QDepth-VLA: Quantized Depth Prediction as Auxiliary Supervision for Vision-Language-Action Models

Yixuan Li
School of Artificial Intelligence, University of Chinese Academy of Sciences
Beijing, China
liyixuan223@mails.ucas.ac.cn

Yuhui Chen
Institute of Automation, Chinese Academy of Sciences
Beijing, China
School of Artificial Intelligence, University of Chinese Academy of Sciences
Beijing, China
chenyuhui2022@ia.ac.cn

Mingcai Zhou
Institute of Automation, Chinese Academy of Sciences
Beijing, China
Beijing Zhongke Huiling Robot Technology Co.
Beijing, China
mingcai.zhou@ia.ac.cn

Haoran Li*
Institute of Automation, Chinese Academy of Sciences
Beijing, China
School of Artificial Intelligence, University of Chinese Academy of Sciences
Beijing, China
lihaoran2015@ia.ac.cn

ABSTRACT

Spatial perception and reasoning are crucial for Vision–Language–Action (VLA) models to accomplish fine-grained manipulation tasks. However, existing approaches often lack the ability to understand and reason over the essential 3D structures necessary for precise control. To address this limitation, we propose QDepth-VLA, a general framework that augments VLA models with an auxiliary depth prediction task. A dedicated depth expert is designed to predict quantized latent tokens of depth maps obtained from a VQ-VAE encoder, enabling the model to learn depth-aware representations that capture critical geometric cues. Experimental results on the simulation benchmarks and real-world tasks demonstrate that QDepth-VLA yields strong spatial reasoning and competitive performance on manipulation tasks.

KEYWORDS

Vision–Language–Action models; Quantized depth prediction; Spatial reasoning; Robotic manipulation

ACM Reference Format:
Yixuan Li, Yuhui Chen, Mingcai Zhou, and Haoran Li. 2026. QDepth-VLA: Quantized Depth Prediction as Auxiliary Supervision for Vision-Language-Action Models. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 10 pages. https://doi.org/10.65109/LJRK3716

1 INTRODUCTION

Large Vision-Language-Action (VLA) models [4, 7, 8, 10, 18] have recently emerged as a powerful paradigm for robotic learning. By grounding pre-trained Vision-Language Models (VLMs) [2, 33, 37] with action-generation capabilities, robots acquire strong generalization across diverse instructions and visual contexts [11]. However, when applied to long-horizon or fine-grained manipulation tasks, these models often exhibit substantial performance degradation [13, 32, 38, 47]. The primary reason lies in a persistent gap between semantic understanding and geometric reasoning [14, 16].

Without reliable 3D understanding, VLAs often misestimate object positions or gripper-object relations, leading to cascading errors during manipulation [32]. Therefore, several recent works have explored incorporating geometric information into VLA models to enable a deeper understanding of the 3D physical environment. These approaches can be broadly generalized into three paradigms: direct 3D feature injection, 2D-projected 3D feature integration and auxiliary 3D information prediction. The first category injects encoded 3D representations, such as point clouds [20] or depth maps [3], into the vision–language backbone or the action head. This strategy typically requires an additional encoder to process 3D features, increasing model complexity and computational cost. While providing explicit geometric cues, it may disrupt the powerful 2D priors learned during large-scale VLM pretraining, leading to degraded visual-language reasoning and understanding. The second category projects 3D features into 2D representations and feeds them into the VLM [21]. Although this preserves pretrained 2D priors, it inevitably introduces information loss in the projection process, which can hinder fine-grained manipulation performance. Compared to these two paradigms, enhancing geometric understanding through auxiliary visual prediction tasks, such as future depth maps estimation [43], offers a more promising alternative. This approach not only preserves the strong 2D priors of pretrained VLMs, but also avoids the need for additional sensory inputs during inference, while encouraging the model to learn 3D-consistent spatial reasoning.

However, existing works that employ depth-map-based visual prediction as auxiliary tasks [43] have not achieved consistent

*Corresponding author.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/LJRK3716

performance improvements, and in some cases even indicate that introducing depth prediction as an auxiliary loss can be detrimental to policy learning due to noisy supervision and weak geometric grounding. The key challenges lie in three aspects. Firstly, the supervision quality of depth maps is often limited by insufficient spatial–temporal consistency across frames [9, 39], introducing substantial noise that weakens geometric grounding. Secondly, pixel-wise depth regression produces highly redundant learning signals, forcing the model to reconstruct every pixel rather than focusing on salient structural cues essential for manipulation. Thirdly, using a vision-language backbone to predict depth maps may interfere with its pre-trained semantic alignment, potentially degrading multimodal reasoning performance.

To address these challenges, we propose **QDepth-VLA**[1], which augments large VLAs by introducing quantized depth prediction as an auxiliary supervision signal. Instead of regressing pixel-wise depth values, QDepth-VLA learns discrete depth representations through vector quantization, capturing salient structural information in a compact and optimization-friendly manner. An independent depth expert is also introduced to predict these quantized depth tokens, enabling the model to leverage geometric cues without interfering with the vision-language backbone's pretrained semantic alignment. Our main contributions are summarized as follows:

(1) We introduce QDepth-VLA, a novel VLA model enhanced with quantized depth information. By integrating a depth prediction task, it internalizes geometric understanding, enabling more accurate reasoning about object spatial relationships.
(2) To facilitate more robust depth learning, we design a specialized *Depth Expert* that predicts quantized depth tokens rather than raw pixel-level depth maps. This formulation effectively mitigates the impact of depth noise and provides a more compact, optimization-friendly supervision signal for geometry-aware policy learning.
(3) Comprehensive experiments on both the Simpler [22] and LIBERO [26] benchmarks demonstrate that QDepth-VLA substantially enhances policy performance, outperforming Open π_0 [31] by 6.1% and 7.7% on average success rate, respectively. Moreover, QDepth-VLA achieves a 10.0% improvement in real-world robotic manipulation, validating its effectiveness and generalizability.

2 RELATED WORKS

2.1 3D-Enhanced VLA

3D spatial information has been widely explored to overcome the limitations of purely 2D-based models. Early efforts typically enhanced spatial perception by either lifting 2D inputs into 3D [14, 16, 17] or directly fusing 2D visual features with 3D point clouds [24, 30, 40].

While these approaches demonstrate that incorporating 3D signals can significantly improve spatial perception and action precision, directly fusing 3D and 2D representations or relying solely on 3D features can disrupt the visual-language alignment established in large-scale VLM pretraining. To mitigate this, two alternative directions have been proposed: (1) Projecting 3D features into 2D space, as in BridgeVLA [21], which renders 3D inputs into multi-view 2D images for compatibility with VLMs. (2) Independent 3D encoders encode geometric information for integration into the action head. This paradigm is employed by PointVLA [20] and GeoVLA [34], where specialized point cloud encoders supply 3D embeddings to modality-specific experts.

