

# ReconVLA: Reconstructive Vision-Language-Action Model as Effective Robot Perceiver

Wenxuan Song<sup>1</sup>, Ziyang Zhou<sup>1</sup>, Han Zhao<sup>2,3</sup>, Jiayi Chen<sup>1</sup>, Pengxiang Ding<sup>2,3</sup>, Haodong Yan<sup>1</sup>, Yuxin Huang<sup>1</sup>, Feilong Tang<sup>4</sup>, Donglin Wang<sup>2</sup>, Haoang Li<sup>1</sup>

<sup>1</sup>The Hong Kong University of Science and Technology (Guangzhou)

<sup>2</sup>Westlake University <sup>3</sup>Zhejiang University <sup>4</sup>Monash University

## Abstract

Recent advances in Vision-Language-Action (VLA) models have enabled robotic agents to integrate multimodal understanding with action execution. However, our empirical analysis reveals that current VLAs struggle to allocate visual attention to target regions. Instead, visual attention is always dispersed. To guide the visual attention grounding on the correct target, we propose **ReconVLA**, a reconstructive VLA model with an implicit grounding paradigm. Conditioned on the model’s visual outputs, a diffusion transformer aims to reconstruct the gaze region of the image, which corresponds to the target manipulated objects. This process prompts the VLA model to learn fine-grained representations and accurately allocate visual attention, thus effectively leveraging task-specific visual information and conducting precise manipulation. Moreover, we curate a large-scale pretraining dataset comprising over 100k trajectories and 2 million data samples from open-source robotic datasets, further boosting the model’s generalization in visual reconstruction. Extensive experiments in simulation and the real world demonstrate the superiority of our implicit grounding method, showcasing its capabilities of precise manipulation and generalization. Our project page is <https://zionchow.github.io/ReconVLA/>.

## 1 Introduction

Recent progress in Vision-Language Models (VLMs) (Awadalla et al. 2023; Liu et al. 2024b) has demonstrated their potential to bridge perceptual and linguistic modalities effectively. Building upon these advances, Vision-Language-Action (VLA) models (Brohan et al. 2023; Zitkovich et al. 2023; Octo Model Team et al. 2024; Niu et al. 2024; Song et al. 2024; Kim et al. 2024) have extended this capability to action execution by integrating multimodal understanding. Benefit of billions of parameters and pre-training on large-scale robot datasets (O’Neill et al. 2024; Fang et al. 2024), these models have shown promise in enabling generalizable skills.

Accurate visual grounding is fundamental to enable precise grasping of VLAs, especially in cluttered environments and long-horizon tasks. To analyze the visual grounding behavior during predicting actions, we visualize the attention map on visual inputs. The results show that traditional VLA models often exhibit dispersed visual attention (Figure 4 Row 1), failing to focus precisely on the target object, which

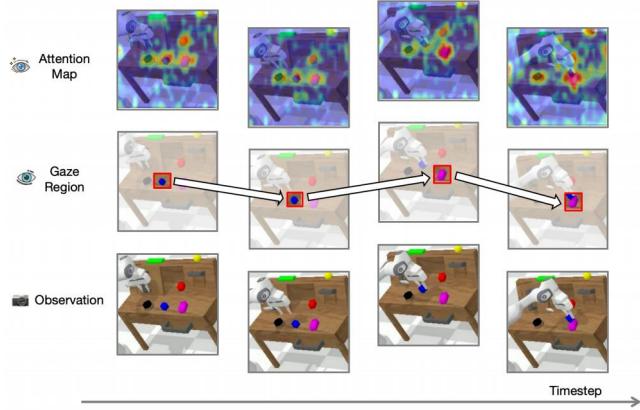
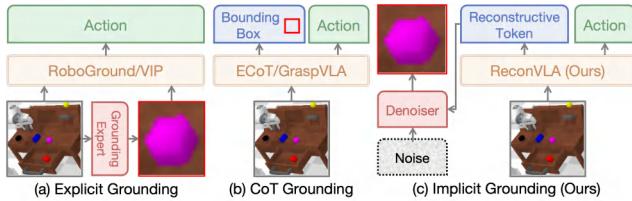


Figure 1: **Visualization of the observation, gaze region, and attention map.** For a long-horizon task “stack blocks” that requires the arm to lift the blue block and put it on the pink one. Although there are several distractors, our model adaptively adjusts the gaze region, guiding the allocation of visual attention to the right target. With the precise visual grounding, it sequentially manipulates different target objects and successfully completes the task.

may further lead to manipulating incorrect objects. The finding raises a critical question: *how can VLA models refine visual attention allocation and further improve visual grounding capabilities?*

Previous visual grounding methods for VLAs usually explicitly input grounded images (Huang et al. 2025; Li et al. 2025) or output bounding boxes (Zawalski et al. 2024; Deng et al. 2025) in a chain-of-thought (CoT) manner. These methods enhance the perception of target regions and improve spatial awareness, while they do not fundamentally refine the attention allocation. Inspired by reconstructive visual instruction tuning (Wang et al. 2024), we introduce an auxiliary visual reconstruction module implemented as a lightweight diffusion transformer (Peebles and Xie 2022). This module is conditioned on the visual outputs of the VLA model and aims to reconstruct the target manipulated region from noise. This process prompts the VLA model to learn fine-grained representations with region-specific information, thereby focusing visual attention on the correct area.



**Figure 2: Conceptual comparison between different paradigms.** (a) **Explicit Grounding:** Employing an external grounding expert and inputting entire images and cropped images (Huang et al. 2025; Li et al. 2025). (b) **CoT Grounding:** Outputting coordinates of bounding boxes before action in a chain-of-thought (CoT) manner (Zawalski et al. 2024; Deng et al. 2025). (c) **Implicit Grounding:** Our ReconVLA directly leverages crucial regions as implicit visual supervision for visual outputs, called reconstructive tokens, through a reconstruction process.

As shown in Figure 1, this mechanism is analogous to the gaze behavior of the human eye, where the eye perceives a small, focused area with sharp clarity while the surrounding regions remain blurred (Stewart 2020). Thus, the target manipulated region is named gaze region.

However, similar to their VLM backbone (Liu et al. 2024b), conventional VLA models are finetuned on vision-language understanding tasks and generate actions in an autoregressive manner, lacking visual generation capabilities. To address this limitation, we curated a pretraining dataset containing over 100k trajectories and 2 million data samples. We select several open-source robotic datasets (Walke et al. 2024; Liu et al. 2024a; Mees et al. 2021) and design an automatic data processing by Grounding DINO (Liu et al. 2024c) to produce pairwise entire images and images of target manipulated regions. Pretraining on this large-scale dataset significantly enhances the model’s generalization ability in visual generation.

