

LLaDA-VLA: Vision Language Diffusion Action Models

Yuqing Wen^{1*} Hebei Li^{1*} Kefan Gu^{2*} Yucheng Zhao^{3†}
Tiancai Wang³ Xiaoyan Sun^{1‡}

¹University of Science and Technology of China,

²Nanjing University, ³Dexmal

Project Page: <https://wenyuqing.github.io/llada-vla/>

Abstract

The rapid progress of auto-regressive vision-language models (VLMs) has inspired growing interest in vision-language-action models (VLA) for robotic manipulation. Recently, masked diffusion models, a paradigm distinct from autoregressive models, have begun to demonstrate competitive performance in text generation and multimodal applications, leading to the development of a series of diffusion-based VLMs (d-VLMs). However, leveraging such models for robot policy learning remains largely unexplored. In this work, we present LLaDA-VLA, the first Vision-Language-Diffusion-Action model built upon pretrained d-VLMs for robotic manipulation. To effectively adapt d-VLMs to robotic domain, we introduce two key designs: (1) a localized special-token classification strategy that replaces full-vocabulary classification with special action token classification, reducing adaptation difficulty; (2) a hierarchical action-structured decoding strategy that decodes action sequences hierarchically considering the dependencies within and across actions. Extensive experiments demonstrate that LLaDA-VLA significantly outperforms state-of-the-art VLAs on both simulation and real-world robots.

1. Introduction

Auto-regressive models (ARMs) have long dominated the development of vision-language models (VLMs) [1, 2, 21, 24, 31, 32], demonstrating strong performance in multimodal understanding and text generation. Their success has naturally inspired vision-language-action models (VLAs) [3, 4, 6, 23, 25, 50, 58], where pretrained VLMs are fine-tuned on robot-specific datasets to generate actions. While effective, ARM-based VLMs rely on sequential to-

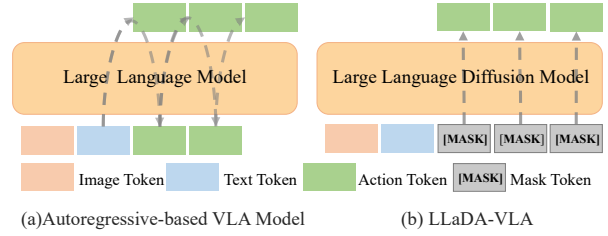


Figure 1. Comparison between Autoregressive-based VLA Model and LLaDA-VLA.

ken generation, which constrains efficiency and limits flexibility due to their inherently unidirectional generation, especially in complex multimodal robotic tasks.

Recently, masked diffusion models (MDMs) [35, 45, 46] have emerged as a competitive alternative to auto-regressive approaches. Instead of generating tokens sequentially, MDMs produce them in parallel and iteratively refine predictions through a discrete diffusion process. A seminal work in language generation, LLaDA [39], demonstrates the remarkable effectiveness and scalability of this paradigm with large-scale pretraining, while subsequent studies, such as LLaDA-V [55] and MMaDA [52], further extend this paradigm to the vision-language domain, forming diffusion-based VLMs (d-VLMs) that achieve performance comparable to leading ARM-based models. Despite this progress, their potential for robotic action generation remains unexplored, motivating our development of LLaDA-VLA, the first Vision-Language-Diffusion-Action model built upon pretrained d-VLMs.

Adapting d-VLMs to robotic tasks poses unique challenges in two aspects. First, a substantial domain gap exists: d-VLMs are trained on large-scale general datasets rich in high-level semantics, whereas VLAs must interpret low-level visual cues to generate precise actions. Second, the masked diffusion paradigm is not naturally suited for generating structured action sequences, as its decoding strategy fails to account for the structural characteristics of action

*This work was done during the internship at Dexmal.

†Project lead.

‡Corresponding author.

sequences during decoding, making it difficult to model the strong hierarchical dependencies of robotic actions and to generate reasonable action trajectories.

To address these challenges, we propose two key strategies. First, a **localized special-token classification strategy** that mitigates the domain gap by restricting the classification space to special action tokens, significantly easing the difficulty of adaptation to robotic environments. Second, a **hierarchical action-structured decoding strategy** explicitly considers the dependencies within and across actions. Concretely, we first estimate action-level confidence scores and rank actions to determine their decoding order. Within each action, we further rank individual tokens by token-level confidence to guide decoding within the action. Together, these designs enable the masked diffusion paradigm to generate coherent and precise action sequences in robotic tasks.

LLaDA-VLA demonstrates state-of-the-art performance on both simulation and real-robot benchmarks. Compared to ARM-based baselines such as OpenVLA [23], it achieves a 0.74 average length improvement on CALVIN and 51.3% average success rate gain on SimplerEnv benchmark. LLaDA-VLA also surpasses methods such as $\pi 0$ [4] and CogACT [25], achieving improvements of 23% and 28% in average success rate on real robots. In summary, our main contributions are:

- We propose the first Vision-Language-Diffusion-Action model (LLaDA-VLA) built on pretrained d-VLMs, establishing a new paradigm for robotic policy learning.
- We design two techniques to make the masked diffusion model well-suited for action generation: a localized special-token classification strategy for easier domain adaptation, and a hierarchical action-structured decoding strategy for seamless integration into action generation.
- Extensive experiments on simulation benchmarks SimplerEnv and CALVIN and the WidowX real robot demonstrate LLaDA-VLA’s superior performance, highlighting the potential of d-VLMs for robotic manipulation.

2. Related Work

2.1. Large Language Diffusion Models

Diffusion models [12, 17, 42, 44, 48] have achieved remarkable progress in the vision domain in recent years, sparking growing interest in extending their capabilities to text generation. However, due to the inherently discrete nature of textual tokens, directly transferring diffusion models that are originally designed to operate in continuous pixel

representation spaces remains challenging. To address this issue, one line of work [8, 14, 36, 54] proposes learning continuous representations for text, while another focuses on developing discrete diffusion models [13, 33, 41, 53, 59]. Among them, masked diffusion models, a specific variant of discrete diffusion, have demonstrated impressive potential. Representative works such as Dream7B [53] and LLaDA [39] leverage large-scale language pretraining and exhibit text generation performance and scaling properties comparable to those of auto-regressive large language models.