Despite these advances, point cloud reconstruction may lose fine-grained object details, and the modality gap between 2D RGB pretraining and 3D geometry remains a persistent challenge. By contrast, depth maps exhibit a much smaller gap with RGB images and thus offer a more natural bridge between 2D and 3D. Recent depth-based approaches have demonstrated this advantage. 3D-CAVLA [3] integrates Region of Interest (RoI) pooling with depth embeddings projected into VLM token space, achieving extraordinary multi-view performance, while 4D-VLA [42] augments visual inputs with 3D coordinate embeddings to support both spatial alignment and temporal reasoning.

Motivated by these insights, we adopt depth maps as the 3D augmentation source. Crucially, instead of directly fusing them with RGB features which risks interfering with pre-trained VLM semantics, we reformulate depth as an auxiliary prediction task. This design enables QDepth-VLA to move beyond passive depth perception toward depth understanding, a capability we elaborate on in the next section.

2.2 Auxiliary Visual Reasoning Tasks for VLA

While depth maps offer a natural bridge between 2D and 3D for enhancing spatial grounding, another promising direction is to strengthen the reasoning capacity of VLAs through auxiliary visual prediction tasks. Instead of passively mapping inputs to actions, policies can be trained to output intermediate signals that make future-oriented reasoning explicit, thereby providing richer supervision during training and improving long-horizon planning at inference.

A series of works focus on predicting future sub-goals, such as generating sub-goal images or short rollouts that visualize task progress. This strategy, as exemplified by CoT-VLA [44] , enhances temporal reasoning by conditioning action generation on both current and predicted states, but incurs high computational cost due to the difficulty of synthesizing realistic RGB predictions. Other researches [8, 18] introduce object-centric signals, such as bounding boxes or spatial relations, which provide structured knowledge of entities and their interactions. More recently, latent future embeddings have been explored, where discrete action tokens predicted in a compressed latent space encode upcoming intentions. AgiBot World Colosseo [1] and UniVLA [6] exemplify this paradigm, showing scalability through large-scale human video pretraining, yet such latent predictions often lack explicit 3D grounding and struggle to capture fine-grained geometry. Finally, some approaches turn to pixel-level 3D supervision, predicting dense depth or semantic maps to reinforce geometric awareness, as in 3D-VLA [46] and DreamVLA [43]. While sometimes effective for strengthening spatial reasoning, these signals are difficult to optimize directly and

[1]Supplementary material and source code are publicly available at: https://github.com/ucasmichael/QDepth-VLA.

may overemphasize redundant low-level cues rather than the most relevant spatial structures.

Different from previous works, our approach unifies 3D information enhancement and visual reasoning by introducing depth codebook prediction—an auxiliary task that brings 3D cues into reasoning in a compact and semantically meaningful way, while remaining naturally aligned with language-conditioned action policies.

3 METHODOLOGY

3.1 Depth Annotation

Since existing VLA datasets such as OXE dataset [29] lack sufficient 3D annotations, we first generate monocular depth estimates for training. To ensure high-quality and spatial–temporal consistent depth sequences, we employ *Video-Depth-Anything* (ViDA) [9], the current state-of-the-art monocular video depth estimation framework built upon a ViT-Large backbone, to acquire depth maps. Specifically, ViDA is applied to the main-view RGB frames from a subset of the OXE [29] and LIBERO [26] datasets to obtain temporally aligned relative depth annotations, providing reliable geometric supervision for depth tokenization and subsequent model training.

3.2 VQ-VAE Reconstruction

To represent depth compactly, we pretrain a Vector-Quantized Variational Autoencoder (VQ-VAE) [35]. Given a depth frame $\mathbf{x}$, the encoder $f_\theta(\cdot)$ produces a latent $\mathbf{z}_e = f_\theta(\mathbf{x})$, which is quantized to the nearest code vector in a codebook $C = \{\mathbf{c}_1, \ldots, \mathbf{c}_K\}$:

$$\mathbf{z}_q = \mathbf{c}_{j^*}, \quad j^* = \arg\min_j \|\mathbf{z}_e - \mathbf{c}_j\|_2^2. \tag{1}$$

We use $K = 256$ codebook entries of dimension $d = 160$, and train the VQ-VAE [35] with the standard objective:

$$\mathcal{L}_{\text{vq}} = \underbrace{\ell_{\text{rec}}(\mathbf{x}, g_\phi(\mathbf{z}_q))}_{\text{reconstruction}} + \underbrace{\|\text{sg}[\mathbf{z}_e] - \mathbf{c}_{j^*}\|_2^2}_{\text{codebook update}} + \beta \underbrace{\|\mathbf{z}_e - \text{sg}[\mathbf{c}_{j^*}]\|_2^2}_{\text{commitment}}, \tag{2}$$

where $\text{sg}[\cdot]$ denotes stop-gradient and $\beta = 0.25$. In practice, we experiment with latent grid resolutions of 16×16 and 32×32. We find that the smaller 16×16 configuration already achieves accurate depth reconstruction while remaining computationally efficient. The VQ-VAE [35] is pretrained independently on each dataset using AdamW [27] with a learning rate of 1×10^{-5} to ensure stable convergence and reconstruction quality. The resulting pretrained model produces discretized depth code indices, which serve as supervisory targets for the depth expert in QDepth-VLA.

3.3 QDepth-VLA Architecture

QDepth-VLA adopts a unified and modular architecture built upon Open π_0 [31], extending its VLA pipeline with an additional depth supervision branch. As shown in Figure 1(a), the model consists of three parameterized modules: a pretrained vision–language model (VLM), an action expert, and a newly introduced *depth expert*. These modules are coordinated through a mixture-of-experts (MoE) structure and a carefully designed hybrid attention mask, enabling QDepth-VLA to jointly reason about geometry and control variants without disrupting pretrained representations.

Table 1: Key configurations of the Action and Depth experts of QDepth-VLA.

	QDepth-VLA	
	Action Expert	**Depth Expert**
Backbone	Transformer	Transformer
Layers / Heads	18 / 8	18 / 8
Hidden dim	1024	1024
Interm. dim	4096	4096
Inputs	Proprio + Action	RGB-Img tokens
Outputs	Actions	Depth tokens

