load imbalance problems that lead to significant performance drawbacks, we also introduce load balancing losses to balance the experts. We then describe the shared expert and load balancing loss in detail.

**Shared Expert** In a standard MoE setup, each expert is typically specialized for a specific subset of the input space. However, *shared expert* are designed to process inputs across different tasks or subsets, providing a common knowledge base that reduces redundancy and enhances parameter efficiency. The gating network assigns weights to both task-specific experts and shared expert, enabling flexible routing.

The output of an MoE layer with shared expert can be expressed as:

$$y = \sum_{i=1}^{N} G_i(x) \cdot E_i(x) + G_s(x) \cdot E_s(x), \tag{7}$$

where:$x$ is the input to the MoE layer, $G_i(x)$ is the gating weight for the $i$-th task-specific expert $E_i(x)$, $G_s(x)$ is the gating weight for the shared expert $E_s(x)$,$N$ is the number of task-specific experts. The shared expert $E_s(x)$ contributes to the output for all inputs, modulated by the gating function $G_s(x)$, which determines its relevance for a given input.

**Load Balancing Loss** Routing imbalances in MoE can degrade performance, so we propose a modified balance loss to distribute workloads evenly. This is defined as:

$$L_{\text{balance}} = \beta \sum_{j=1}^{m} \frac{m}{LT} \sum_{t=1}^{T} I(t,j) \cdot \frac{1}{T} \sum_{t=1}^{T} P(t,j), \tag{8}$$

where $\beta$ is the tuning factor, $T$ represents the length of action sequence, $I(t,j)$ indicates whether token $t$ selects expert $j$,

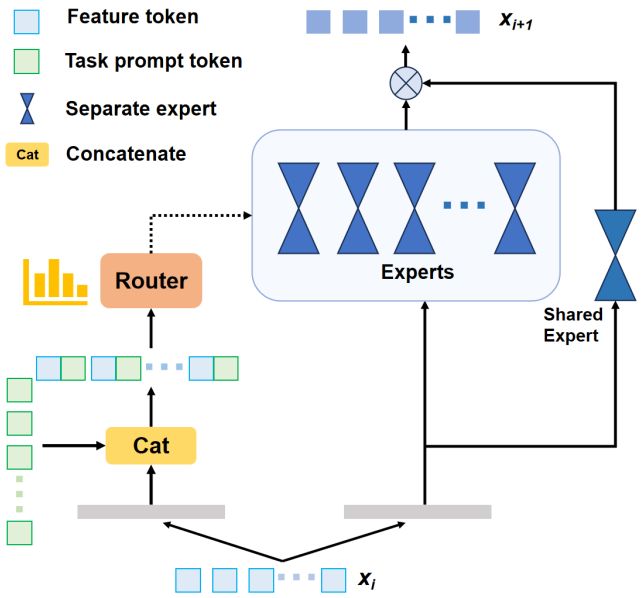


Figure 2: Illustration of the MoE framework. Task-Instruction Gate significantly improves VLA instruction following by selecting experts via instruction prompt token

and $P(t,j)$ denotes the likelihood of token $t$ being assigned to expert $j$. This adjustment helps ensure all experts are utilized effectively.

## Task-Instruction Gate

Compared to traditional MoE gate where experts are selected by feature tokens, the Task-Instruction Gate significantly improves the instruction following ability of VLAs. By using the instruction prompt token $x_{\text{inst}}$ as the gating token, the Task-Instruction Gate ensures that expert selection is directly guided by the task instructions (Figure 2), improving the model's responsiveness to diverse instructions. This design enhances the ability to handle complex, instruction-driven robotic tasks. The Task-Instruction Gate employs a standard gating mechanism parameterized by $\Theta$, which computes a weighted combination of expert outputs based on a softmax operation over the instruction-driven gating scores.