Beyond the text domain, large language diffusion models have also demonstrated promising performance in various multimodal settings [26, 52, 55, 57]. In the vision-language domain, LaViDA [26] employs a discrete diffusion transformer with multi-view image encoding and masked-denoising training. LLaDA-V [55] integrates masked diffusion with visual instruction tuning, enabling parallel decoding and controllable infilling. MMaDA [52] further proposes a unified diffusion transformer that aligns reasoning across text and vision through chain-of-thought supervision and reinforcement learning. Beyond the vision-language domain, DIFFA [57] extends diffusion-based LLMs to audio understanding tasks by aligning spoken content with textual representations, while LLaDA-MedV [9] adapts masked diffusion for medical applications, integrating imaging and textual data for diagnostic reasoning. Despite these advancements, applying large language diffusion models to robotic manipulation remains largely unexplored, leaving significant avenues for future research.

2.2. Vision-Language-Action Models

Building a generalist robotic policy [7, 10, 15, 16, 18–20, 29, 37, 43, 47] has long been a challenging and highly desired goal. In recent years, the rapid advances in auto-regressive Vision-Language Models (VLMs) [1, 21, 24, 31, 32] have inspired researchers to leverage their strong multimodal understanding and generalization capabilities to develop robotic policies, commonly referred to as Vision-Language-Action (VLA) models [3, 4, 6, 11, 23, 25, 28, 56, 58]. Among them, RT-2 [6] is a pioneering work that jointly fine-tunes pretrained VLMs on a combination of web-scale VQA data and robot demonstrations, achieving strong multi-task performance across diverse manipulation tasks. Building upon this idea, OpenVLA [23] is introduced as the first open-source VLA model, further promoting research progress in this area. Subsequently, LLARVA [40] improves action prediction by incorporating trajectory annotations; CogACT [25] introduces a large diffusion action head to predict continuous control commands; and $\pi 0$ [4] adopts a flow-matching strategy combined with carefully curated large-scale multi-task datasets, achieving remarkable performance.

Despite these advances, existing VLA models are al-

most exclusively built on auto-regressive VLMs, leaving the potential of diffusion-based VLMs largely unexplored. In this work, we investigate how to build a Vision-Language-Diffusion-Action model based on pretrained d-VLMs. A concurrent work [30] shares partial inspiration with us, but their approach still relies on auto-regressive VLMs.

3. Method

We present the methodology of LLaDA-VLA, starting with a brief overview of the masked diffusion model in Section 3.1, followed by the architecture and key designs of LLaDA-VLA in Section 3.2.

3.1. Preliminary: Mask Diffusion Models

Masked diffusion models (MDMs) [13, 39, 41, 53] define a generative paradigm based on a forward–reverse diffusion process over discrete tokens, fundamentally different from auto-regressive models. Within this paradigm, large language diffusion models such as LLaDA have demonstrated competitive performance comparable to auto-regressive counterparts. In the forward process, given an input sequence $\mathbf{x}_0 = [\mathbf{x}_0^i]_{i=1}^N$, $\mathbf{x}_0^i \in \{0, 1, \dots, \mathcal{V} - 1\}^N$ of length N with vocabulary size \mathcal{V} , each token in \mathbf{x}_0 is independently replaced by a special mask token $[M]$ with probability t , formulated as:

$$q_{t|0}(\mathbf{x}_t | \mathbf{x}_0) = \prod_{i=0}^{N-1} q_{t|0}(\mathbf{x}_t^i | \mathbf{x}_0^i), \quad (1)$$

$$q_{t|0}(\mathbf{x}_t^i | \mathbf{x}_0^i) = \begin{cases} 1 - t, & \mathbf{x}_t^i = \mathbf{x}_0^i \\ t, & \mathbf{x}_t^i = [M]. \end{cases} \quad (2)$$

In a reverse process, MDMs gradually transform masked tokens into meaningful content, starting from a fully masked sequence. Given $0 < s < t < 1$, each sampling step in the reverse process can be formulated as:

$$q_{s|t}(\mathbf{x}_s | \mathbf{x}_t) = \prod_{i=0}^{N-1} q_{s|t}(\mathbf{x}_s^i | \mathbf{x}_t^i), \quad (3)$$

$$q_{s|t}(\mathbf{x}_s^i | \mathbf{x}_t^i) = \begin{cases} 1, & \mathbf{x}_t^i \neq [M], \mathbf{x}_s^i = \mathbf{x}_t^i \\ \frac{s}{t}, & \mathbf{x}_t^i = [M], \mathbf{x}_s^i = [M], \\ \frac{t-s}{t} p_\theta(\mathbf{x}_0^i | \mathbf{x}_t^i), & \mathbf{x}_t^i = [M], \mathbf{x}_s^i \neq [M], \\ 0, & \text{otherwise,} \end{cases} \quad (4)$$

where p_θ is modeled by a Transformer as mask predictor. This means that for each reverse sampling step, a fraction of $(1 - s/t)$ tokens are predicts through mask predictor $p_\theta(\mathbf{x}_0^i | \mathbf{x}_t^i)$, while a fraction s/t of tokens are remasked and deferred for re-predict in subsequent sampling steps. A typical strategy, as adopted in LLaDA, is to choose the s/t

tokens for remask with the lowest confidence scores (i.e., the smallest logits).

The mask predictor p_θ is trained to predict mask tokens with cross-entropy loss only computed on masked tokens:

$$\mathcal{L}(\theta) \triangleq -\mathbb{E}_{t, \mathbf{x}_0, \mathbf{x}_t} \left[\frac{1}{t} \sum_{i=1}^L \mathbf{1}[\mathbf{x}_t^i = M] \log p_\theta(\mathbf{x}_0^i | \mathbf{x}_t^i) \right] \quad (5)$$

3.2. LLaDA-VLA

In this section, we provide a detailed description of LLaDA-VLA. We first introduce the overall model architecture, including the structure of our VLA framework and the processing of its inputs and outputs. We then present two key designs proposed in this work: localized special-token classification and hierarchical action-structured decoding strategy.