By leveraging the aforementioned techniques, we develop the **Reconstructive Vision-Language-Action Model (ReconVLA)**. It takes current images, language instructions, and robot proprioception as inputs. During training, the gaze regions of input images are processed into latent tokens via a frozen visual tokenizer, which preserves detailed visual information and enables high-fidelity reconstruction. To better learn the latent information, we train a diffusion transformer learning to recover the latent tokens guided by reconstructive tokens. The diffusion denosing effectively models the conditional distribution of visual observation.

Experiments in long-horizon tasks demonstrate that our implicit grounding method is more effective than other visual grounding paradigms. Besides, visualizations of visual attention prove that our ReconVLA demonstrates directive visual attention and leads to precise manipulation. Then, ablation studies prove the generalization through large-scale pretraining. Comprehensive comparison with other popular methods shows that our ReconVLA yields superior performance. Finally, we conduct real-world experiments and evaluate the generalization to unseen objects. This demon-

strates that our ReconVLA has the potential to facilitate the real-world deployment of VLAs.

In summary, our key contributions are as follows:

- We propose ReconVLA, a reconstructive VLA model with an implicit grounding paradigm. The reconstruction of gaze regions prompts the model toward precise visual attention allocation and fine-grained representation learning, thereby enhancing visual grounding capabilities and executing precise manipulation.
- We constructed a large-scale robot pretraining dataset, containing more than 100k trajectories, 2 million data samples. Pretraining on this dataset enhances the model’s generalization of visual reconstruction capabilities.
- Extensive experiments in simulation and the real world show the superiority of our implicit grounding methods and the capabilities of precise manipulation and generalization for unseen targets.

## 2 Related Work

**Action-centric Vision-language-action Models.** Building upon foundational advancements on pretrained VLMs (Beyer et al. 2024; Lu et al. 2024; Liu et al. 2024b), VLAs (Brohan et al. 2023; Zitkovich et al. 2023; Ding et al. 2024; Octo Model Team et al. 2024; Tong et al. 2025; Zhao et al. 2025a,b; Cui et al. 2025; Song et al. 2025c,a) learn to generate executable actions supervised by actions. Within them, RoboFlamingo (Li et al. 2024) models sequential history information with an explicit policy head. OpenVLA (Kim et al. 2024) is the first open-source VLA model with large-scale robotic pretraining (O’Neill et al. 2024). VLAs (Zhao et al. 2025b) expands the modality with audio. UniVLA (Bu et al. 2025) learns task-centric latent actions from web-scale videos and adapts to different downstream tasks. These models only supervise action outputs, while our ReconVLA supervise visual outputs as auxiliary tasks, thus enhancing visual perception.

**Generative Methods for Manipulation.** Previous works have explored image or video generation models for robotic control. Unipi (Du et al. 2023) first generates future images and extracts action from generated images. SuSIE (Black et al. 2024) generates subgoals with an image-editing diffusion model and executes them using a language-agnostic policy. CLOVER (Bu et al. 2024) produces visual plans to guide a closed-loop policy using error measurements. GR-1 (Wu et al. 2024) first combines generative methods with VLAs. It proposes a GPT-style model for visual robot manipulation by leveraging large-scale video pre-training to predict future images and robot actions. 3D-VLA (Zhen et al. 2024) further integrates depth information as guidance for vision-language-action reasoning and planning. GEVRM (Zhang et al. 2025) generates future images for a goal-conditioned policy in a closed-loop manner. These methods (Tian et al. 2024; Guo et al. 2024; Wang et al. 2025; Cen et al. 2025) predict future frames to learn from dynamics, thereby enhancing the model’s planning capability. In contrast, our approach reconstructs target regions of the current image to achieve precise perception and manipulation.

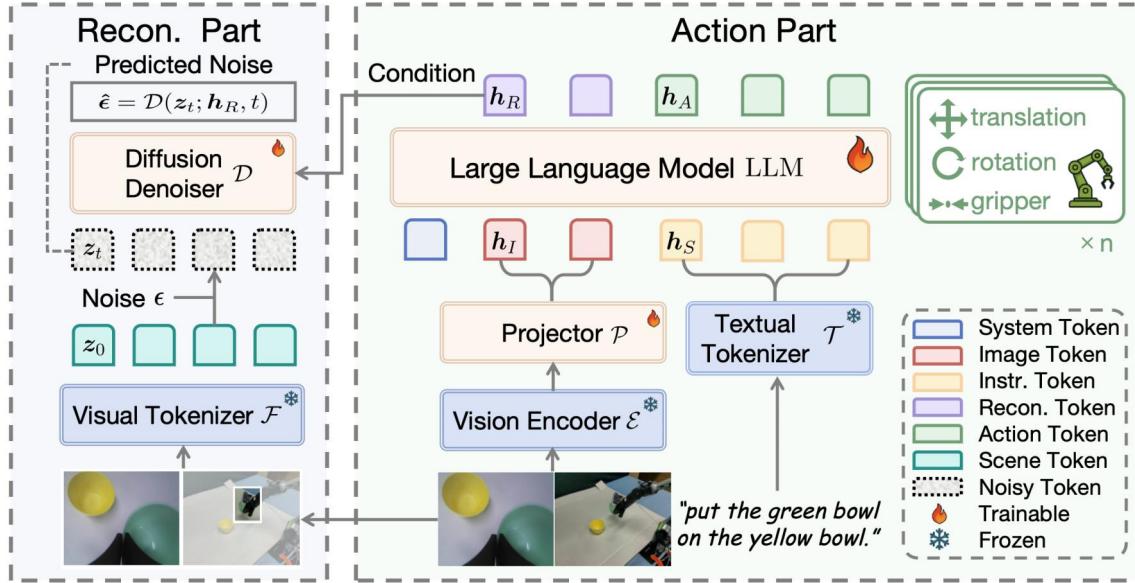


Figure 3: **Architecture of our ReconVLA.** Our model consists of a reconstructive part and an action part. The input includes multi-view images and a text instruction. For the action part, the model outputs discrete action tokens. For the reconstruction part, our ReconVLA is guided to output reconstructive tokens, which are conditions of the denoising process to reconstruct the scene tokens  $z_0$  from noisy  $z_t$ . The scene tokens are tokenized images of gaze regions. This supervision enables our ReconVLA to enhance visual grounding and fine-grained comprehension capabilities, which contribute to precise manipulation.