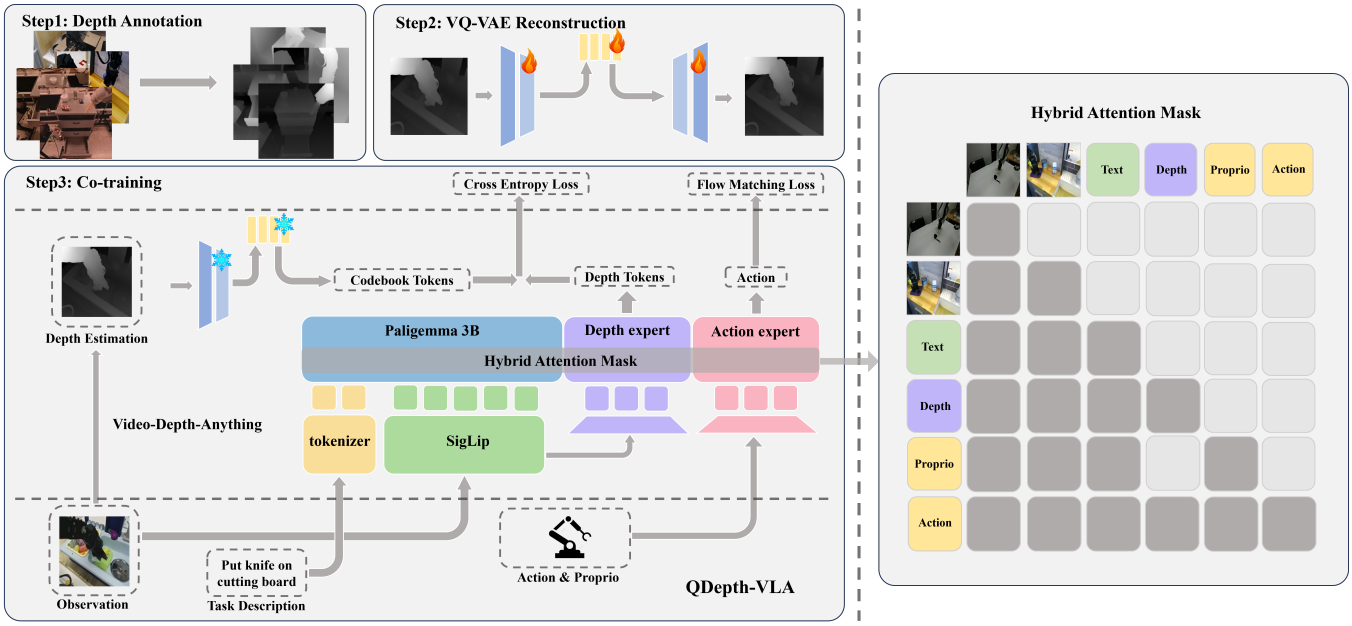
We choose PaliGemma-3B [2] as VLM backbone, which integrates SigLIP-based [41] vision encoding with Gemma's [28] language modeling capability. Input instructions are first tokenized using Gemma's [28] tokenizer, while the main-view RGB image is processed by the SigLIP [41] image encoder to obtain 256 visual tokens. These image tokens are concatenated with 20 text prefix tokens and fed into the Gemma [28] decoder under full block attention to produce multimodal embeddings that capture both spatial and semantic cues. This pretrained VLM remains trainable during the training stage, allowing geometric adaptation to manipulation environment.

The action expert is a transformer-based module responsible for translating multimodal embeddings and proprioceptive states into executable robot actions. It consists of stacked transformer layers with MLP-based encoders and decoders that integrate visual–language context from the VLM with proprioceptive features. This module, which is built upon the original Open π_0 [31] action head, functions as the core control head of QDepth-VLA.

To incorporate geometric reasoning, QDepth-VLA introduces a dedicated depth expert, architecturally aligned with the action expert (illustrated in Table 1). It takes the visual embeddings from the SigLIP [41] encoder as input, before language fusion to avoid semantic interference. These embeddings are projected through a lightweight MLP, processed by a transformer backbone, and then passed to a shallow CNN decoder that predicts 256 depth tokens. Each predicted token corresponding to a latent vector is then aligned with the quantized tokens produced by the pretrained VQ-VAE [35] encoder over its codebook. The pretrained VQ-VAE [35] decoder is subsequently used to reconstruct the spatial depth map from these latent tokens when required. This discrete formulation enables QDepth-VLA to capture compact, structured geometric representations while maintaining optimization stability.

As for hybrid attention mechanism, existing designs typically employ a standard causal attention structure, as seen in DreamVLA [43] and CoT-VLA [44]. However, since depth modalities inherently contain noise, directly fusing them under causal attention may introduce undesirable interference, potentially degrading action generation quality [43]. To address this issue, we redesign the hybrid attention mechanism (Figure 1(b)) to more effectively regulate cross-modal information flow among text, image, depth, proprioception, and action tokens. To be specific:

(1) Text and image tokens attend only within their modality to preserve pretrained semantic grounding.



(a) An Overview of QDepth-VLA (b) The Design of Hybrid Attention Mask

Figure 1: An overview of QDepth-VLA. (a) The overall architecture and training pipeline, where depth supervision is incorporated via a depth expert and latent prediction module. In co-training, the VQ-VAE [35] encoder and codebook are frozen, while PaLI-Gemma 3B [2], the action expert, depth expert, SigLIP [41], and tokenizer are trainable. (b) The proposed hybrid attention mask, which integrates depth and visual tokens to enhance spatial reasoning and manipulation performance.

(2) Depth tokens attend to both image and text tokens, contextualizing geometric features with visual semantics.

(3) Action tokens attend to all preceding modalities, integrating fused perceptual and geometric cues for policy generation.

This hierarchical attention design allows depth to enhance spatial understanding while preventing over-interference with the pretrained VLM and keeping computation efficient.

3.4 Co-Training Procedures

3.4.1 Quantized Depth Supervision. During joint training, the depth expert predicts latent depth tokens, and these tokens are then used to compute logits with the encoded image features over the VQ-VAE [35] codebook:

$$\ell_{i,k} = -\frac{1}{\tau}\|x_i - c_k\|_2^2, \tag{3}$$

where i indexes latent spatial positions , k indexes codebook entries ($K = 256$) and τ represents temperature factor. A cross-entropy loss is applied using ground-truth code indices z_i^* obtained from the pretrained VQ-VAE [35]:

$$\mathcal{L}_{\text{depth}} = -\frac{1}{B \cdot N}\sum_{i=1}^{B \cdot N} \log \frac{\exp(\ell_{i,z_i^*})}{\sum_{k=1}^{K}\exp(\ell_{i,k})}, \tag{4}$$

where B is batch size and N the number of latent tokens per frame. This loss encourages the visual encoder to learn geometry-aware embeddings aligned with the quantized depth representation.

3.4.2 Action Modeling. Based on the underlying VLA backbone, the action prediction objective is as follows:

The Conditional Flow Matching (CFM) action loss [25] is identical to that of π_0 [4]:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{p(A_t|O_t),\, q(\hat{A}_t^\lambda|A_t)}\|f_\theta(\hat{A}_t^\lambda, O_t) - g(\hat{A}_t^\lambda|A_t)\|_2^2, \tag{5}$$

where the action chunk $A_t = [a_t, a_{t+1}, \ldots, a_{t+H-1}]$ is conditioned on the observation $O_t = [I_t, \ell_t, s_t]$, which includes the RGB image, language instruction, and end-effector state. Notably, $\hat{A}_t^\lambda$ denotes noisy action samples generated from a diffusion-like process:

$$\hat{A}_t^\lambda = \lambda A_t + (1-\lambda)\eta, \quad \eta \sim \mathcal{N}(0, I), \tag{6}$$

and the corresponding noise distribution and flow target are defined as:

$$q(\hat{A}_t^\lambda \mid A_t) = \mathcal{N}(\lambda A_t, (1-\lambda)I), \quad g(\hat{A}_t^\lambda \mid A_t) = \eta - A_t. \tag{7}$$

This formulation enables the model to approximate a continuous-time flow field that transports noisy actions toward their clean ground-truth counterparts.