3.2.1 Model Architecture

Vision-language Modules: As shown in Figure. 2 (a), LLaDA-VLA consists of three main components: a language backbone, a vision encoder, and a projector. Following LLaDA-V [55], we use LLaDA [39] as the language backbone, and adopt SigLIP-2 [49] as the vision encoder and MLP as projector, respectively. The model takes two inputs: a language instruction that specifies the robot’s task and a front-view RGB image. The vision encoder processes the image to extract visual features, which are then projected by the MLP into the shared space with the text tokens. The visual and text tokens are then concatenated and fed into the large language diffusion model to generate the robot action sequence in a manner that is detailed in Section 3.2.2 and Section 3.2.3.

Action Tokenization and Chunking: To enable the language model to generate robot actions, we discretize continuous action values into bins with bin size of \mathcal{V}_a . We augment the original vocabulary with \mathcal{V}_a additional special tokens $\mathcal{S} = \{s_0, s_1, \dots, s_{\mathcal{V}_a-1}\}$, where $\mathcal{V}_a \ll \mathcal{V}$, to represent these discrete action tokens, resulting in a vocabulary size of $\mathcal{V}_{total} = \mathcal{V} + \mathcal{V}_a$. Therefore, a per-timestep action is represented by $D = 7$ special action tokens: three for positional displacements, three for rotational changes, and one for the gripper open/close state. To generate multi-step trajectories, the model predicts an action chunk spanning K consecutive timesteps, yielding a robot action sequence of $K \times D$ special action tokens. These tokens can be de-tokenized to recover the original continuous values, allowing the model to produce executable robot trajectories.

3.2.2 Localized Special-token Classification

The original training objective of pretrained d-VLMs is to perform full-vocabulary classification for each token. To

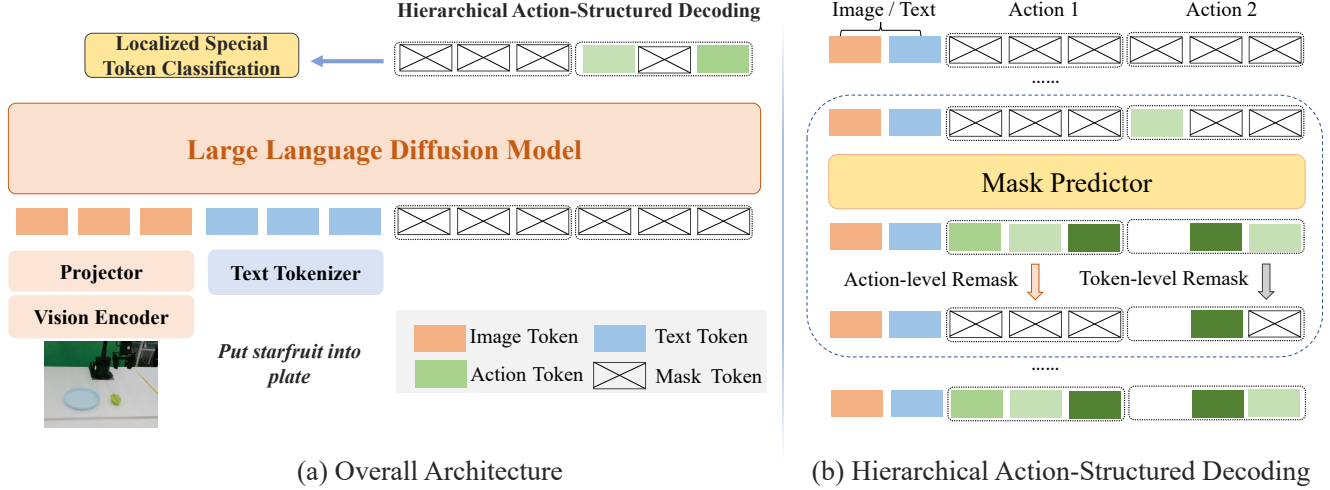


Figure 2. **Overview of LLaDA-VLA.** (a) Overall architecture. Visual features extracted by the vision encoder are projected into the text space and concatenated with text tokens. Together with masked tokens, they are fed into a large language diffusion model to generate action sequences via Localized Special-Token Classification and further refined with Hierarchical Action-Structured Decoding. (b) Hierarchical Action-Structured Decoding strategy. Starting from a fully masked action sequence (except vision and text prompts), the model iteratively predicts masked tokens, performing action-level and token-level remasking based on confidence until the full sequence is decoded.

generate actions, however, the model is required to predict only the special action tokens. Therefore, retaining the full-vocabulary classification objective would complicate the learning process and hinder the adaptation of pre-trained d-VLMs to robotic action generation. To address this, we introduce a localized special-token classification mechanism. Instead of performing classification over the entire vocabulary with size \mathcal{V}_{total} , we focus the classification only on the special action tokens \mathcal{S} . During training, each original token label $y_i \in \mathcal{V}$ is mapped to a local class $l_i \in \{0, \dots, \mathcal{V}_a - 1\}$ as:

$$l_i = \begin{cases} \text{map}(y_i), & \text{if } y_i \in \mathcal{S}, \\ -100, & \text{otherwise (ignored in loss),} \end{cases} \quad (6)$$

where $\text{map}(\cdot)$ denotes the mapping from the original token index in the full vocabulary to the local class index. We then choose only the logits on the special action tokens $z_i = \text{logits}[i, \mathcal{S}] \in \mathbb{R}^{\mathcal{V}_a}$ and compute the token-level cross-entropy loss only on masked positions with the training objective:

$$L_{\text{token}} = \frac{1}{|M|} \sum_{i \in M} \text{CE}(z_i, l_i), \quad (7)$$

where $\text{CE}(\cdot)$ denotes the standard cross-entropy loss and M is the set of valid masked positions. During inference, we predict over the target token subset \mathcal{S} only, and map the predicted local class indexes back to the original token indexes to get the action tokens. This localized classification objective concentrates learning on action-relevant tokens, thereby improving the accuracy for action generation and making the training easier.