The gating decision is formalized as follows:

$$G(x_{\text{inst}}; \Theta) = \text{softmax}(W_g x_{\text{inst}} + b_g), \tag{9}$$

where $x_{\text{inst}}$ is the instruction vector, and $W_g$ and $b_g$ are the weight matrix and bias of the gating network, respectively. The output of the MoE system with Task-Instruction Gate is computed as a weighted sum of expert outputs:

$$F_{\text{MoE}}(x; \Theta, \{W_i\}_{i=1}^{N}) = \sum_{i=1}^{N} G(x_{\text{inst}}; \Theta)_i \cdot F_i(x; W_i), \tag{10}$$

where $F_i(x; W_i)$ represents the output of the $i$-th expert parameterized by $W_i$, and $N$ is the total number of experts.

| WidowX Robot | Method | Put Spoon on Towel | Put Carrot on Plate | Stack Green Block on Yellow Block | Put Eggplant in Yellow Basket | Average |
|---|---|---|---|---|---|---|
| SIMPLER (Visual Matching) | RoboVLMs | **40.0** | 21.6 | 13.3 | 58.7 | 33.4 |
| | Octo-Small | 21.6 | 10.0 | 4.0 | 56.7 | 23.1 |
| | OpenVLA | 0.0 | 0.0 | 0.0 | 4.0 | 1.0 |
| | CogACT | 21.6 | 37.5 | 4.0 | 77.5 | 35.2 |
| | **DiTEA(Ours)** | 28.8 | **41.2** | **13.3** | **79.7** | **40.8** |

Table 1. Evaluation results on the WidowX robot in the SIMPLER Visual Matching setting.

## Training Objective

Our DiTEA is jointly trained or fine-tuned end-to-end by minimizing the mean-squared error (MSE) between the predicted noise outputs from the action module and the corresponding ground truth noise. The loss function is formulated as:

$$\mathcal{L}_{\text{MSE}} = \mathbb{E}_{\sim\mathcal{N}(0,1),x} \left\| \hat{\epsilon}^i - \epsilon \right\|_2^2, \tag{11}$$

where $\hat{\epsilon}^i$ represents the predicted noise for the noisy action sequence $(\hat{a}_t^i, \hat{a}_{t+1}^i, \ldots, \hat{a}_{t+M}^i)$ at the $i$-th denoising step, and $\epsilon$ is the corresponding ground truth noise, ensuring the model learns to predict both the current action and multiple future actions smoothly.

# Experiment

**Dataset** We use the BridgeData V2 dataset as our primary training dataset for our simulation experiments. It contains 60,096 trajectories collected across 24 environments on a publicly available low-cost robot. BridgeData V2 is compatible with a wide variety of open-vocabulary, multi-task learning methods conditioned on goal images or natural language instructions, providing extensive task and environment variability. DROID is a diverse robot manipulation dataset comprising 76k demonstration trajectories (350 hours of interaction data) collected across 564 scenes and 86 tasks by 50 data collectors spanning North America, Asia, and Europe over 12 months. Prior research has demonstrated that training with DROID yields policies with superior performance and enhanced generalization capabilities. For our real-robot experiments, we employ this dataset for pre-training followed by fine-tuning with our own collected data. We utilize the full open-sourced dataset along with the publicly available policy learning code, while also referencing their detailed hardware replication guide for our experimental setup.

**Implementation Details** The model has a batch size of 256, with 8 diffusion steps per sample, and is initialized with pre-trained vision and language module weights from (Kim et al. 2024). The vision module (i.e., DINOv2 and SigLIP), the language module (i.e., LLAMA-2), and the action module are trained end-to-end with a constant learning rate of $2 \times 10^{-5}$. In the simulation experiments training, we use the Bridge dataset for fine-tuning, with about 20K training steps. Training is performed on 8 NVIDIA A100 GPUs using PyTorch's Fully Shared Data Parallel (FSDP) framework. For fine-tuning, we freeze the VLM and vision modules, and train only DiT. In real-world experiments, we first use the DROID dataset for full-parametric pre-training (without freezing any parameter), and then fine-tune it with the real-world data obtained from teleoperation by training only DiT and freezing the other modules, also on 8 NVIDIA A100 GPUs, with a constant learning rate of $2 \times 10^{-5}$, and using PyTorch's Fully Shared Data Parallel (FSDP) framework.