3.4.3 Co-Training Objectives. The total loss combines the action and depth objectives:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{action}} + \lambda_t \cdot \mathcal{L}_{\text{depth}}, \tag{8}$$

where $\lambda_t = \lambda_0 \cdot \gamma^t$ exponentially decays over training steps, with $\lambda_0 = 0.01$. This co-training schedule enables the model to first establish stable geometric alignment before gradually focusing on action refinement.

Table 2: Results of QDepth-VLA on the LIBERO benchmark.

View Setting	Category	Method	Spatial	Object	Goal	Long	Avg
Single-view VLA	General VLA	OpenVLA finetuned [19]	84.7	88.4	79.2	53.7	76.5
		CoT-VLA-7B [44]	87.5	91.6	87.6	69.0	81.1
		Open π_0 [31]	77.2	84.0	83.6	66.0	77.7
	3D-cloud-enhanced VLA	SpatialVLA [30]	**88.2**	89.9	78.6	55.5	78.1
	Depth-enhanced VLA	3D-CAVLA [3]	86.1	**94.7**	82.9	66.8	82.6
		QDepth-VLA (ours)	86.0	88.8	**94.0**	**72.6**	**85.4**
Multi-view VLA	General VLA	Diffusion Policy [12]	78.3	92.5	68.3	50.5	72.4
		Octo finetuned [15]	78.9	85.7	84.6	51.1	75.1
		π_0-FAST finetuned [4]	96.4	96.8	88.6	60.2	85.5
		π_0 finetuned [4]	96.8	98.8	95.8	85.2	94.2
		UniVLA [6]	96.5	96.8	95.6	92.0	95.2
	3D-cloud-enhanced VLA	GeoVLA [34]	**98.4**	99.0	96.6	**96.6**	97.7
	Depth-enhanced VLA	3D-CAVLA [3]	98.2	**99.8**	**98.2**	96.1	**98.1**
		4D-VLA [42]	88.9	95.2	90.9	79.1	88.6
		DreamVLA [43]	97.5	94.0	89.5	89.5	92.6
		QDepth-VLA (ours)	97.6	96.6	95.2	90.0	94.9

3.4.4 Optimization Setup. QDepth-VLA is trained using the AdamW [27] optimizer with decoupled weight decay. We set the learning rate for both the action expert and the VLM backbone to 5×10^{-5}. A cosine learning rate scheduler with 200 warm-up steps and a cycle length of 10^7 steps is applied, ensuring stable optimization throughout training.

4 EXPERIMENTS

In this section, we conduct comprehensive experiments across both simulation and real-world settings to evaluate the effectiveness of our approach. Specifically, we aim to address the following three questions:

(1) Can depth supervision effectively enhance VLA performance in long-horizon and pick-and-place tasks, particularly those requiring fine-grained manipulation?

(2) Is depth supervision more effective than pixel-level depth prediction?

(3) Does the proposed hybrid attention mask contribute to performance gains?

4.1 Simulation Experiments

4.1.1 Training Recipe. The QDepth-VLA based on Open π_0 [31] is initially pre-trained for 9 epochs on the Fractal dataset [5], followed by 20 epochs of pre-training on the LIBERO-90 dataset [26]. After pre-training, the model is further fine-tuned on the four LIBERO subsets — Spatial, Object, Goal and Long [26] for around 50 epochs. For the Simpler benchmark [22], the model is instead trained from scratch, first using the Bridge dataset [36] for 13 epochs, and then the Fractal dataset [5] for an additional 9 epochs.

All experiments are conducted using the Fully Sharded Data Parallel (FSDP) training strategy on 8 × NVIDIA H20 GPUs. A per-GPU batch size of 32 is used, yielding a global batch size of 1024 with gradient accumulation, and the action chunk size is fixed at 4.

4.1.2 Evaluation Setup. For evaluation on LIBERO [26], we adopt its four benchmark suites (Spatial, Object, Goal and Long). Following the preprocessing method in [19], image resolution is first normalized to 256×256 and then resized to 224×224 as model input. We also apply a 180-degree rotation to all images and use only the main-view RGB observations. Each task is evaluated over 50 rollouts, with the average success rate reported.

On the Simpler benchmark [22], the evaluation covers two distinct settings: (1) models trained on the Bridge dataset [36] are tested on tasks involving the WidowX250 robot, and (2) models trained on the Fractal dataset [5] are tested on tasks for the Google Robot. We adopt the visual matching configuration from Simpler [22], evaluating each task across multiple initial positions with 10 rollouts per configuration. Depending on the number of configurations, the total number of evaluations per task ranges from 240 to 2400.

4.1.3 Main Results.

LIBERO Benchmark. QDepth-VLA adopts a *single-view* setting, where the visual input consists of only one RGB image. This contrasts with *multi-view* models, which take multiple images as input, including temporally adjacent frames from historical observations.

As shown in Table 2, QDepth-VLA consistently outperforms single-view baselines across the LIBERO [26] suites. It achieves stronger performance on both fine-grained and long-horizon tasks, reaching 94.0% on the Goal tasks and 72.6% on the Long tasks, surpassing the single-view baseline CoT-VLA [44] by 6.4% and 3.6%, respectively. Compared with Open π_0 [31], QDepth-VLA shows consistent improvements across all four subsets (Spatial, Object, Goal and Long), with the largest gain of 8.8% observed on the Spatial tasks.

While QDepth-VLA operates with only a single RGB observation, its average success rate remains competitive with multi-view VLAs.

Table 3: Results of QDepth-VLA on Simpler benchmark(Google Robot tasks)

	Pick Coke Can	Move Near	Open/Close Drawer	Open Top Drawer and Put Apple In	Avg
RT-2-X [48]	78.7	77.9	25.0	-	60.7
Octo-Base [15]	17.0	4.2	22.7	-	16.8
OpenVLA [19]	16.3	46.2	35.6	-	27.7
RoboVLM finetuned [23]	77.3	61.7	43.5	-	63.4
SpatialVLA finetuned [30]	86.0	77.9	57.4	-	75.1
Open π_0 [31]	97.5	**87.1**	**68.0**	32.9	71.4
QDepth-VLA (ours)	**98.3**	81.4	58.0	**62.6**	**75.1**

Table 4: Results of QDepth-VLA on Simpler benchmark(WidowX250 Robot tasks)

	Put Carrot on Plate	Put Eggplant in Basket	Put Spoon on Towel	Stack Block	Avg
Octo-Base [15]	8.3	43.1	12.5	0.0	16.0
OpenVLA [19]	0.0	4.1	0.0	0.0	1.0
RoboVLM finetuned [23]	25.0	58.3	29.2	12.5	31.3
SpatialVLA finetuned [30]	25.0	**100.0**	16.7	29.2	42.7
Open π_0 [31]	**61.3**	89.6	73.7	15.8	60.0
QDepth-VLA (ours)	57.5	95.0	**82.0**	**39.6**	**68.5**

Specifically, QDepth-VLA achieves a mean success rate only 0.1% lower than π_0-FAST [31], while exceeding 4D-VLA [42] by 3.1% and DreamVLA [43] by 4.5% on the Goal tasks. Moreover, it surpasses π_0-FAST [4] by 12.4% on the more challenging Long tasks. Although leading multi-view models such as 3D-CAVLA [3], GeoVLA [34] and UniVLA [6] achieve higher overall results, our experimental results demonstrate that depth-augmented supervision effectively compensates for the lack of multi-view observations and brings single-view VLAs closer to multi-view performance levels.