3.2.3 Hierarchical Action-Structured Decoding

In the original LLaDA decoding process, as illustrated in Figure 2 (b), starting from a fully masked sequence, the model first predicts the masked tokens and retains those with high confidence, while remasks the remaining tokens. In the subsequent diffusion step, the remasked tokens are re-predicted, and again a fraction of tokens with high confidence are retained and others remasked. By iteratively repeating this predict–remask–predict process, the model gradually refines its generated output until obtaining the desired result.

However, this conventional decoding strategy cannot be directly applied to generating action chunks. This is primarily because it treats all output tokens equally, ignoring the structured dependencies between them. In contrast, an action chunk exhibits both intra-action and inter-action correlations. To explicitly capture such correlations, we introduce a hierarchical action-structured decoding strategy that accounts these dependencies during generation. We first compute *action-level confidence scores* for each action by summing the confidence of tokens within each action:

$$C_a^{(i)} = \sum_{j=1}^D c_{i,j}, \quad (8)$$

where $c_{i,j}$ denotes the confidence of the j -th token within the i -th action, and D is the number of tokens per action. At each hierarchical decoding step, we first rank all actions within the predicted action chunk according to their action-level confidence scores $C_a^{(i)}$. The action with the

Table 1. Performance comparison on the WidowX robot in the SimplierEnv *Visual Matching* setting. We compare success rates (%) on 4 tasks. LLaDA-VLA achieves the best average success rate.

Method	Put Spoon on Towel	Put Carrot on Plate	Stack Green on Yellow	Put Eggplant in Basket	Average
RT-1-X [50]	0.0	4.2	0.0	0.0	1.1
Octo-Base [37]	15.8	12.5	0.0	41.7	17.5
Octo-Small [37]	41.7	8.2	0.0	56.7	26.7
OpenVLA [23]	4.2	0.0	0.0	12.5	4.2
Cog-ACT [25]	71.7	50.8	15.0	67.5	51.3
DiscreteDiffusionVLA [30]	37.5	-	20.8	29.2	29.2
LLaDA-VLA	56.9	76.3	30.6	58.3	55.5

Table 2. Comparison of methods on CALVIN ABC-D setting. We report the average success rate over 1000 rollouts per task, along with the average number of tasks completed consecutively to accomplish five instructions (Avg. Len.). LLaDA-VLA consistently outperforms previous methods.

Method	Task completed in a row					Avg. Len. ↑
	1	2	3	4	5	
Roboflamingo [29]	82.4	61.9	46.6	33.1	23.5	2.47
Susie [5]	87.0	69.0	49.0	38.0	26.0	2.69
GR-1 [51]	85.4	71.2	59.6	49.7	40.1	3.06
3D Diffusor Actor [22]	92.2	78.7	63.9	51.2	41.2	3.27
OpenVLA [23]	91.3	77.8	62.0	52.1	43.5	3.27
LLaDA-VLA	95.6	87.8	79.5	73.9	64.5	4.01

highest confidence C_a is selected for partial preservation and the rest of the actions are remasked (action-level remask). Within this selected action, its tokens are further ranked by a token-level confidences. Only a subset of high-confidence tokens is retained with the rest remasked (token-level remask). The remasked tokens are then regenerated in subsequent diffusion steps. This hierarchical decoding procedure ensures that the trajectory is generated in an action-wise manner, preserving the structural integrity while allowing further refinement within each individual action. In this way, the model can effectively generate more coherent and reasonable action trajectories.

4. Experiment

4.1. Experiment Setup

4.1.1 Dataset

We evaluate LLaDA-VLA in the simulation environments SimplierEnv [27] and CALVIN [38], as well as on a real-world WidowX robot.

SimplierEnv. SimplierEnv [27] is designed to closely mirror real-world physical dynamics and visual appearances, providing a more faithful evaluation of robotic policies. In our experiments, we evaluate on the WidowX robot under the SimplierEnv Visual Matching setting, which minimizes the gap between simulation and reality, thereby closely approximating real-world conditions. We focus on four tasks: *Put*

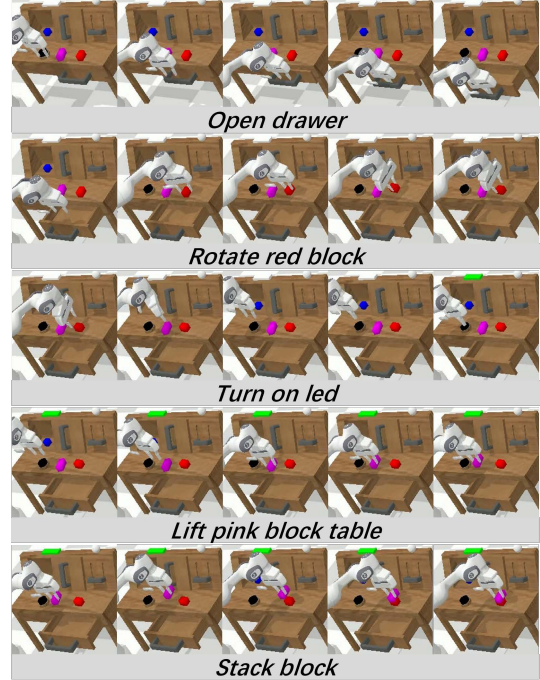


Figure 3. Qualitative results of LLaDA-VLA on CALVIN tasks.

Spoon on Towel, Put Carrot on Plate, Stack Green on Yellow, and Put Eggplant in Basket. Task success rate is used as the primary evaluation metric.