## Simulated Evaluation

We use the full BridgeV2 data to finetune DiTEA training for inference on four tasks related to SIMPLER's WidowX robot.

**Evalution Settings** SIMPLER is a benchmark designed to bridge the gap between simulation and real-world robotic manipulation. It provides standardized evaluation protocols to rigorously assess the sim-to-real transfer capabilities of learning-based policies. SIMPLER offers an evaluation setting called Visual Matching, which closely replicates real-world tasks by minimizing discrepancies between the simulated and real environments. For the WidowX robot, this evaluation framework includes four tasks: 1) Put spoon on towel; 2) Put carrot on plate; 3) Stack green block on yellow block; and 4) Put eggplant in yellow basket. Task performance is quantified through success rate measurements.

**Results** We compare DiTEA with four existing models (Table 1), and the results show that on SIMPLER, except for the *Put spoon on towel* task, DiTEA achieves the highest success rate for the other three tasks, and the average success rate is 40.8%, which is higher than the other models. The experiment proves that our DiTEA has obvious advantages in multi-task learning.

## Real-World Evaluation

We perform real-world experiments on a single-armed Franka Research 3 robot equipped with two cameras, a ZED mini for the wrist and a ZED 2i for the third view (Figure 3). For the real-world experiments, we use the DROID dataset for pre-training, and then use GELLO (Wu et al. 2024) teleoperation to capture data from the real machine for fine-tuned post-training. Each task was fine-tuned on 60-100 trajectories of varying difficulty. The evaluation was performed on prepared tasks.

**Evalution Settings** We categorize the tasks into three groups: 1) *Pick and Place*: Grasping and placing three distinct objects, including corn, potato, and cabbage. 2) *Stack*:

| Method | Pick and Place | | | | Stack | | | Open and Close | | | Task (All) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Corn | Potato | Cabbage | Avg. | Cube | Bowl | Avg. | Drawer | Box | Avg. | Avg. |
| OpenVLA | 15.0 | 17.5 | 0.0 | 10.8 | 0.0 | 5.0 | 2.5 | 4.2 | 6.5 | 5.4 | 6.2 |
| CogACT | 67.5 | 52.2 | 72.5 | 64.1 | 6.1 | 24.3 | 15.2 | 19.8 | 25.5 | 22.7 | 34.0 |
| **DiTEA(Ours)** | **76.5** | **65.6** | **87.5** | **76.5** | **10.2** | **28.2** | **19.2** | **22.1** | **27.5** | **24.8** | **40.2** |

Table 2. Real-world evaluation with the Franka Research 3 Robot across three tasks. All models are pre-trained on DROID and then fine-tuned on our collected data.

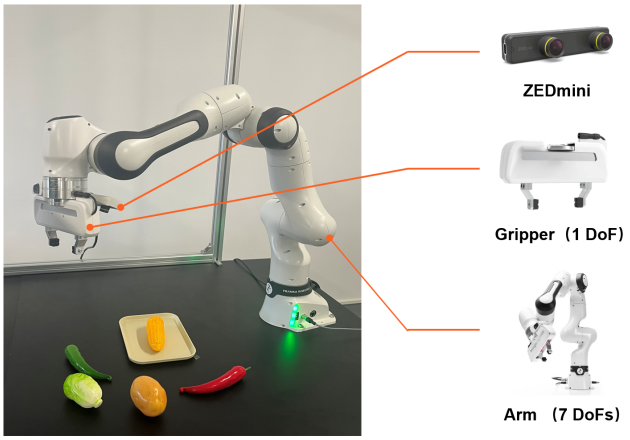


Figure 3: Franka robot setup.

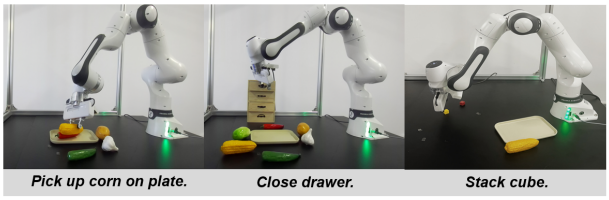


Figure 4: Real-world evaluation environments of Franka robot.