By extension, we further implement a *multi-view* variant of QDepth-VLA while maintaining the same setting that predicts latent depth tokens corresponding only to the current main-view image. As shown in Table 2, the multi-view QDepth-VLA consistently outperforms single-view baselines. It achieves an average success rate of 94.9%, surpassing DreamVLA [43] by 0.1% and π_0 [4] by 0.8% on the Spatial tasks, and reaches 90.0% success on the Long tasks. While 3D-CAVLA [3] and GeoVLA [34] achieve higher average success rates, they require explicit point cloud or depth map inputs during inference — modalities that QDepth-VLA does not rely on.

These results reveal that the QDepth-VLA generalizes effectively to multi-view configurations, further enhancing geometric perception and long-horizon reasoning.

Simpler Benchmark. Tables 3 and 4 present the experimental results of QDepth-VLA in the Simpler [22] simulation environment.

As shown in Table 3, QDepth-VLA achieves a success rate of 98.3% on the pick coke can task, surpassing Open π_0 [31] by 0.8% and SpatialVLA [30] by 12.3%. On the more complex open top drawer and put apple in task, QDepth-VLA also attains 62.6%, outperforming Open π_0 [31] by a large margin of 29.7%. This substantial improvement on long-horizon tasks can be attributed to enhanced spatial perception and object localization provided by depth-guided supervision, which improves the model's ability to accurately identify and grasp target objects. As a result, the success probability of intermediate manipulation steps—such as grasping or placing—is increased, leading to higher overall task completion rates.

Table 5: Real-World Evaluation on the Piper Arm

	task1	task2	task3	task4	Avg
ACT [45]	20.0	0.0	0.0	0.0	5.0
Open π_0 [31]	50.0	40.0	40.0	0.0	32.5
QDepth-VLA (ours)	**70.0**	40.0	**50.0**	**10.0**	**42.5**

In Table 4, QDepth-VLA consistently achieves high success rates across various manipulation tasks. Notably, on the stack block task—which demands precise spatial reasoning and fine-grained control—QDepth-VLA reaches a success rate of 39.6%, surpassing SpatialVLA [30] by 10.4%. Furthermore, on the Put Eggplant in Basket and Put Spoon on Towel tasks, QDepth-VLA achieves 95.0% and 82.0% success rates, respectively, outperforming Open π_0 [31] by 5.4% and 8.3%. These improvements highlight the effectiveness of quantized depth supervision in enhancing spatial reasoning and manipulation precision, particularly for tasks involving object placement and coordination in cluttered 3D environments.

Depth Reconstruction Visualization . To further validate the effectiveness of our depth supervision, we visualize depth reconstructions by passing the quantized predicted features through a trained VQ-VAE [35] decoder. As shown in Figure 3, the reconstructions preserve structural details and align well with object boundaries, demonstrating that the learned depth representations capture spatial geometry in a meaningful way.

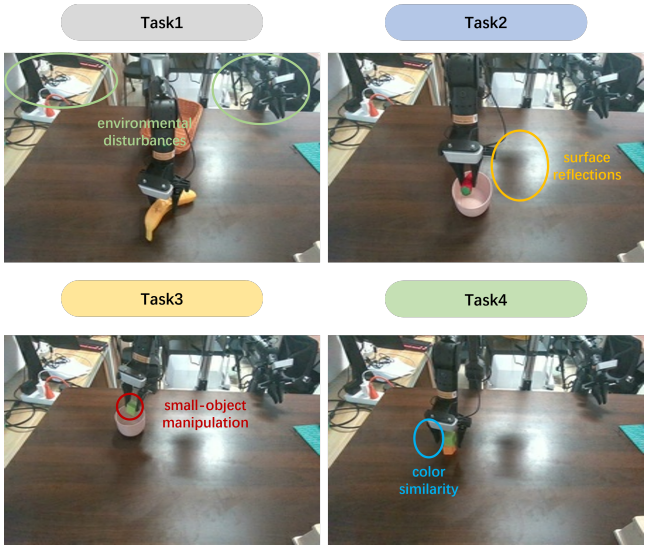


Figure 2: An overview of the main camera view in our real-world task. The environment presents significant challenges for policy learning, including complex lighting conditions, low visual contrast between the gripper and tabletop, and various environmental disturbances that can obscure critical geometric details.

4.2 Real-Robot Experiments

4.2.1 Environment Setup. In the real-world experiments, we employ a 6-DoF Piper robotic arm, with a RealSense D455 camera positioned directly in front of the arm. The training hyperparameters are kept consistent with those used in the simulation experiments, except that the action chunk size is set to 16 to achieve faster execution speed. We select four tasks for evaluation:

- **Task1:** pick the banana into the yellow basket
- **Task2:** put the chili into the bowl
- **Task3:** put the green block into the bowl
- **Task4:** stack the green block on top of the yellow block

For each task, we collect 50 trajectories and fine-tune the model separately on the corresponding dataset. During testing, each task is evaluated over 10 trials.As shown in Figure 2, all experiments are conducted on a dark-colored wooden desk, where the surface color is visually similar to the robot gripper. This setup increases the difficulty of the tasks by introducing additional perceptual ambiguity.

4.2.2 Main Results. We evaluate QDepth-VLA on a series of pick-and-place tasks with varying difficulty to assess its spatial perception and localization ability. We also compare our method against representative baselines such as ACT [45]. As shown in Table 5, QDepth-VLA consistently outperforms ACT [45] across all tasks, while ACT [45] fails to perform reliably in these complex real-world environments, QDepth-VLA achieves robust success rates. Compared to our baseline Open π_0 [31], QDepth-VLA achieves a 20.0% improvement on the simple task of picking a banana, and further achieves gains of 10.0% on both Task 3 and Task 4, demonstrating stronger generalization to challenging scenarios and tasks.

4.3 Ablation Study

To evaluate the contribution of each proposed component, we perform a series of controlled ablation experiments in the Simpler [22] simulation environment. Four ablated variants are considered: (1) removing the depth supervision signal by setting its loss weight to zero (**w/o Depth Loss**); (2) removing the dedicated depth prediction branch (**w/o Depth Expert**); (3) replacing latent depth token prediction with pixel-wise regression (**w/o Latent Prediction**); and (4) substituting the proposed hybrid attention mask with a standard version that enforces proprioception-to-depth attention (**w/o Hybrid Attn**). Quantitative results are reported in Table 6.