CALVIN. We train and evaluate LLaDA-VLA on CALVIN, an open-source benchmark for long-horizon, language-conditioned robotic manipulation. CALVIN comprises four distinct environments (A, B, C, D) and provides diverse sensory inputs commonly used in visuomotor control, including RGB-D observations from fixed and gripper-mounted cameras, proprioceptive states, and vision-based tactile feedback. Following the ABC-D protocol, we train on environments A, B, and C, and evaluate generalization on environment D. We report both success rate and average episode length over five consecutive tasks.

Real-World WidowX Robot. For real-world experiments, we use a WidowX 250S robotic arm equipped with an Intel RealSense D435 camera positioned in front of the setup to capture third-person visual observations. LLaDA-VLA is evaluated on eight tasks, comprising four seen tasks and four generalization tasks. The seen tasks include three short-horizon manipulations—*Banana on Plate*, *Strawberry in Bowl*, and *Starfruit on Plate*—and one long-horizon task, *Banana and Strawberry in Bowl*. The generalization tasks involve *Cube on Plate* (introducing an unseen object), *Strawberry in Box* (featuring an unseen container), *Cube in Box* (involving both an unseen object and container), and a long-horizon OOD task, *Banana and Starfruit in Bowl* (performed in the presence of distractor ob-

Table 3. Comparison with previous methods on real robot. We compare the success rates (%) across four tasks. LLaDA-VLA achieves the best performance.

Method	Banana on Plate	Strawberry in Bowl	Starfruit on Plate	Banana&Strawberry in Bowl	Average Success Rate
π_0 [4]	50%	30%	40%	20%	35%
CogACT [25]	40%	30%	30%	20%	30%
LLaDA-VLA	50%	70%	70%	40%	58%

Table 4. Performance on unseen tasks for real robot evaluating the generalization capability. Comparing the success rate across these OOD tasks, LLaDA-VLA shows better performance than π_0 .

Method	Cube on Plate	Strawberry in Box	Cube in Box	Banana&Strawberry in Bowl(dist)	Average Success Rate
π_0 [4]	30	20	10	0	15
LLaDA-VLA	50	60	50	0	40

Table 5. Ablation on the localized special-token classification(LSC) and the hierarchical action-structured decoding(HAD) on CALVIN ABC-D setting.

Method	Task completed in a row					Avg. Len. \uparrow
	1	2	3	4	5	
baseline	86.2	62.4	46.1	34.1	24.7	2.64
+ LSC	91.6	80.1	67.2	57.3	46.4	3.43(+0.79)
+ HAD	95.6	87.8	79.5	73.9	64.5	4.01(+0.58)

Table 6. Ablation on the action chunk size on CALVIN ABC-D setting.

Method	Task completed in a row					Avg. Len. \uparrow
	1	2	3	4	5	
3	95.0	86.8	78.2	70.4	59.9	3.90
5	95.6	87.8	79.5	73.9	64.5	4.01
8	91.1	79.5	69.0	61.3	52.4	3.53
10	90.8	76.5	65.8	55.7	46.0	3.36

jects). Each task is executed over 10 independent trials.

4.1.2 Training and Evaluation Details

We use LLaDA-V [55] as our pretrained weights, an open-source d-VLM. All our experiments are fine-tuned for 3 epochs, with a learning rate of $2e-5$ and a batch size of 128. We introduce 32 additional special tokens into the vocabulary for classification. Since we adopt a fixed-length output setting, we remove the EOS token. In our main experiments, the action chunk size is set to 5, and the model predicts delta actions. During inference, we use 10 iterative diffusion steps, with 2 iterations per action. We adopt the dllm-cache [34] method to accelerate the decoding.

4.2. Quantitative Results

4.2.1 Performance Comparison

SimplerEnv. We train and evaluate LLaDA-VLA on SimplerEnv, as shown in Table 1. LLaDA-VLA achieves an

average performance of 55.1, which is significantly higher than OpenVLA, a typical autoregressive VLA model, with 50.9% improvements and also surpasses more advanced methods such as CogAct with 4.2 points gain. Such superior performance validates the effectiveness of the LLaDA-VLA framework.

CALVIN. We further compare LLaDA-VLA with several representative works on another widely adopted simulation benchmark, CALVIN, as summarized in Table 2. In particular, LLaDA-VLA outperforms OpenVLA with a notable performance improvement of 0.74. It also demonstrates superior performance compared to other prominent approaches, such as GR1 [51] and RoboFlamingo [29]. These results provide compelling evidence of LLaDA-VLA’s effectiveness, highlighting the promising value of incorporating the masked diffusion paradigm to the VLA framework.

Real Robot. Beyond simulation, we further evaluate LLaDA-VLA on real-robot experiments, comparing it with state-of-the-art approaches π_0 and CogACT. As shown in Table 3, LLaDA-VLA achieves an average success rate of 58% across four real-world manipulation tasks, including up to 70% success rates on the *Strawberry in Bowl* and *Starfruit on Plate* tasks. These results consistently outperform both CogACT and π_0 , demonstrating the strong efficacy of LLaDA-VLA on real world setting.

4.2.2 Generalization Capability

We evaluate the generalization capability of LLaDA-VLA on four out-of-distribution (OOD) real-robot tasks, which involve unseen objects, containers, and distractors. As shown in Table 4, LLaDA-VLA demonstrates strong generalization performance, achieving, for example, a 60% success rate on *Strawberry in Box* task, which is much higher than the success rate of π_0 . LLaDA-VLA attains a gain of 25% average success over π_0 , further highlighting its outstanding generalization ability.