Stacking cubes and plates, with a particular focus on stacking cubes, which requires precise manipulation and posing a significant challenge. 3) *Open and Close*: Opening drawer and box, pausing for a few seconds, and then closing them.

**Results** We conduct three main experiments in the real world to fully evaluate the multifaceted benefits of our DiTEA model.

*Results of Muti-tasking Experiments:* In real-world experiments, we evaluated our approach across eight distinct tasks, as presented in Table 2, which includes three main categories: Pick and Place (corn, potato, cabbage), Stack (cube, bowl), and Open and Close (drawer, box). Compared to OpenVLA and CogACT, our method demonstrates significantly improved success rates across all eight tasks, with particularly notable enhancements in more challenging tasks, such as cube stacking and cabbage pick-and-place. Our approach achieves the highest average success rates in every sub-task, culminating in an overall average of 40.2%, which substantially outperforms CogACT and OpenVLA and CogACT. This consistent superiority suggest that our DiTEA not only excels in individual tasks but also achieves remarkable performance in multi-task learning, demonstrating robust generalization capabilities.

| Method | Unseen Colors | Unseen Shapes | Unseen Categories | Avg. |
|---|---|---|---|---|
| OpenVLA | 0.0 | 5.0 | 0.0 | 1.7 |
| CogACT | 45.0 | 61.6 | 12.5 | 39.7 |
| **DiTEA (Ours)** | **53.5** | **71.3** | **15.0** | **46.6** |

Table 3. Real-world generalization evaluation with the Franka Research 3 Robot on unseen colors, shapes and categories.

*Results of generalization experiments:* To thoroughly evaluate the generalization capability of our DiTEA model, we conducted a series of experiments using the Franka Research 3 Robot, focusing on its performance with objects exhibiting unseen colors, shapes, and categories. These tests were designed to assess the model's zero-shot learning ability—its capacity to handle tasks involving objects or attributes not encountered during the fine-tuning phase. The results, as detailed in Table 3, highlight DiTEA's superior performance compared to two baseline methods, OpenVLA and CogACT, across all evaluated metrics.

*Results of Instruction-Following Experiments:* We assess DiTEA's ability to comprehend and follow diverse instructions across three task categories: 1) Basic tasks with explicit instructions (e.g., "stack the green block on the yellow block"), where success requires precise object interaction and spatial reasoning; 2) Complex tasks involving multi-step reasoning or implicit semantics (e.g., "place the leftmost object into the red zone"), designed to evaluate higher-level instruction parsing and contextual awareness; 3) Distractor tasks featuring semantically similar or visually confusable objects (e.g., "pick up the spoon from among the utensils"), testing robustness to perceptual ambiguity. Across all categories, DiTEA shows differentiated performance compared to existing approaches: while it underperforms the autoregressive model OpenVLA, it significantly outperforms CogACT based on diffusion models. The performance gap between DiTEA and CogACT can be attributed to autoregressive models' inherent strength in temporal coherence . These results shown in Figure 5 collectively demonstrate