4.3.1 w/o Depth Loss. In this variant, the depth loss weight is set to zero while preserving the full model capacity. This configuration isolates the contribution of the depth supervision signal without altering the overall parameter scale.

As shown in Table 6, performance decreases from 68.5% to 65.6% on average. The degradation is most pronounced in the Carrot (-9.6%) and Eggplant (-12.5%) tasks, both requiring coarse spatial grounding. Conversely, slight improvements are observed in Spoon (+7.2%) and Block (+3.3%) tasks. Notably, in the Block task, the depth difference between successful and unsuccessful stacking outcomes can be small, as partially misaligned or unstable stacks may still exhibit similar depth profiles. As a result, depth prediction errors do not necessarily correlate strongly with task failure, and removing the auxiliary depth objective can occasionally simplify optimization and yield modest performance gains.

Overall, these results suggest that depth supervision provides informative depth cues that are beneficial for spatial reasoning beyond the primary control objective.

4.3.2 w/o Depth Expert. Eliminating the dedicated depth branch results in the largest overall performance degradation (-8.5%), as presented in Table 6. The most significant drop occurs in the Stack Block Task (-23.8%), where precise 3D alignment is critical. Substantial declines are also observed in the Eggplant (-5.4%) and Spoon (-8.3%) tasks, indicating that fine-grained spatial reasoning relies heavily on an explicit and specialized depth pathway.

4.3.3 w/o Latent Prediction. As shown in Table 6, replacing latent depth prediction with direct pixel-wise regression lowers average performance to 64.6% (-3.9%). The largest impact is on Eggplant (-14.6%) task, whereas other tasks are only mildly affected. This validates our design choice: quantized latent tokens encourage abstraction of geometric cues, while pixel prediction entangles the model with redundant local detail that is less relevant for manipulation.

4.3.4 w/o Hybrid Attention. In this variant, the proposed hybrid attention mask is replaced with a DreamVLA-style [43] configuration, removing dynamic and semantic modalities. This setting tests whether relative depth maps can enhance proprioceptive state perception and thereby improve action generation quality. As expected, the performance declines by 5.5% on average, with the most substantial drop on the Carrot Task (-15.8%). This result indicates that enforcing proprioception-to-depth attention introduces noise rather than useful guidance, as relative depth lacks absolute positional encoding necessary for stable control.

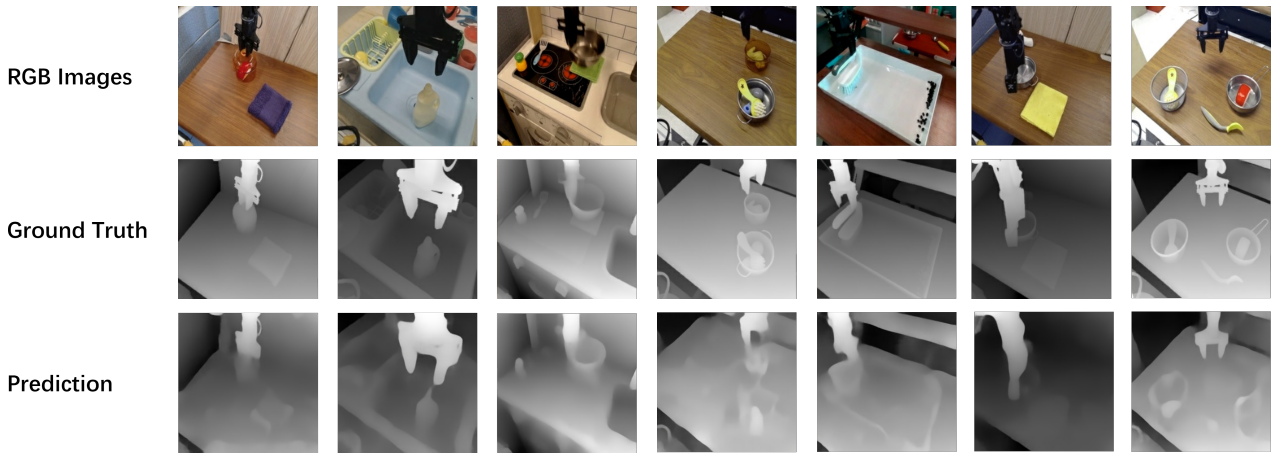


Figure 3: Depth reconstruction results from the QDepth-VLA. Features predicted by depth expert are decoded using a trained VQ-VAE [35] decoder. The reconstructions demonstrate QDepth-VLA's ability to learn critical depth map features, including object and gripper boundaries, underscoring the success of its depth supervision.

Table 6: Ablation study of QDepth-VLA on Simpler tasks. The table reports success rates (%) with module ablations. A checkmark (✓) indicates the module is present, while a cross (✗) indicates it is removed, changed or set to zero manually.

Model	Depth Loss	Depth Expert	Latent Pred	Hybrid Attn	Tasks				Avg
					Carrot	Eggplant	Spoon	Block	
QDepth-VLA	✓	✓	✓	✓	57.5	95.0	82.0	39.6	68.5
w/o Depth Loss	✗	✓	✓	✓	47.9	82.5	89.2	42.9	65.6
w/o Depth Expert	✓	✗	✓	✓	61.3	89.6	73.7	15.8	60.0
w/o Latent Prediction	✓	✓	✗	✓	54.6	80.4	82.0	41.3	64.6
w/o Hybrid Attn [43]	✓	✓	✓	✗	41.7	89.6	78.8	42.0	63.0

4.3.5 Analysis of the Computational Cost. QDepth-VLA introduces a modest and well-contained computational overhead in terms of model size, runtime, and data annotation. Compared to Open π_0 [31], the parameter count increases from 2.606B to 2.924B (+12.2%), mainly due to the additional depth expert and its projection layers, while the core vision-language-action backbone and policy head remain unchanged. At inference time, we observe only minimal runtime overhead in real-robot experiments, with wall-clock control latency remaining comparable to Open π_0 [31]. Depth supervision is obtained via Video-Depth-Anything-Large [9] as an offline preprocessing step, requiring approximately 4 hours to annotate the Bridge and Fractal datasets. The added computational cost represents a favorable trade-off given the consistent gains in spatial reasoning and task success.

Overall, we find the following answers to the three research questions brought up at the beginning of this section,

- Depth supervision effectively enhances VLA performance, especially on long-horizon and fine-grained pick-and-place tasks. In particular, tasks such as stacking and precise placement benefit significantly, indicating improved spatial reasoning.
- Compared to pixel-level regression, quantized depth supervision proves more effective, as it reduces redundancy and focuses learning on salient geometric structures, leading to more stable training and stronger downstream performance.
- The proposed hybrid attention mask consistently contributes to performance gains, particularly in placement tasks, by selectively routing depth cues into the policy network and improving cross-modal feature alignment.