**Visual Grounding Methods for Manipulation.** Explicit grounding methods often take the grounded image as extra inputs to serve as auxiliary observation (Figure 2 (a)). RoboGround (Huang et al. 2025) employs LISA (Lai et al. 2024) as a high-level segmenter to extract the target object and background based on instructions and feeds them as part of the observation into the VLA model. Similarly, VIP (Li et al. 2025) uses YOLOv11 (Khanam and Hussain 2024) to segment the target object, which is then enlarged and provided to a transformer-based policy. However, these models rely on external expert models and do not fundamentally enhance the visual grounding capabilities of the policy itself. ECoT (Zawalski et al. 2024) and GraspVLA (Deng et al. 2025) (Figure 2 (b)) adopt a chain-of-thought approach, sequentially outputting bounding boxes and actions, which simultaneously trains the grounding capability and provides richer information for action output through causal attention. In contrast to these prior methods, our ReconVLA directly reconstructs the target manipulation region from the visual outputs (Figure 2 (c)), thereby implicitly performing grounding while encouraging the model to learn fine-grained representations of the target region. This process emulates the human eye’s spontaneous ability to focus on salient regions within the field of view.

### 3 Method

#### 3.1 Preliminaries

To establish the foundation of our method, we first formalize the typical formulation and architecture of VLA models in the context of robotic manipulation (Brohan et al.

2023; Zitkovich et al. 2023; Kim et al. 2024; Zhao et al. 2025b; Song et al. 2025b). Given a pair of images and text instructions  $(I, S)$ , the VLA model  $\Lambda$  predicts the actions  $\mathcal{A} = \Lambda(I, S)$ .

**Architecture.** A regular VLA mainly consists of a large language model LLM, a vision encoder  $\mathcal{E}$ , the tokenizer  $\mathcal{T}$ , and an action detokenizer  $\mathcal{Q}$ . The tuple  $(I, S)$  are processed into image tokens  $h_I$  and text tokens  $h_S$  by  $\mathcal{E}$  and  $\mathcal{T}$  separately. These tokens are then fed into the LLM to generate action tokens  $a$ . Finally, the action detokenizer  $\mathcal{Q}$  maps  $a$  into executable action  $\mathcal{A}$  for robotic control. The whole process can be formulated as:

$$\mathcal{A} = \mathcal{Q}(a) = \mathcal{Q}(\text{LLM}(h_I, h_S)) = \mathcal{Q}(\text{LLM}(\mathcal{E}(I), \mathcal{T}(S))). \quad (1)$$

Specifically, the action tokens are generated in an autoregressive manner:

$$p(a) = \prod_{i=1}^N p_{\text{LLM}}(a_i | a_{1 \sim i-1}; h_I; h_S), \quad (2)$$

where  $i$  denotes the  $i$ -th action token and  $N$  denotes the total number of action tokens.

#### 3.2 Reconstructive Vision-Language-Action Model

With observation of the dispersed attention, we aim to guide VLAs’ visual attention to focus on the correct target. Our philosophy is to construct an auxiliary visual supervision, realized by setting a reconstructive visual signal. The supervising signal serves as conditions to guide a diffusion de-noising process to reconstruct the target manipulated region.

Formally, we present the Reconstructive Vision-Language-Action Model (ReconVLA), grounded in this framework.

**Reconstruction Target.** When manipulating objects, humans receive a global view of the scene. However, visual perception primarily focuses on a small part of it, namely the region intended to be manipulated. This behavior is known as gaze. Similarly, the reconstruction target of our ReconVLA is the target manipulated region, which we refer to as the **gaze region**. The gaze region not only helps the model focus on the correct target among multiple affordable regions, but also enhances the detailed perception of these regions. Besides, the mechanism implicitly facilitates sub-task planning in long-horizon tasks by focusing on and switching to different sub-goals.

**Loss Function.** The overall training objectives of ReconVLA include **(i)** the autoregressive action prediction supervised by demonstration data, and **(ii)** another reconstructive term supervised by the visual features of gaze regions, *i.e.*,  $\mathcal{L}_{\text{ReconVLA}} = \mathcal{L}_{\text{VLA}}^{\text{action}} + \mathcal{L}_{\text{VLA}}^{\text{visual}}$ , where the  $\mathcal{L}_{\text{VLA}}^{\text{action}}$  is cross-entropy loss and the  $\mathcal{L}_{\text{VLA}}^{\text{visual}}$  is a measurement between reconstructive tokens  $\mathbf{h}_R$  and reconstruction targets  $I'$ .

**Latent Visual Reconstruction.** To construct a region-specific visual supervision signal from an RGB input with spatial information redundancy (He et al. 2022), we design a denoising process to reconstruct tokens with low-level features of gaze regions. This process encourages the model to fully capture intrinsic features instead of cloning explicit RGB values (Chen et al. 2023; Song and Ermon 2020; Karras et al. 2022; Yang et al. 2024b).

Figure 3 illustrates that our ReconVLA utilizes the visual tokenizer  $\mathcal{F}$  to extract target scene tokens  $\mathbf{z}_0 = \mathcal{F}(I')$ . Specifically, we employ a continuous variational autoencoder (VAE) (Kingma and Welling 2022) in (Rombach et al. 2022) as the visual tokenizer  $\mathcal{F}$  because of its visual fidelity and ability to capture fine-grained image features. The denoiser  $\mathcal{D}$  tries to predict the noise and recover  $\mathbf{z}_0$  from noisy tokens  $\mathbf{z}_t$  conditioned on the reconstructive tokens  $\mathbf{h}_R = \text{LLM}(\mathbf{h}_I)$ . The reconstructive objective function is formalized following a diffusion process (Ho, Jain, and Abbeel 2020):

$$\mathcal{L}_{\text{VLA}}^{\text{visual}}(\mathbf{h}_R, I') = \mathbb{E}_{t, \epsilon} [ \|\mathcal{D}(\mathbf{z}_t; \mathbf{h}_R, t) - \epsilon\|^2 ], \quad (3)$$

where  $t$  denotes the diffusion timesteps. The denoiser  $\mathcal{D}$  consists of a stack of Transformer encoder blocks (Vaswani 2017) with self-attention modules to capture the correlations between noisy tokens and reconstructive tokens.