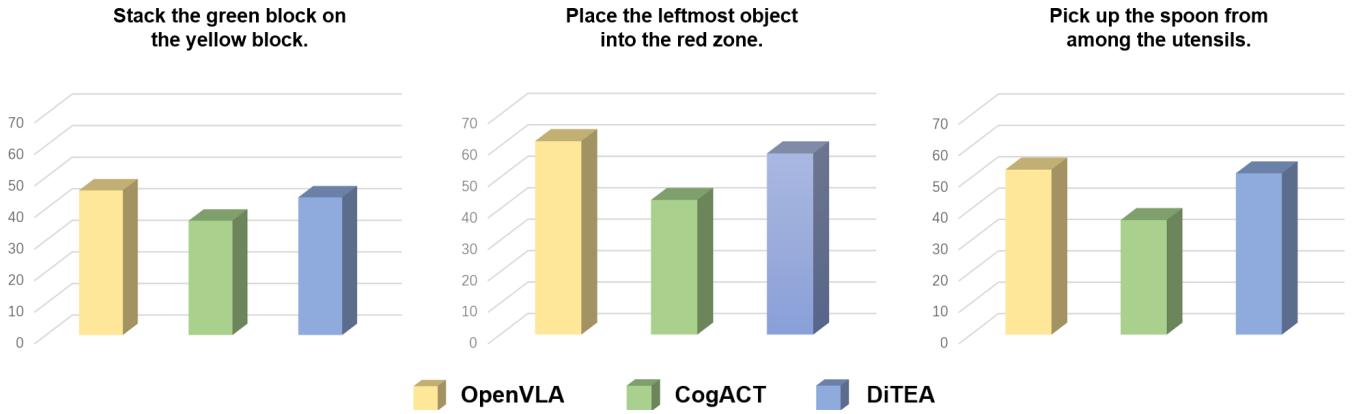


Figure 5: Comparison of instruction-following capabilities demonstrated across the three types of tasks. Success rates are evaluated through a multi-stage scoring protocol that measures both task completion and semantic alignment. In object manipulation tasks, partial credit is assigned for intermediate steps, enabling fine-grained assessment beyond binary success metrics.

that DiTEA partially addresses critical limitations in existing diffusion-based VLAs: it shows improvement over diffusion models' struggle to adapt to varied language instructions.

## Ablation Study

We employ SIMPLER evaluation on the WidowX robot for all ablation studies. The ablation study presented in the Table 4 analyzes the effects of three components, namely load balancing loss, task instruction gate, and shared expert, on the model's overall performance. While the baseline configuration ($Ex_0$) which incorporates none of the three components yields a success rate of merely 40.2%, enabling only expert ($Ex_{1-1}$) boosts the rate to 42.4%, and activating solely the task instruction gate ($Ex_{1-2}$) elevates performance to 43.8%. Their combined activation ($Ex_{1-3}$) further propels the success rate to 48.7%, a leap highlighting significant synergy between the two. In load balancing loss scenarios, this component alone ($Ex_{2-1}$) yields a success rate of 41.1%. When combined with shared expert ($Ex_{2-2}$), performance rises to 44.2%, and pairing it with the task instruction gate ($Ex_{2-3}$) further boosts results to 47.9%. Ultimately, integrating all three components ($Ex_{2-4}$) culminates in the highest success rate of 51.5%—an 11.3% improvement over the baseline—confirming the value of each individual component.

Additionally, we conducted an experiment by introducing a +Params baseline model ($Ex_{3-1}$), where we scaled its action head to match the count of our DiTEA model. The comparative results demonstrate that the performance is attributable to the MoE architecture itself, rather than just the increase in parameters.

| Methods | LB loss | TIG | Shared expert | +Params | Success rate |
|---|---|---|---|---|---|
| $Ex_0$ | × | × | × | × | 40.2% |
| $Ex_{1-1}$ | × | × | ✓ | × | 42.4% |
| $Ex_{1-2}$ | × | ✓ | × | × | 43.8% |
| $Ex_{1-3}$ | × | ✓ | ✓ | × | 48.7% |
| $Ex_{2-1}$ | ✓ | × | × | × | 41.1% |
| $Ex_{2-2}$ | ✓ | × | ✓ | × | 44.2% |
| $Ex_{2-3}$ | ✓ | ✓ | × | × | 47.9% |
| $Ex_{2-4}$ | ✓ | ✓ | ✓ | × | **51.5%** |
| $Ex_{3-1}$ | × | × | × | ✓ | 45.4% |

Table 4. Ablation study on load balancing loss, task instruction gate, shared expert and parameter count on SIMPER simulation environment.