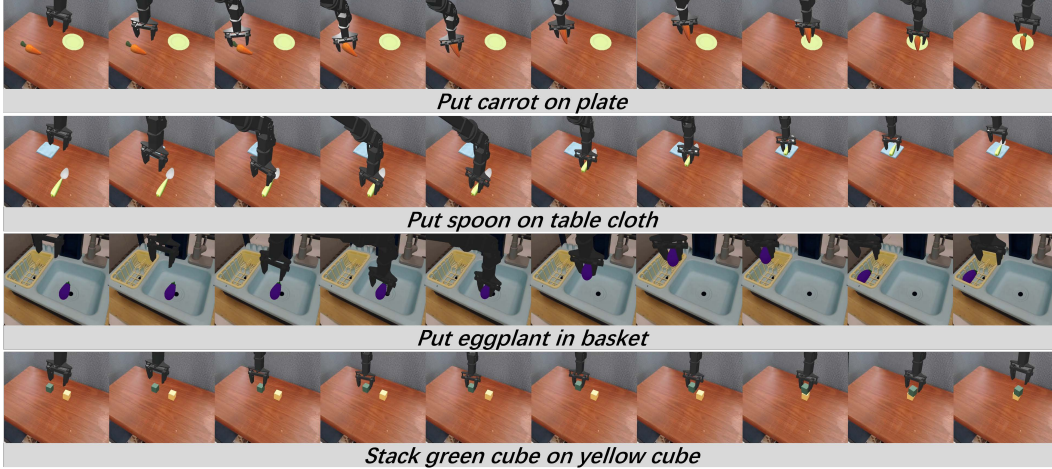


Figure 4. Qualitative results of LLaDA-VLA on SimplerEnv tasks.

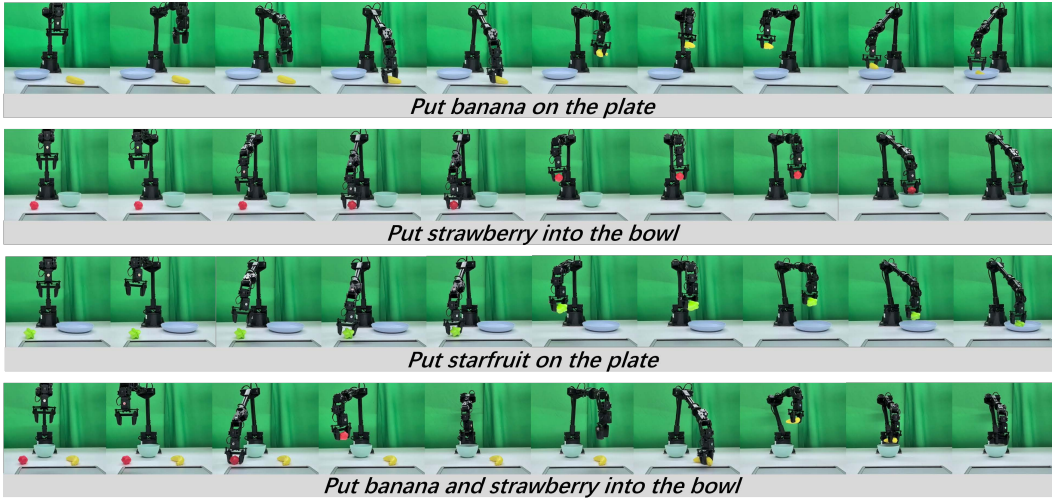


Figure 5. Qualitative results of LLaDA-VLA on real-world in-domain tasks.

4.2.3 Ablation Study

To validate the effectiveness of our proposed designs, we conduct ablation studies on the two core components of LLaDA-VLA: the localized special-token classification mechanism and the hierarchical action-structured decoding strategy. Additionally, we investigate how the length of the generated action sequences affects performance under the diffusion paradigm. All experiments in this section are carried out on the CALVIN benchmark.

Localized Special-token Classification. As shown in Table 5, we first establish a baseline model without the localized special-token classification or the hierarchical action-structured decoding. This baseline yields suboptimal results, with an average episode length of only 2.54 and a success rate of 86.2% on the first task, highlighting the significant challenges of directly adapting vanilla d-VLMs to

robotic manipulation. When we incorporate the localized special-token classification strategy into the baseline, the performance improves substantially by 0.79. This demonstrates that focusing the model on action-specific tokens effectively reduces the adaptation difficulty, transforming the full-vocabulary classification problem into a more tractable task and enabling more efficient training.

Hierarchical Action-structured Decoding strategy. We investigate the effectiveness of the hierarchical action-structured decoding strategy. As shown in Table 5, incorporating this decoding approach brings a substantial improvement of 0.58 points over the vanilla decoding strategy. This significant gain demonstrates that explicitly modeling both intra-action and inter-action dependencies through hierarchical decoding effectively benefit the diffusion model’s outputs, enabling the generation of more coherent action

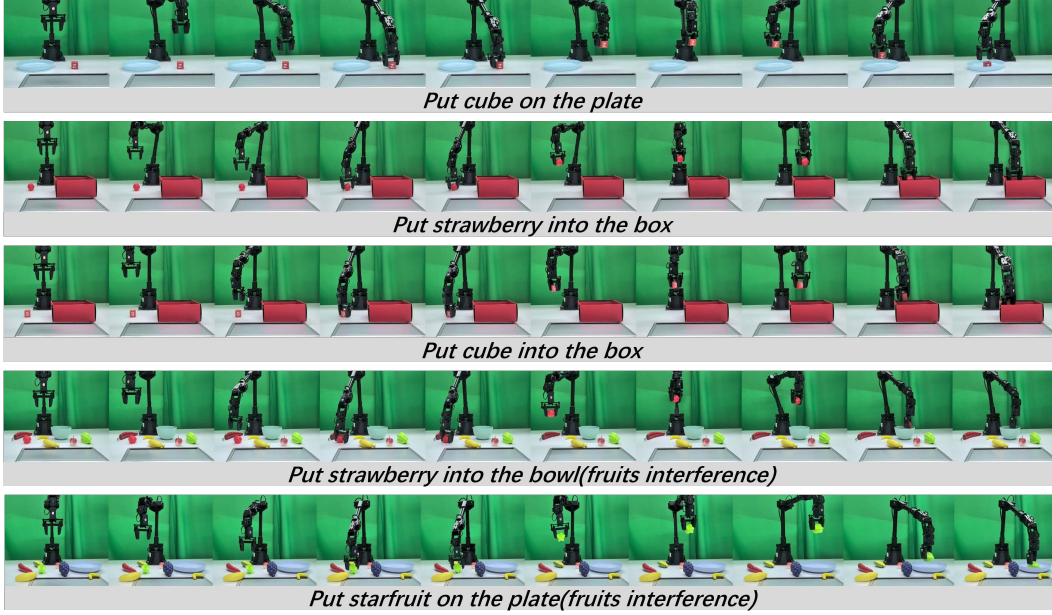


Figure 6. Qualitative results of LLaDA-VLA on real-world out-of-domain tasks.

trajectories and markedly improving task success rates.