To ensure that the VLA model processes visual tokens corresponding to the instructed target, it is necessary to guarantee that image tokens attend to instruction tokens. Thus, we prepend a set of instruction tokens before the image tokens, enabling the image tokens to fuse information from these prefix texts through causal attention. Experimental results show that this interleaved format achieves our objective without degrading the model’s inherent language modeling capability.

**Implementation Details.** In this paper, we construct our ReconVLA based on a pretrained vision-language model

LLaVA-7b (Liu et al. 2024b), which uses Qwen2-7b (Yang et al. 2024a) as the LLM backbone and siglip-so400m-patch14-384 (Zhai et al. 2023) as the vision encoder.

### 3.3 Visual Pretraining

The reconstruction capability of the VLA model is inherently limited, as its VLM backbone is primarily trained on vision-language understanding tasks. To enhance its ability to ground and reconstruct specific regions, we design a pretraining process for reconstruction tasks on a large-scale robot dataset.

**Dataset.** To build a foundational reconstruction capability, we constructed the pre-training dataset based on large-scale open-source robotic datasets BridgeData V2 (Walke et al. 2023), along with high-quality simulation datasets LIBERO (Liu et al. 2024a) and CALVIN (Mees et al. 2021). Given an image-text pair, we fine-tune Grounding DINO (Liu et al. 2024c), which is the state-of-the-art open-vocabulary object detector, to segment out the gaze region that the robot is instructed to interact with. The cropped images and original images are organized in a pairwise manner. In this way, we obtain an annotated visual pretraining dataset containing over 100k trajectories and 2 million samples.

**Training.** During the pretraining process, we perform gradient backpropagation both on the reconstruction loss and action loss to keep the consistency of the optimization target. This process equips our VLM with generalized visual reconstruction capabilities and facilitates the model’s deployment to diverse environments and tasks. After pretraining, we finetune our model on specific tasks to precisely align vision-language comprehension and visual reconstruction capabilities with manipulation capabilities on the corresponding action space.

## 4 Experiments

In this section, we structure the experiments to answer the following questions:

- Does our **implicit grounding** approach outperform other visual grounding paradigms? (see Section 4.2)
- Does the gazing mechanism contribute to visual grounding and further improve the precise manipulation? (see Section 4.3)
- Does our proposed **pretraining** stage improve the generalization of visual generation, and how do other proposed **key designs** in ReconVLA influence the overall performance? (see Section 4.4)
- Can ReconVLA effectively manage long-horizon tasks compared with other competitive methods? (see Section 4.5)
- Can ReconVLA realize **generalized** manipulation on **unseen** targets in **real-world** tasks? (see Section 4.6)

### 4.1 Simulation Environment

The CALVIN benchmark (Mees et al. 2021) is built on top of the PyBullet (Coumans and Bai 2016–2019) simulator and involves a Franka Panda Robot arm that manipulates the

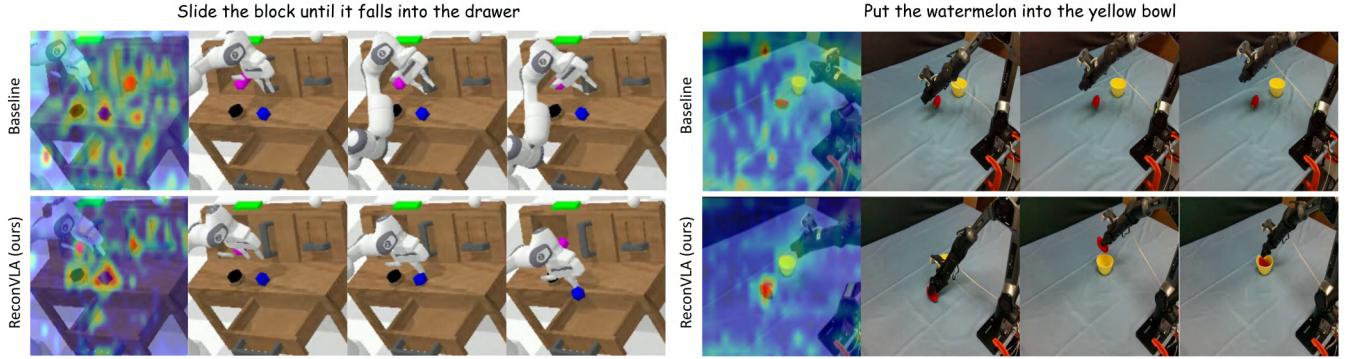


Figure 4: **Qualitative comparison of attention maps on CALVIN (Mees et al. 2021) and the real world.** **Row 1:** The baseline exhibits dispersed attention patterns or predominantly attends to an incorrect region, leading to inaccurate actions. **Row 2:** With auxiliary visual supervision signals, ReconVLA forces the model to *focus on specific image contents* with higher attention values and precisely move to the target region, thus successfully completing the task.

Paradigm	Success Rate (%)					
	1/5	2/5	3/5	4/5	5/5	Avg. Len
Baseline	88.8	76.1	63.7	57.0	49.0	3.36
EG	94.4	82.5	70.9	62.2	50.2	3.61
CG	47.0	14.3	1.6	0.0	0.0	0.63
IG (ours)	<b>95.6</b>	<b>87.6</b>	<b>76.9</b>	<b>69.3</b>	<b>64.1</b>	<b>3.95</b>

Table 1: Comparison among different paradigms, including Explicit Grounding (EG), CoT Grounding (CG), and our Implicit Grounding (IG). The comparison is conducted on the CALVIN ABC→D.

scene. CALVIN consists of 34 tasks and 4 different environments (A, B, C and D). The CALVIN long-horizon challenge is a sequential task comprising five subtasks. We report the success rates for each subtask and the average completed length across all five tasks. The method is evaluated over 500 rollouts to ensure a fair comparison. The metrics of CALVIN are the success rates of each sub-task and the average length of all sequential 5 sub-tasks.

## 4.2 Paradigm Comparison

We implement different visual grounding paradigms in Figure 2 on the same baseline to conduct a fair comparison.

**Explicit Grounding (EG).** We choose a finetuned YOLOv11 (Khanam and Hussain 2024) as the detector to recognize the target object at each timestep. Then we crop out the recognized object region from the image and resize it. Then the resized and original images are jointly fed to the VLA model to guide object manipulation.

**Chain-of-Thought Grounding (CG).** For data preparation, we preprocess the images with the detector to get the coordinates of the bounding box. Then, we reformulate the training dataset and modify the outputs in a CoT format: Bbox [x1 x2 y1 y2] + action sequence. The input remains the original images. In this way, the VLA model learns to

ground the target object and output actions with the grounding information.