## Conclusion

We introduced DiTEA, a Diffusion Transformer-based Mixture-of-Experts Vision-Language-Action model that effectively addresses persistent challenges of skill forgetting and enhanced instruction-following capability in diffusion-based robotic manipulation. By integrating a novel Action MoE module governed by a neurobiologically inspired Task-Instruction Gate, DiTEA achieves promising multi-task performance while significantly improving instruction following ability. The approach remains constrained by high computational demands during training, limiting large-scale pre-training opportunities. This restriction consequently impedes zero-shot adaptation to unseen operational scenarios, where task-specific fine-tuning remains essential for optimal performance. Future work will explore efficient MoE scaling strategies and pursue large-scale pre-training to advance cross-task generalization for general-purpose robotic agents.

## Acknowledgments

This work was primarily supported by the Institute of Automation, Chinese Academy of Sciences, and Peking University. We are grateful for their guidance.

## References

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

Black, K.; Brown, N.; Driess, D.; Esmail, A.; Equi, M.; Finn, C.; Fusai, N.; Groom, L.; Hausman, K.; Ichter, B.; et al.

2024. $\pi_0$: A Vision-Language-Action Flow Model for General Robot Control. *arXiv preprint arXiv:2410.24164*.

Bousmalis, K.; Vezzani, G.; Rao, D.; Devin, C.; Lee, A. X.; Bauzá, M.; Davchev, T.; Zhou, Y.; Gupta, A.; Raju, A.; et al. 2023. Robocat: A self-improving generalist agent for robotic manipulation. *arXiv preprint arXiv:2306.11706*.

Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Dabis, J.; Finn, C.; Gopalakrishnan, K.; Hausman, K.; Herzog, A.; Hsu, J.; et al. 2022. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*.

Cai, Y.; Yang, S.; Li, M.; Chen, X.; Mao, Y.; Yi, X.; and Yang, W. 2023. Task2Morph: Differentiable Task-Inspired Framework for Contact-Aware Robot Design. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 452–459. IEEE.

Chen, X.; Djolonga, J.; Padlewski, P.; Mustafa, B.; Changpinyo, S.; Wu, J.; Ruiz, C. R.; Goodman, S.; Wang, X.; Tay, Y.; et al. 2024a. On scaling up a multilingual vision and language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14432–14444.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024b. 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*, 24185–24198.

Chi, C.; Xu, Z.; Feng, S.; Cousineau, E.; Du, Y.; Burchfiel, B.; Tedrake, R.; and Song, S. 2023. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 02783649241273668.

Dai, D.; Deng, C.; Zhao, C.; Xu, R.; Gao, H.; Chen, D.; Li, J.; Zeng, W.; Yu, X.; Wu, Y.; et al. 2024. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*.

Ding, P.; Zhao, H.; Zhang, W.; Song, W.; Zhang, M.; Huang, S.; Yang, N.; and Wang, D. 2024. Quar-vla: Vision-language-action model for quadruped robots. In *European Conference on Computer Vision*, 352–367. Springer.

Fan, Z.; Sarkar, R.; Jiang, Z.; Chen, T.; Zou, K.; Cheng, Y.; Hao, C.; Wang, Z.; et al. 2022. $M^3$vit: Mixture-of-experts vision transformer for efficient multi-task learning with model-accelerator co-design. *Advances in Neural Information Processing Systems*, 35: 28441–28457.

Fedus, W.; Zoph, B.; and Shazeer, N. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120): 1–39.

Fei, Z.; Fan, M.; Yu, C.; Li, D.; and Huang, J. 2024. Scaling diffusion transformers to 16 billion parameters. *arXiv preprint arXiv:2407.11633*.

Huang, R.; Zhu, S.; Du, Y.; and Zhao, H. 2025. MoE-Loco: Mixture of Experts for Multitask Locomotion. *arXiv preprint arXiv:2503.08564*.

Huang, S.; Zhang, Z.; Liang, T.; Xu, Y.; Kou, Z.; Lu, C.; Xu, G.; Xue, Z.; and Xu, H. 2024. Mentor: Mixture-of-experts network with task-oriented perturbation for visual reinforcement learning. *arXiv preprint arXiv:2410.14972*.