Action Chunk Size. The action chunk size has a significant impact on the model’s performance for generating trajectories of varying lengths. As shown in Table 6, moderately increasing the chunk size can improve task success rates, as generating longer segments at once encourages smoother motion. However, when the chunk size becomes too large, the mask prediction task becomes more difficult due to the increased number of tokens to be generated, which may reduce overall accuracy. Therefore, for LLaDA-VLA, it is crucial to select an appropriate action chunk size to balance trajectory smoothness and prediction accuracy, to achieve optimal performance.

4.3. Qualitative Results

In this section, we present visualizations of several tasks in both simulation and real-world, demonstrating LLaDA-VLA’s ability to accomplish a wide range of robotic tasks.

Simulation results. In CALVIN, LLaDA-VLA successfully completes multi-step manipulation tasks as shown in Figure 3. This demonstrates LLaDA-VLA’s capability to complete long-horizon tasks. In SimplerEnv, as illustrated in Figure 4, LLaDA-VLA accurately localizes target objects and reliably grasps and places them into the correct locations, showcasing its good performance. Overall, these results demonstrate that LLaDA-VLA is capable of stably executing long-horizon tasks while also achieving precise object-level manipulation.

Real-world results. On in-domain tasks, such as *Banana on Plate* or *Strawberry in Bowl*, the robot reliably executes actions, as shown in Figure 5. LLaDA-VLA also exhibits

strong generalization to out-of-domain tasks, as illustrated in Figure 6. It is capable of grasping previously unseen objects, such as cubes, and placing items into novel targets, such as a paper box. Even in the presence of multiple distractor objects, LLaDA-VLA successfully completes tasks, for example, accurately picking and placing a strawberry despite surrounding fruits. These qualitative results demonstrate the versatility and robustness of LLaDA-VLA in real-world scenarios.

5. Conclusion

We present LLaDA-VLA, the first Vision-Language-Diffusion-Action model built upon pretrained diffusion-based Vision-Language Models (d-VLMs). In this paper, we explore how to effectively leverage d-VLMs for robotic manipulation. To address the domain gap between d-VLMs and the VLA, we propose a localized special-token classification strategy to reduce the adaptation difficulty and enable more efficient model training. Furthermore, to allow the masked diffusion paradigm to generate structured action trajectories, we introduce a hierarchical action-structured decoding strategy, which enables the model to produce action sequences in a hierarchical manner, resulting in more coherent and plausible outcomes. With these designs, LLaDA-VLA achieves state-of-the-art performance, demonstrating remarkable effectiveness across multiple simulation benchmarks as well as real-robot experiments. These results provide a solid foundation for exploring the application of d-VLMs in robotic manipulation and paving the way for future research in this direction.

References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1, 2
- [2] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 1
- [3] Suneel Belkhale, Tianli Ding, Ted Xiao, Pierre Sermanet, Quon Vuong, Jonathan Tompson, Yevgen Chebotar, Debiddatta Dwibedi, and Dorsa Sadigh. Rt-h: Action hierarchies using language. *arXiv preprint arXiv:2403.01823*, 2024. 1, 2
- [4] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. *pi*-0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. 1, 2, 6
- [5] Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pretrained image-editing diffusion models. *arXiv preprint arXiv:2310.10639*, 2023. 5
- [6] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. 1, 2
- [7] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022. 2
- [8] Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. *arXiv preprint arXiv:2208.04202*, 2022. 2
- [9] Xuanzhao Dong, Wenhui Zhu, Xiwen Chen, Zhipeng Wang, Peijie Qiu, Shao Tang, Xin Li, and Yalin Wang. Llada-medv: Exploring large language diffusion models for biomedical image understanding. *arXiv preprint arXiv:2508.01617*, 2025. 2
- [10] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palme: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 2
- [11] Zhekai Duan, Yuan Zhang, Shikai Geng, Gaowen Liu, Joschka Boedecker, and Chris Xiaoxuan Lu. Fast ecot: Efficient embodied chain-of-thought via thoughts reuse. *arXiv preprint arXiv:2506.07639*, 2025. 2
- [12] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 2
- [13] Shansan Gong, Shivam Agarwal, Yizhe Zhang, Jiacheng Ye, Lin Zheng, Mukai Li, Chenxin An, Peilin Zhao, Wei Bi, Jiawei Han, et al. Scaling diffusion language models via adaptation from autoregressive models. *arXiv preprint arXiv:2410.17891*, 2024. 2, 3
- [14] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*, 2022. 2
- [15] Ankit Goyal, Valts Blukis, Jie Xu, Yijie Guo, Yu-Wei Chao, and Dieter Fox. Rvt-2: Learning precise manipulation from few demonstrations. *arXiv preprint arXiv:2406.08545*, 2024. 2
- [16] Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. Rvt: Robotic view transformer for 3d object manipulation. In *Conference on Robot Learning*, pages 694–710. PMLR, 2023. 2
- [17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2
- [18] Stephen James, Kentaro Wada, Tristan Laidlow, and Andrew J Davison. Coarse-to-fine q-attention: Efficient learning for visual robotic manipulation via discretisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13739–13748, 2022. 2
- [19] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, pages 991–1002. PMLR, 2022. 2
- [20] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. Vima: General robot manipulation with multimodal prompts. *arXiv preprint arXiv:2210.03094*, 2(3):6, 2022. 2
- [21] Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic vlms: Investigating the design space of visually-conditioned language models. In *Forty-first International Conference on Machine Learning*, 2024. 1, 2
- [22] Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. *arXiv preprint arXiv:2402.10885*, 2024. 5
- [23] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 1, 2, 5
- [24] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 2