**Results.** As shown in Table 1, EG gets relatively higher success rates than baseline. This indicates that explicit grounding as input helps better comprehension of spatial relationships. However, the simple concatenation of entire and cropped images introduces visual information redundancy, which limits model performance. CG performance is even worse. This suggests that bounding boxes in coordinate form are insufficient to effectively guide the model in precisely manipulating target locations. Additionally, directly outputting precise coordinates and action values together presents training challenges for VLA models.

Our implicit grounding method gets the highest success rates, which demonstrates the superiority of our method over other paradigms. From the perspective of *training mechanism*, the advantage stems from our implicit grounding learning framework, which enables the model to precisely attend to visual information at target objects, thereby achieving precise manipulation. From the perspective of *architecture*, our model directly supervises visual outputs, eliminating the need for additional inputs or outputs. This design yields a simple yet effective training and inference pipeline.

## 4.3 In-depth Analysis

To better explore the influence of the gazing mechanism, we conduct qualitative experiments of visual attention and its effect on fine-grained manipulation tasks.

**Attention Visualization** Figure 4 demonstrates that the implementation of  $\mathcal{L}_{\text{VLA}}$  enables the alignment of attention closely with the gaze region, which corresponds to the target object. For the instruction “*put the watermelon into the yellow bowl*”, the attention of baseline is highly dispersed, with the third-view image attention mostly focused on irrelevant positions and resulting in the task failure. In contrast, ReconVLA successfully concentrates attention on the correct target, *i.e.*, the watermelon. This demonstrates that our method brings precise visual grounding, which facilitates task success.

Recon.	Gaze Region	Pretrain	Splits	Task completed in a row (%)					Average Length
				1	2	3	4	5	
✓	✓	✓	ABC→D	<b>95.6</b>	<b>87.6</b>	<b>76.9</b>	<b>69.3</b>	<b>64.1</b>	<b>3.95</b>
✓	✓	✗	ABC→D	96.8	86.9	76.9	64.9	58.2	3.85
✓	✗	✗	ABC→D	89.8	80.3	67.7	56.6	46.5	3.42
✗	✗	✗	ABC→D	88.8	76.1	63.7	57.0	49.0	3.36

Table 2: Ablation results of the proposed techniques using the reconstructive part, gaze region, and pretraining.

Category	Method	Splits	Success Rate (%)					Avg. Len
			1/5	2/5	3/5	4/5	5/5	
Generative Methods	UniPi (Du et al. 2023) ( <i>NIPS’23</i> )	ABC→D	56.0	16.0	8.0	8.0	4.0	0.92
	SuSIE (Black et al. 2024) ( <i>ICLR’24</i> )	ABC→D	87.0	69.0	49.0	38.0	26.0	2.69
	GEVRM (Zhang et al. 2025) ( <i>ICLR’25</i> )	ABC→D	92.0	70.0	54.0	41.0	26.0	2.83
	GR-1 (Wu et al. 2024) ( <i>ICLR’24</i> )	ABC→D	85.4	71.2	59.6	49.7	40.1	3.06
	Vidman (Wen et al. 2024) ( <i>NIPS’24</i> )	ABC→D	91.5	76.4	68.2	59.2	46.7	3.42
Large VLA Models	CLOVER (Bu et al. 2024) ( <i>NIPS’24</i> )	ABC→D	96.0	83.5	70.8	57.5	45.4	3.53
	VLAS (Zhao et al. 2025b) ( <i>ICLR’25</i> )	ABC→D	87.2	64.2	40.9	28.1	19.6	2.40
	RoboFlamingo (Li et al. 2024) ( <i>ICLR’24</i> )	ABC→D	82.4	61.9	46.6	33.1	23.5	2.47
	OpenVLA (Kim et al. 2024) ( <i>CoRL’24</i> )	ABC→D	91.3	77.8	62.0	52.1	43.5	3.27
Reconstructive Methods	UniVLA (Bu et al. 2025) ( <i>RSS’25</i> )	ABC→D	95.5	85.8	75.4	66.9	56.5	3.80
	<b>ReconVLA (ours)</b>	ABC→D	<b>95.6</b>	<b>87.6</b>	<b>76.9</b>	<b>69.3</b>	<b>64.1</b>	<b>3.95</b>

Table 3: Comparison with various manipulation models on CALVIN ABC→D in success rates and average length.

Category	Method	Splits	Success Rate (%)					Avg. Len
			1/5	2/5	3/5	4/5	5/5	
Generative Methods	3D-VLA (Zhen et al. 2024) ( <i>ICML’24</i> )	ABCD→D	44.7	16.3	8.1	1.6	0	0.70
	GR-1 (Wu et al. 2024) ( <i>ICLR’24</i> )	ABCD→D	94.9	89.6	84.4	78.9	73.1	4.21
Large VLA Models	VLAS (Zhao et al. 2025b) ( <i>ICLR’25</i> )	ABCD→D	94.2	84.0	73.2	64.3	54.6	3.70
	RoboFlamingo (Li et al. 2024) ( <i>ICLR’24</i> )	ABCD→D	96.4	89.6	82.4	74.0	66.0	4.08
Reconstructive Methods	<b>ReconVLA (ours)</b>	ABCD→D	<b>98.0</b>	<b>90.0</b>	<b>84.5</b>	<b>78.5</b>	<b>70.5</b>	<b>4.23</b>

Table 4: Comparison with various manipulation models on CALVIN ABCD→D in success rates and average length.

**Precise Manipulation.** Among all tasks, the “stack block” task is the most challenging, which requires the robot to lift one block and precisely stack it on the other block. While our baseline achieves only 59.3% on this task, our gazing mechanism attains a success rate as high as 79.5%, representing a **20.2%** increase. This significant improvement highlights the enhanced action accuracy of our gazing mechanism through precise visual grounding.

#### 4.4 Ablation Study

We perform ablation studies of the proposed techniques using the reconstructive part, gaze region, and pretraining on large-scale robotic datasets in Table 2. We observe that **pretraining** leads to a significant improvement in success rates. This is because, in unseen test environments, grounding the target object and performing reconstruction is inherently challenging and poses a generalization challenge to the model’s generative capability. Pretraining on large-scale datasets substantially enhances the model’s generalization ability during visual reconstruction. Furthermore, reconstructing the **gaze region** to be manipulated, rather than

the entire image, proves to be more effective. This guides the model’s visual attention to focus on the target object, thereby avoiding manipulation of incorrect targets. Notably, models trained to **reconstruct** the entire image still outperform the baseline, which can be attributed to the enhanced holistic visual attention. However, in unseen scenarios, reconstructing the entire image with pixel redundancy is extremely challenging, which further limits the performance improvements.