Kanwisher, N. 2010. Functional specificity in the human brain: a window into the functional architecture of the mind. *Proceedings of the national academy of sciences*, 107(25): 11163–11170.

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

Kim, M. J.; Pertsch, K.; Karamcheti, S.; Xiao, T.; Balakrishna, A.; Nair, S.; Rafailov, R.; Foster, E.; Lam, G.; Sanketi, P.; et al. 2024. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*.

Li, Q.; Liang, Y.; Wang, Z.; Luo, L.; Chen, X.; Liao, M.; Wei, F.; Deng, Y.; Xu, S.; Zhang, Y.; Wang, X.; Liu, B.; Fu, J.; Bao, J.; Chen, D.; Shi, Y.; Yang, J.; and Guo, B. 2024a. CogACT: A Foundational Vision-Language-Action Model for Synergizing Cognition and Action in Robotic Manipulation. arXiv:2411.19650.

Li, S.; Wang, J.; Dai, R.; Ma, W.; Ng, W. Y.; Hu, Y.; and Li, Z. 2024b. Robonurse-vla: Robotic scrub nurse system based on vision-language-action model. *arXiv preprint arXiv:2409.19590*.

Li, X.; Liu, M.; Zhang, H.; Yu, C.; Xu, J.; Wu, H.; Cheang, C.; Jing, Y.; Zhang, W.; Liu, H.; et al. 2023. Vision-language foundation models as effective robot imitators. *arXiv preprint arXiv:2311.01378*.

Li, Y.; Jiang, S.; Hu, B.; Wang, L.; Zhong, W.; Luo, W.; Ma, L.; and Zhang, M. 2025. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, C.; Zhang, J.; Li, C.; Zhou, Z.; Wu, S.; Huang, S.; and Duan, H. 2025. TTF-VLA: Temporal Token Fusion via Pixel-Attention Integration for Vision-Language-Action Models. arXiv:2508.19257.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Ma, Y.; Yu, Z.; Lin, X.; Xie, W.; and Shen, L. 2024. BIG-MoE: Bypass Isolated Gating MoE for Generalized Multimodal Face Anti-Spoofing. *arXiv preprint arXiv:2412.18065*.

McClelland, J. L.; McNaughton, B. L.; and O'Reilly, R. C. 1995. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3): 419.

Mengara, A. G. M.; and Moon, Y.-k. 2025. CAG-MoE: Multimodal Emotion Recognition with Cross-Attention Gated Mixture of Experts. *Mathematics*, 13(12): 1–37.

Murphy, A. C.; Bertolero, M. A.; Papadopoulos, L.; Lydon-Staley, D. M.; and Bassett, D. S. 2020. Multimodal network dynamics underpinning working memory. *Nature communications*, 11(1): 3035.

Mustafa, B.; Riquelme, C.; Puigcerver, J.; Jenatton, R.; and Houlsby, N. 2022. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35: 9564–9576.

Nair, S.; Rajeswaran, A.; Kumar, V.; Finn, C.; and Gupta, A. 2022. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

O'Reilly, R. C.; and Rudy, J. W. 2001. Conjunctive representations in learning and memory: principles of cortical and hippocampal function. *Psychological review*, 108(2): 311.

Riquelme, C.; Puigcerver, J.; Mustafa, B.; Neumann, M.; Jenatton, R.; Susano Pinto, A.; Keysers, D.; and Houlsby, N. 2021. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34: 8583–8595.

Schneegans, S.; and Bays, P. M. 2017. Neural architecture for feature binding in visual working memory. *Journal of Neuroscience*, 37(14): 3913–3925.

Shentu, Y.; Wu, P.; Rajeswaran, A.; and Abbeel, P. 2024. From llms to actions: Latent codes as bridges in hierarchical robot control. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 8539–8546. IEEE.

Shine, J. M.; Koyejo, O.; Bell, P. T.; Gorgolewski, K. J.; Gilat, M.; and Poldrack, R. A. 2015. Estimation of dynamic functional connectivity using Multiplication of Temporal Derivatives. *NeuroImage*, 122: 399–407.