- [25] Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, et al. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024. 1, 2, 5, 6
- [26] Shufan Li, Konstantinos Kallidromitis, Hritik Bansal, Akash Gokul, Yusuke Kato, Kazuki Kozuka, Jason Kuen, Zhe Lin, Kai-Wei Chang, and Aditya Grover. Lavidia: A large diffusion language model for multimodal understanding. *arXiv preprint arXiv:2505.16839*, 2025. 2
- [27] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, et al. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024. 5
- [28] Xinghang Li, Peiyan Li, Minghuan Liu, Dong Wang, Jirong Liu, Bingyi Kang, Xiao Ma, Tao Kong, Hanbo Zhang, and Huaping Liu. Towards generalist robot policies: What matters in building vision-language-action models. *arXiv preprint arXiv:2412.14058*, 2024. 2
- [29] Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. Vision-language foundation models as effective robot imitators. *arXiv preprint arXiv:2311.01378*, 2023. 2, 5, 6
- [30] Zhixuan Liang, Yizhuo Li, Tianshuo Yang, Chengyue Wu, Sitong Mao, Liua Pei, Xiaokang Yang, Jiangmiao Pang, Yao Mu, and Ping Luo. Discrete diffusion vla: Bringing discrete diffusion to action decoding in vision-language-action policies. *arXiv preprint arXiv:2508.20072*, 2025. 3, 5
- [31] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 1, 2
- [32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1, 2
- [33] Xiaoran Liu, Zhigeng Liu, Zengfeng Huang, Qipeng Guo, Ziwei He, and Xipeng Qiu. Longllada: Unlocking long context capabilities in diffusion llms. *arXiv preprint arXiv:2506.14429*, 2025. 2
- [34] Zhiyuan Liu, Yicun Yang, Yaojie Zhang, Junjie Chen, Chang Zou, Qingyuan Wei, Shaobo Wang, and Linfeng Zhang. dllm-cache: Accelerating diffusion large language models with adaptive caching. *arXiv preprint arXiv:2506.06295*, 2025. 6
- [35] Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*, 2023. 1
- [36] Rabeeh Karimi Mahabadi, Hamish Ivison, Jaesung Tae, James Henderson, Iz Beltagy, Matthew E Peters, and Arman Cohan. Tess: Text-to-text self-conditioned simplex diffusion. *arXiv preprint arXiv:2305.08379*, 2023. 2
- [37] Oier Mees, Dibya Ghosh, Karl Pertsch, Kevin Black, Homer Rich Walke, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, et al. Octo: An open-source generalist robot policy. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024. 2, 5
- [38] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 7(3):7327–7334, 2022. 5
- [39] Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025. 1, 2, 3
- [40] Dantong Niu, Yuvan Sharma, Giscard Biamby, Jerome Quenum, Yutong Bai, Baifeng Shi, Trevor Darrell, and Roei Herzig. Llarva: Vision-action instruction tuning enhances robot learning. In *8th Annual Conference on Robot Learning*. 2
- [41] Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. *arXiv preprint arXiv:2406.03736*, 2024. 2, 3
- [42] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 2
- [43] Ilija Radosavovic, Baifeng Shi, Letian Fu, Ken Goldberg, Trevor Darrell, and Jitendra Malik. Robot learning with sensorimotor pre-training. In *Conference on Robot Learning*, pages 683–693. PMLR, 2023. 2
- [44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2
- [45] Subham Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184, 2024. 1
- [46] Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. Simplified and generalized masked diffusion for discrete data. *Advances in neural information processing systems*, 37:103131–103167, 2024. 1
- [47] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR, 2023. 2
- [48] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2
- [49] Michael Tschanen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 3

- [50] Quan Vuong, Sergey Levine, Homer Rich Walke, Karl Pertsch, Anikait Singh, Ria Doshi, Charles Xu, Jianlan Luo, Liam Tan, Dhruv Shah, et al. Open x-embodiment: Robotic learning datasets and rt-x models. In *Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition@ CoRL2023*, 2023. 1, 5
- [51] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. *arXiv preprint arXiv:2312.13139*, 2023. 5, 6
- [52] Ling Yang, Ye Tian, Bowen Li, Xincheng Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*, 2025. 1, 2
- [53] Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*, 2025. 2, 3
- [54] Jiasheng Ye, Zaixiang Zheng, Yu Bao, Lihua Qian, and Mingxuan Wang. Dinoiser: Diffused conditional sequence learning by manipulating noises. *arXiv preprint arXiv:2302.10025*, 2023. 2
- [55] Zebin You, Shen Nie, Xiaolu Zhang, Jun Hu, Jun Zhou, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. Llada-v: Large language diffusion models with visual instruction tuning. *arXiv preprint arXiv:2505.16933*, 2025. 1, 2, 3, 6
- [56] Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*, 2024. 2
- [57] Jiaming Zhou, Hongjie Chen, Shiwan Zhao, Jian Kang, Jie Li, Enzhi Wang, Yujie Guo, Haoqin Sun, Hui Wang, Aobo Kong, et al. Diffa: Large language diffusion models can listen and understand. *arXiv preprint arXiv:2507.18452*, 2025. 2
- [58] Zhongyi Zhou, Yichen Zhu, Minjie Zhu, Junjie Wen, Ning Liu, Zhiyuan Xu, Weibin Meng, Ran Cheng, Yaxin Peng, Chaomin Shen, et al. Chatvla: Unified multimodal understanding and robot control with vision-language-action model. *arXiv preprint arXiv:2502.14420*, 2025. 1, 2
- [59] Fengqi Zhu, Rongzhen Wang, Shen Nie, Xiaolu Zhang, Chunwei Wu, Jun Hu, Jun Zhou, Jianfei Chen, Yankai Lin, Ji-Rong Wen, et al. Llada 1.5: Variance-reduced preference optimization for large language diffusion models. *arXiv preprint arXiv:2505.19223*, 2025. 2