#### 4.5 Comparison with State-of-the-arts

**Compared Methods.** We compare our model with generative methods that predict future images (UniPi, SuSIE, CLOVER, 3D-VLA, GR-1, Vidman, GEVRM), and large VLA models (RoboFlamingo, VLAS, OpenVLA, UniVLA), as introduced in Section 2.

**Results.** In the basic ABCD→D tasks, our ReconVLA achieves competitive performance, successfully completing an average of 4.23 out of 5 consecutive tasks, with a success rate of 98.0% on the first task. This indicates that our gazing mechanism provides a flexible planning ability

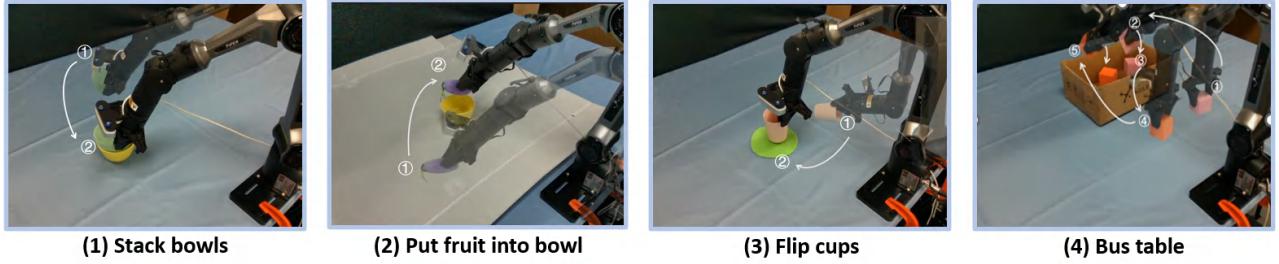


Figure 5: **Real-world Setup of four representative tasks.** We use a 6-DoF AgileX PiPer robotic arm with a 1-DoF parallel gripper and a RealSense D515 depth camera as Eye-on-Base and an ORBBEC Dabai depth camera as Eye-on-Hand. We selected four representative and practically meaningful tasks: (1) *Stack bowls*, (2) *Put fruit into bowl*, (3) *Flip cups*, (4) *Bus table*.

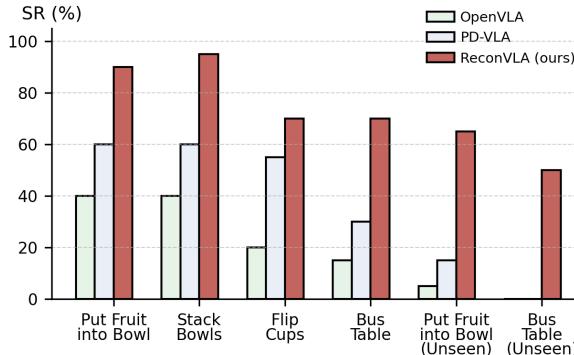


Figure 6: **Real-world multi-task results.** We report success rates (SR) across 4 different tasks as well as 2 unseen tasks.

to realize a better operation schedule in long-horizon tasks. The ABC→D tasks challenge the generalization for unseen backgrounds. Our method surpasses all generative methods, including the popular GR-1 with over 20% success rates on the last sub-task. This indicates that besides the generative model that predicts the future images, enhancing the perception of the current observation is equally valuable for robot manipulation. With comparable parameter amounts, our method outperforms OpenVLA by 20.6% and UniVLA by 7.6% on the last sub-task, which indicates the effectiveness of our implicit grounding learning strategy.

#### 4.6 Multi-task Experiments in the Real World

**Setup.** We conducted real-world experiments using a 6-DoF AgileX PiPer robotic arm with a 1-DoF parallel gripper. Besides, we use a RealSense D515 depth camera as Eye-on-Base and an ORBBEC Dabai depth camera as Eye-on-Hand for visual inputs.

**Tasks.** We select four representative tasks: *Put fruit into bowl*, *Stack bowls*, *Flip cups*, and *Bus table*. To enhance the model’s generalization ability, each task includes variations in target objects and background colors. We collect 150 trajectories per task on average. For evaluation, each model is tested on each task with 20 trials, and the success rate is used

as the performance metric. For unseen tasks, we replace the target object with unseen ones.

**Results.** ReconVLA consistently outperforms both popular OpenVLA and strong PD-VLA across the four real-world tasks, achieving the highest success rate in each case. In particular, ReconVLA achieves a success rate close to or exceeding 90% on both the *Put Fruit into Bowl* and *Stack Bowls* tasks. OpenVLA shows limited effectiveness in executing fine-grained manipulation tasks (e.g., *flip cup* and *bus table*), while our ReconVLA achieves significant performance improvements through precise visual grounding.

In unseen tasks, where the target objects are absent from the training data, both OpenVLA and PD-VLA methods exhibit nearly 0% success rates. Benefiting from large-scale mix-data pretraining, our ReconVLA can still successfully ground the target objects and complete the intended actions, demonstrating the advantage of our approach’s **visual generalization** capability.

## 5 Conclusion

In this paper, we analyze and reveal the dispersed visual attention in traditional VLAs, which limits the precise manipulation. Then, we propose a reconstructive vision-language-action model (ReconVLA), a novel framework trained in an implicit grounding paradigm. Our ReconVLA successfully realizes accurate visual attention allocation and further enhances manipulation skills. We further construct a large-scale pretraining dataset for ReconVLA to generalize on diverse scenes and unseen objects. Extensive experiments in simulation and the real world show the superiority of our implicit grounding methods.