Spaak, E.; Watanabe, K.; Funahashi, S.; and Stokes, M. G. 2017. Stable and dynamic coding for working memory in primate prefrontal cortex. *Journal of neuroscience*, 37(27): 6503–6516.

Stone, A.; Xiao, T.; Lu, Y.; Gopalakrishnan, K.; Lee, K.-H.; Vuong, Q.; Wohlhart, P.; Kirmani, S.; Zitkovich, B.; Xia, F.; et al. 2023. Open-world object manipulation using pre-trained vision-language models. *arXiv preprint arXiv:2303.00905*.

Suzuki, K.; Kamiwano, Y.; Chiba, N.; Mori, H.; and Ogata, T. 2023. Multi-Timestep-Ahead Prediction with Mixture of Experts for Embodied Question Answering. In *International Conference on Artificial Neural Networks*, 243–255. Springer.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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

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

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

Wen, J.; Zhu, Y.; Zhu, M.; Tang, Z.; Li, J.; Zhou, Z.; Liu, X.; Shen, C.; Peng, Y.; and Feng, F. 2025b. DiffusionVLA: Scaling Robot Foundation Models via Unified Diffusion and Autoregression. In *Forty-second International Conference on Machine Learning*.

Wu, H.; Jing, Y.; Cheang, C.; Chen, G.; Xu, J.; Li, X.; Liu, M.; Li, H.; and Kong, T. 2023. Unleashing large-scale video generative pre-training for visual robot manipulation. *arXiv preprint arXiv:2312.13139*.

Wu, P.; Shentu, Y.; Yi, Z.; Lin, X.; and Abbeel, P. 2024. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 12156–12163. IEEE.

Wu, W.; Liu, F.; Li, H.; Hu, Z.; Dong, D.; Chen, C.; and Wang, Z. 2025. Mixture-of-Experts Meets In-Context Reinforcement Learning. *arXiv preprint arXiv:2506.05426*.

Xue, F.; Zheng, Z.; Fu, Y.; Ni, J.; Zheng, Z.; Zhou, W.; and You, Y. 2024. Openmoe: An early effort on open mixture-of-experts language models. *arXiv preprint arXiv:2402.01739*.

Yang, Z.; Chai, Y.; Jia, X.; Li, Q.; Shao, Y.; Zhu, X.; Su, H.; and Yan, J. 2025. DriveMoE: Mixture-of-Experts for Vision-Language-Action Model in End-to-End Autonomous Driving. *arXiv preprint arXiv:2505.16278*.

Yu, Q.; Huang, S.; Yuan, X.; Jiang, Z.; Hao, C.; Li, X.; Chang, H.; Wang, J.; Liu, L.; Li, H.; et al. 2024. UniAff: A unified representation of affordances for tool usage and articulation with vision-language models. *arXiv preprint arXiv:2409.20551*.

Yuan, K.; and Li, Z. 2022. Multi-expert synthesis for versatile locomotion and manipulation skills. *Frontiers in Robotics and AI*, 9: 970890.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11975–11986.

Zhao, H.; Song, W.; Wang, D.; Tong, X.; Ding, P.; Cheng, X.; and Ge, Z. 2025. More: Unlocking scalability in reinforcement learning for quadruped vision-language-action models. *arXiv preprint arXiv:2503.08007*.

Zhao, T. Z.; Kumar, V.; Levine, S.; and Finn, C. 2023. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*.

Zhou, Y.; Lei, T.; Liu, H.; Du, N.; Huang, Y.; Zhao, V.; Dai, A. M.; Le, Q. V.; Laudon, J.; et al. 2022. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35: 7103–7114.

Zitkovich, B.; Yu, T.; Xu, S.; Xu, P.; Xiao, T.; Xia, F.; Wu, J.; Wohlhart, P.; Welker, S.; Wahid, A.; et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, 2165–2183. PMLR.

Zoph, B.; Bello, I.; Kumar, S.; Du, N.; Huang, Y.; Dean, J.; Shazeer, N.; and Fedus, W. 2022. St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*.