5 CONCLUSION

In this paper, we introduced QDepth-VLA, a new vision-language-action model that incorporates depth supervision and hybrid attention to enhance spatial perception and long-horizon reasoning. Through extensive experiments in both simulation (Simpler and LIBERO) and real-world manipulation tasks, we demonstrate that depth supervision significantly improves manipulation performance. In summary, our work demonstrates that predicting quantized depth tokens at the current timestep is an effective way to enhance policy learning. Extending this approach to predict future depth tokens for improved reasoning and exploring more efficient VAE-based depth representations for enhanced perception present two promising directions for future research.

REFERENCES

[1] AgiBot-World-Contributors, Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xu Huang, Shu Jiang, Yuxin Jiang, Cheng Jing, Hongyang Li, Jialun Li, Chiming Liu, Yi Liu, Yuxiang Lu, Jianlan Luo, Ping Luo, Yao Mu, Yuehan Niu, Yixuan Pan, Jiangmiao Pang, Yu Qiao, Guanghui Ren, Cheng-Xing Ruan, Jiaqi Shan, Yongjian Shen, Chengshi Shi, Mi Shi, Modi Shi, Chonghao Sima, Jia-Yi Song, Huijie Wang, Wenhao Wang, Dafeng Wei, Chengen Xie, Guo-Liang Xu, Junchi Yan, Cunbiao Yang, Lei Yang, Shu-Xiang Yang, Maoqing Yao, Jiansheng Zeng, Chi Zhang, Qingli Zhang, Bin Zhao, Chengyu Zhao, Jiaqi Zhao, and Jianchao Zhu. 2025. AgiBot World Colosseo: A Large-scale Manipulation Platform for Scalable and Intelligent Embodied Systems. *ArXiv* abs/2503.06669 (2025). https://doi.org/10.48550/arXiv.2503.06669

[2] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel M. Salz, Maxim Neumann, Ibrahim M. Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey A. Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Martin Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bovsnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiao-Qi Zhai. 2024. PaliGemma: A versatile 3B VLM for transfer. *ArXiv* abs/2407.07726 (2024).

[3] V.S.K. Pandi V Bhat, Yu-Hsiang Lan, Prashanth Krishnamurthy, Ramesh Karri, and Farshad Khorrami. 2025. 3D CAVLA: Leveraging Depth and 3D Context to Generalize Vision Language Action Models for Unseen Tasks. *ArXiv* abs/2505.05800 (2025).

[4] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. 2025. $\pi 0$: A Vision-Language-Action Flow Model for General Robot Control. In *Proceedings of Robotics: Science and Systems*. LosAngeles, CA, USA. https://doi.org/10.15607/RSS.2025.XXI.010

[5] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael S. Ryoo, Grecia Salazar, Pannag R. Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong T. Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. 2023. RT-1: Robotics Transformer for Real-World Control at Scale. In *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*. https://doi.org/10.15607/RSS.2023.XIX.025

[6] Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. 2025. UniVLA: Learning to Act Anywhere with Task-centric Latent Actions. *ArXiv* abs/2505.06111 (2025). https://doi.org/10.48550/arXiv.2505.06111

[7] Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, Hanbo Zhang, and Minzhao Zhu. 2024. GR-2: A Generative Video-Language-Action Model with Web-Scale Knowledge for Robot Manipulation. *ArXiv* abs/2410.06158 (2024).

[8] Chi-Lam Cheang, Sijin Chen, Zhongren Cui, Yingdong Hu, Liqun Huang, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Xiao Ma, Hao Niu, Wenxuan Ou, Wanli Peng, Zeyu Ren, Haixin Shi, Jiawen Tian, Hongtao Wu, Xin Xiao, Yuyang Xiao, Jiafeng Xu, and Yichu Yang. 2025. GR-3 Technical Report. *ArXiv* abs/2507.15493 (2025).

[9] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. 2025. Video Depth Anything: Consistent Depth Estimation for Super-Long Videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*. 22831–22840. https://openaccess.thecvf.com/content/CVPR2025/html/Chen_Video_Depth_Anything_Consistent_Depth_Estimation_for_Super-Long_Videos_CVPR_2025_paper.html

[10] Yaran Chen, Wenbo Cui, Yuanwen Chen, Mining Tan, Xinyao Zhang, Dongbin Zhao, and He Wang. 2023. RoboGPT: an intelligent agent of making embodied long-term decisions for daily instruction tasks. *CoRR* abs/2311.15649 (2023). https://doi.org/10.48550/ARXIV.2311.15649

[11] Yuhui Chen, Shuai Tian, Shugao Liu, Yingting Zhou, Haoran Li, and Dongbin Zhao. 2025. ConRFT: A Reinforced Fine-tuning Method for VLA Models via Consistency Policy. In *Proceedings of Robotics: Science and Systems*. LosAngeles, CA, USA. https://doi.org/10.15607/RSS.2025.XXI.019

[12] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. 2023. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*. https://doi.org/10.15607/RSS.2023.XIX.026

[13] Yiguo Fan, Pengxiang Ding, Shuanghao Bai, Xinyang Tong, Yuyang Zhu, Hongchao Lu, Fengqi Dai, Wei Zhao, Yang Liu, Siteng Huang, Zhaoxin Fan, Badong Chen, and Donglin Wang. 2025. Long-VLA: Unleashing Long-Horizon Capability of Vision Language Action Model for Robot Manipulation. *ArXiv* abs/2508.19958 (2025).

[14] Théophile Gervet, Zhou Xian, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 2023. Act3D: 3D Feature Field Transformers for Multi-Task Robotic Manipulation. In *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA*, Vol. 229. 3949–3965. https://proceedings.mlr.press/v229/gervet23a.html

[15] Dibya Ghosh, Homer Rich Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Quan Vuong, Ted Xiao, Pannag R. Sanketi, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. 2024. Octo: An Open-Source Generalist Robot Policy. In *Robotics: Science and Systems XX, Delft, The Netherlands, July 15-19, 2024*. https://doi.org/10.15607/RSS.2024.XX.090

[16] Ankit Goyal, Valts Blukis, Jie Xu, Yijie Guo, Yu-Wei Chao, and Dieter Fox. 2024. RVT-2: Learning Precise Manipulation from Few Demonstrations. In *Robotics: Science and Systems XX, Delft, The Netherlands, July 15-19, 2024*. https://doi.org/10.15607/RSS.2024.XX.055

[17] Ankit Goyal, Jie Xu, Yijie Guo, Valts Blukis, Yu-Wei Chao, and Dieter Fox. 2023. RVT: Robotic View Transformer for 3D Object Manipulation. In *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA*, Vol. 229. 694–710. https://proceedings.mlr.press/v229/goyal23a.html

[18] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. 2025. $\pi 0.5$: a Vision-Language-Action Model with Open-World Generalization. *ArXiv* abs/2504.16054 (2025).

[19] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Paul Foster, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. 2024. OpenVLA: An Open-Source Vision-Language-Action Model. In *Conference on Robot Learning, 6-9 November 2024, Munich, Germany*, Vol. 270. 2679–2713. https://proceedings.mlr.press/v270/kim25c.html

[20] Chengmeng Li, Junjie Wen, Yan Peng, Yaxin Peng, Feifei Feng, and Yichen Zhu. 2025. PointVLA: Injecting the 3D World into Vision-Language-Action Models. *ArXiv* abs/2503.07511 (2025).