## References

- Awadalla, A.; Gao, I.; Gardner, J.; Hessel, J.; Hanafy, Y.; Zhu, W.; Marathe, K.; Bitton, Y.; Gadre, S.; Sagawa, S.; Jitsev, J.; Kornblith, S.; Koh, P. W.; Ilharco, G.; Wortsman, M.; and Schmidt, L. 2023. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. *arXiv preprint arXiv:2308.01390*.
- Beyer, L.; Steiner, A.; Pinto, A. S.; Kolesnikov, A.; Wang, X.; Salz, D.; Neumann, M.; Alabdulmohsin, I.; Tschannen,

- M.; Bugliarello, E.; et al. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*.
- Black, K.; Nakamoto, M.; Atreya, P.; Walke, H. R.; Finn, C.; Kumar, A.; and Levine, S. 2024. Zero-Shot Robotic Manipulation with Pre-Trained Image-Editing Diffusion Models. In *The Twelfth International Conference on Learning Representations*.
- Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Dabis, J.; Finn, C.; ; et al. 2023. RT-1: Robotics Transformer for Real-World Control at Scale. *Proceedings of Robotics: Science and Systems*.
- Bu, Q.; Yang, Y.; Cai, J.; Gao, S.; Ren, G.; Yao, M.; Luo, P.; and Li, H. 2025. UniVLA: Learning to Act Anywhere with Task-centric Latent Actions. *arXiv preprint arXiv:2505.06111*.
- Bu, Q.; Zeng, J.; Chen, L.; Yang, Y.; Zhou, G.; Yan, J.; Luo, P.; Cui, H.; Ma, Y.; and Li, H. 2024. Closed-loop visuomotor control with generative expectation for robotic manipulation. *Advances in Neural Information Processing Systems*, 37: 139002–139029.
- Cen, J.; Yu, C.; Yuan, H.; Jiang, Y.; Huang, S.; Guo, J.; Li, X.; Song, Y.; Luo, H.; Wang, F.; et al. 2025. WorldVLA: Towards Autoregressive Action World Model. *arXiv preprint arXiv:2506.21539*.
- Chen, M.; Huang, K.; Zhao, T.; and Wang, M. 2023. Score Approximation, Estimation and Distribution Recovery of Diffusion Models on Low-Dimensional Data. *arXiv:2302.07194*.
- Coumans, E.; and Bai, Y. 2016–2019. PyBullet, a Python module for physics simulation for games, robotics and machine learning. <http://pybullet.org>.
- Cui, C.; Ding, P.; Song, W.; Bai, S.; Tong, X.; Ge, Z.; Suo, R.; Zhou, W.; Liu, Y.; Jia, B.; et al. 2025. OpenHelix: A Short Survey, Empirical Analysis, and Open-Source Dual-System VLA Model for Robotic Manipulation. *arXiv preprint arXiv:2505.03912*.
- Deng, S.; Yan, M.; Wei, S.; Ma, H.; Yang, Y.; Chen, J.; Zhang, Z.; Yang, T.; Zhang, X.; Cui, H.; et al. 2025. Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data. *arXiv preprint arXiv:2505.03233*.
- 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.
- Du, Y.; Yang, S.; Dai, B.; Dai, H.; Nachum, O.; Tenenbaum, J.; Schuurmans, D.; and Abbeel, P. 2023. Learning universal policies via text-guided video generation. *Advances in neural information processing systems*, 36: 9156–9172.
- Fang, H.-S.; Fang, H.; Tang, Z.; Liu, J.; Wang, C.; Wang, J.; Zhu, H.; and Lu, C. 2024. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 653–660. IEEE.
- Guo, Y.; Hu, Y.; Zhang, J.; Wang, Y.-J.; Chen, X.; Lu, C.; and Chen, J. 2024. Prediction with Action: Visual Policy Learning via Joint Denoising Process. *arXiv:2411.18179*.
- He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. *arXiv:2006.11239*.
- Huang, H.; Chen, X.; Chen, Y.; Li, H.; Han, X.; Wang, Z.; Wang, T.; Pang, J.; and Zhao, Z. 2025. RoboGround: Robotic Manipulation with Grounded Vision-Language Priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 22540–22550.
- Karras, T.; Aittala, M.; Aila, T.; and Laine, S. 2022. Elucidating the Design Space of Diffusion-Based Generative Models. *arXiv:2206.00364*.
- Khanam, R.; and Hussain, M. 2024. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*.
- Kim, M. J.; Pertsch, K.; Karamcheti, S.; Xiao, T.; Balakrishna, A.; Nair, S.; Rafailov, R.; Foster, E. P.; Sanketi, P. R.; Vuong, Q.; et al. 2024. OpenVLA: An Open-Source Vision-Language-Action Model. In *8th Annual Conference on Robot Learning*.
- Kingma, D. P.; and Welling, M. 2022. Auto-Encoding Variational Bayes. *arXiv:1312.6114*.
- Lai, X.; Tian, Z.; Chen, Y.; et al. 2024. LISA: Reasoning Segmentation via Large Language Model. *arXiv:2308.00692*.
- Li, X.; Liu, M.; Zhang, H.; Yu, C.; Xu, J.; Wu, H.; Cheang, C.; Jing, Y.; Zhang, W.; Liu, H.; et al. 2024. Vision-Language Foundation Models as Effective Robot Imitators. In *The Twelfth International Conference on Learning Representations*.
- Li, Z.; Ren, L.; Yang, J.; Zhao, Y.; Wu, X.; Xu, Z.; Bai, X.; and Zhao, H. 2025. VIP: Vision Instructed Pre-training for Robotic Manipulation. In *Forty-second International Conference on Machine Learning*.
- Liu, B.; Zhu, Y.; Gao, C.; Feng, Y.; Liu, Q.; Zhu, Y.; and Stone, P. 2024a. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.
- Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024c. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, 38–55. Springer.
- Lu, H.; Liu, W.; Zhang, B.; Wang, B.; Dong, K.; Liu, B.; Sun, J.; Ren, T.; Li, Z.; Yang, H.; et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.
- Mees, O.; Hermann, L.; Rosete-Beas, E.; and Burgard, W. 2021. CALVIN: A Benchmark for Language-Conditioned Policy Learning for Long-Horizon Robot Manipulation Tasks. *IEEE Robotics and Automation Letters*.

- Niu, D.; Sharma, Y.; Biamby, G.; Quenum, J.; Bai, Y.; Shi, B.; Darrell, T.; and Herzig, R. 2024. LLARVA: Vision-Action Instruction Tuning Enhances Robot Learning. *arXiv preprint arXiv:2406.11815*.
- Octo Model Team; Ghosh, D.; Walke, H.; Pertsch, K.; Black, K.; Mees, O.; Dasari, S.; Hejna, J.; Xu, C.; Luo, J.; Kreiman, T.; Tan, Y.; Chen, L. Y.; Sanketi, P.; Vuong, Q.; Xiao, T.; Sadigh, D.; Finn, C.; and Levine, S. 2024. Octo: An Open-Source Generalist Robot Policy. In *Proceedings of Robotics: Science and Systems*. Delft, Netherlands.
- O’Neill, A.; Rehman, A.; Maddukuri, A.; Gupta, A.; Padalkar, A.; Lee, A.; Pooley, A.; Gupta, A.; Mandlikar, A.; Jain, A.; et al. 2024. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 6892–6903. IEEE.
- Peebles, W.; and Xie, S. 2022. Scalable Diffusion Models with Transformers. *arXiv preprint arXiv:2212.09748*.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv:2112.10752*.
- Song, W.; Chen, J.; Ding, P.; Huang, Y.; Zhao, H.; Wang, D.; and Li, H. 2025a. CEED-VLA: Consistency Vision-Language-Action Model with Early-Exit Decoding. *arXiv preprint arXiv:2506.13725*.
- Song, W.; Chen, J.; Ding, P.; Zhao, H.; Zhao, W.; Zhong, Z.; Ge, Z.; Ma, J.; and Li, H. 2025b. Accelerating Vision-Language-Action Model Integrated with Action Chunking via Parallel Decoding. *arXiv preprint arXiv:2503.02310*.
- Song, W.; Chen, J.; Li, W.; He, X.; Zhao, H.; Cui, C.; Su, P. D. S.; Tang, F.; Cheng, X.; Wang, D.; et al. 2025c. Rationalvla: A rational vision-language-action model with dual system. *arXiv preprint arXiv:2506.10826*.
- Song, W.; Zhao, H.; Ding, P.; Cui, C.; Lyu, S.; Fan, Y.; and Wang, D. 2024. Germ: A generalist robotic model with mixture-of-experts for quadruped robot. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 11879–11886. IEEE.
- Song, Y.; and Ermon, S. 2020. Generative Modeling by Estimating Gradients of the Data Distribution. *arXiv:1907.05600*.
- Stewart, E. E. M. e. a. 2020. A review of interactions between peripheral and foveal vision. *Journal of Vision*, 20(12): 2–2.
- Tian, Y.; Yang, S.; Zeng, J.; Wang, P.; Lin, D.; Dong, H.; and Pang, J. 2024. Predictive inverse dynamics models are scalable learners for robotic manipulation. *arXiv preprint arXiv:2412.15109*.
- Tong, X.; Ding, P.; Fan, Y.; Wang, D.; Zhang, W.; Cui, C.; Sun, M.; Zhao, H.; Zhang, H.; Dang, Y.; Huang, S.; and Lyu, S. 2025. QUART-Online: Latency-Free Large Multimodal Language Model for Quadruped Robot Learning. *arXiv:2412.15576*.
- Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.
- Walke, H.; Black, K.; Lee, A.; Kim, M. J.; Du, M.; Zheng, C.; Zhao, T.; Hansen-Estruch, P.; Vuong, Q.; He, A.; Myers, V.; Fang, K.; Finn, C.; and Levine, S. 2024. BridgeData V2: A Dataset for Robot Learning at Scale. *arXiv:2308.12952*.
- Walke, H. R.; Black, K.; Zhao, T. Z.; Vuong, Q.; Zheng, C.; Hansen-Estruch, P.; He, A. W.; Myers, V.; Kim, M. J.; Du, M.; et al. 2023. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, 1723–1736. PMLR.
- Wang, H.; Zheng, A.; Zhao, Y.; Wang, T.; Ge, Z.; Zhang, X.; and Zhang, Z. 2024. Reconstructive Visual Instruction Tuning. *arXiv:2410.09575*.
- Wang, Y.; Li, X.; Wang, W.; Zhang, J.; Li, Y.; Chen, Y.; Wang, X.; and Zhang, Z. 2025. Unified Vision-Language-Action Model. *arXiv preprint arXiv:2506.19850*.
- Wen, Y.; Lin, J.; Zhu, Y.; Han, J.; Xu, H.; Zhao, S.; and Liang, X. 2024. Vidman: Exploiting implicit dynamics from video diffusion model for effective robot manipulation. *Advances in Neural Information Processing Systems*, 37: 41051–41075.
- Wu, H.; Jing, Y.; Cheang, C.; Chen, G.; Xu, J.; Li, X.; Liu, M.; Li, H.; and Kong, T. 2024. Unleashing Large-Scale Video Generative Pre-training for Visual Robot Manipulation. *ICLR*.
- Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; Dong, G.; Wei, H.; Lin, H.; Tang, J.; Wang, J.; Yang, J.; Tu, J.; Zhang, J.; Ma, J.; Yang, J.; Xu, J.; Zhou, J.; Bai, J.; He, J.; Lin, J.; Dang, K.; Lu, K.; Chen, K.; Yang, K.; Li, M.; Xue, M.; Ni, N.; Zhang, P.; Wang, P.; Peng, R.; Men, R.; Gao, R.; Lin, R.; Wang, S.; Bai, S.; Tan, S.; Zhu, T.; Li, T.; Liu, T.; Ge, W.; Deng, X.; Zhou, X.; Ren, X.; Zhang, X.; Wei, X.; Ren, X.; Liu, X.; Fan, Y.; Yao, Y.; Zhang, Y.; Wan, Y.; Chu, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; Guo, Z.; and Fan, Z. 2024a. Qwen2 Technical Report. *arXiv:2407.10671*.
- Yang, R.; Wang, Z.; Jiang, B.; and Li, S. 2024b. The Convergence of Variance Exploding Diffusion Models under the Manifold Hypothesis.
- Zawalski, M.; Chen, W.; Pertsch, K.; Mees, O.; Finn, C.; and Levine, S. 2024. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*.
- Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid Loss for Language Image Pre-Training. *arXiv:2303.15343*.
- Zhang, H.; Ding, P.; Lyu, S.; Peng, Y.; and Wang, D. 2025. GEVRM: Goal-Expressive Video Generation Model For Robust Visual Manipulation. In *The Thirteenth International Conference on Learning Representations*.
- Zhao, H.; Song, W.; Wang, D.; Tong, X.; Ding, P.; Cheng, X.; and Ge, Z. 2025a. MoRE: Unlocking Scalability in Reinforcement Learning for Quadruped Vision-Language-Action Models. *arXiv preprint arXiv:2503.08007*.
- Zhao, W.; Ding, P.; Zhang, M.; Gong, Z.; Bai, S.; Zhao, H.; and Wang, D. 2025b. VLAS: Vision-Language-Action Model With Speech Instructions For Customized Robot Manipulation. *International Conference on Learning Representations (ICLR)*.

Zhen, H.; Qiu, X.; Chen, P.; Yang, J.; Yan, X.; Du, Y.; Hong, Y.; and Gan, C. 2024. 3D-VLA: A 3D Vision-Language-Action Generative World Model. In *ICML*.

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.