[21] Peiyan Li, Yixiang Chen, Hongtao Wu, Xiao Ma, Xiangnan Wu, Yan Huang, Liang Wang, Tao Kong, and Tieniu Tan. 2025. BridgeVLA: Input-Output Alignment for Efficient 3D Manipulation Learning with Vision-Language Models. *ArXiv* abs/2506.07961 (2025).

[22] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Oier Mees, Karl Pertsch, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. 2024. Evaluating Real-World Robot Manipulation Policies in Simulation. In *Conference on Robot Learning, 6-9 November 2024, Munich, Germany*, Vol. 270. 3705–3728. https://proceedings.mlr.press/v270/li25c.html

[23] Xinghang Li, Peiyan Li, Minghuan Liu, Dong Wang, Jirong Liu, Bingyi Kang, Xiao Ma, Tao Kong, Hanbo Zhang, and Huaping Liu. 2024. Towards Generalist Robot Policies: What Matters in Building Vision-Language-Action Models. *ArXiv* abs/2412.14058 (2024).

[24] Tao Lin, Gen Li, Yilei Zhong, Yanwen Zou, and Bo Zhao. 2025. Evo-0: Vision-Language-Action Model with Implicit Spatial Understanding. *ArXiv* abs/2507.00416 (2025).

[25] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. 2023. Flow Matching for Generative Modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/forum?id=PqvMRDCJT9t

[26] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. 2023. LIBERO: Benchmarking Knowledge Transfer for Lifelong Robot Learning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*. http://papers.nips.cc/paper_files/paper/2023/hash/8c3c666820ea055a77726d66fc7d447f-Abstract-Datasets_and_Benchmarks.html

[27] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. https://openreview.net/forum?id=Bkg6RiCqY7

[28] Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, and et al. 2024. Gemma: Open Models Based on Gemini Research and Technology. *arXiv preprint arXiv:2403.08295* abs/2403.08295 (2024). https://doi.org/10.48550/arXiv.2403.08295

[29] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, et al. 2024. Open X-Embodiment: Robotic Learning Datasets and RT-X Models : Open X-Embodiment Collaboration. In *IEEE International Conference on Robotics and Automation, ICRA 2024, Yokohama, Japan, May 13-17, 2024*. 6892–6903. https://doi.org/10.1109/ICRA57147.2024.10611477

[30] Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yani Ding, Zhigang Wang, Jiayuan Gu, Bin Zhao, Dong Wang, and Xuelong Li. 2025. SpatialVLA: Exploring Spatial Representations for Visual-Language-Action Model. *ArXiv* abs/2501.15830 (2025).

[31] Allen Z. Ren. 2024. *open-pi-zero: Re-implementation of the π_0 vision-language-action (VLA) model.* https://github.com/allenzren/open-pi-zero

[32] Wenxuan Song, Ziyang Zhou, Han Zhao, Jiayi Chen, Pengxiang Ding, Haodong Yan, Yuxin Huang, Feilong Tang, Donglin Wang, and Haoang Li. 2025. ReconVLA: Reconstructive Vision-Language-Action Model as Effective Robot Perceiver. *ArXiv* abs/2508.10333 (2025).

[33] Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, R. Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim M. Alabdulmohsin, Lucas Beyer, and Xiao-Qi Zhai. 2024. PaliGemma 2: A Family of Versatile VLMs for Transfer. *ArXiv* abs/2412.03555 (2024).

[34] Lin Sun, Bin Xie, Yingfei Liu, Hao Shi, Tiancai Wang, and Jiale Cao. 2025. GeoVLA: Empowering 3D Representations in Vision-Language-Action Models. *ArXiv* abs/2508.09071 (2025).

[35] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural Discrete Representation Learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. 6306–6315. https://proceedings.neurips.cc/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fbc-Abstract.html

[36] Homer Rich Walke, Kevin Black, Tony Z. Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, Abraham Lee, Kuan Fang, Chelsea Finn, and Sergey Levine. 2023. BridgeData V2: A Dataset for Robot Learning at Scale. In *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA*, Vol. 229. 1723–1736. https://proceedings.mlr.press/v229/walke23a.html

[37] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Ke-Yang Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *ArXiv* abs/2409.12191 (2024).

[38] Liudi Yang, Yang Bai, George Eskandar, Fengyi Shen, Mohammad Altillawi, Dong Chen, Soumajit Majumder, Ziyuan Liu, Gitta Kutyniok, and Abhinav Valada. 2025. RoboEnvision: A Long-Horizon Video Generation Model for Multi-Task Robot Manipulation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2025, Hangzhou, China, October 19-25, 2025*. 21281–21288. https://doi.org/10.1109/IROS60139.2025.11246352

[39] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024. Depth Anything V2. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*. http://papers.nips.cc/paper_files/paper/2024/hash/26cfdcd8fe6fd75cc53e92963a656c58-Abstract-Conference.html

[40] Rujia Yang, Geng Chen, Chuan Wen, and Yang Gao. 2025. FP3: A 3D Foundation Policy for Robotic Manipulation. *ArXiv* abs/2503.08950 (2025).

[41] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid Loss for Language Image Pre-Training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. 11941–11952. https://doi.org/10.1109/ICCV51070.2023.01100

[42] Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Yanpeng Zhou, Yuan Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. 2025. 4D-VLA: Spatiotemporal Vision-Language-Action Pretraining with Cross-Scene Calibration. *ArXiv* abs/2506.22242 (2025). https://api.semanticscholar.org/CorpusID:280010742

[43] Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, He Wang, Zhizheng Zhang, Li Yi, Wenjun Zeng, and Xin Jin. 2025. DreamVLA: A Vision-Language-Action Model Dreamed with Comprehensive World Knowledge. *ArXiv* abs/2507.04447 (2025).

[44] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, Ankur Handa, Tsung-Yi Lin, Gordon Wetzstein, Ming-Yu Liu, and Donglai Xiang. 2025. CoT-VLA: Visual Chain-of-Thought Reasoning for Vision-Language-Action Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*. 1702–1713. https://openaccess.thecvf.com/content/CVPR2025/html/Zhao_CoT-VLA_Visual_Chain-of-Thought_Reasoning_for_Vision-Language-Action_Models_CVPR_2025_paper.html

[45] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. 2023. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. In *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*. https://doi.org/10.15607/RSS.2023.XIX.016

[46] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 2024. 3D-VLA: A 3D Vision-Language-Action Generative World Model. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024 (Proceedings of Machine Learning Research, Vol. 235)*. 61229–61245. https://proceedings.mlr.press/v235/zhen24a.html

[47] Junjie Zhu, Huayu Liu, Jin Wang, Bangrong Wen, Kaixiang Huang, Xiao-Fei Li, Haiyun Zhan, and Guodong Lu. 2025. Bridging VLM and KMP: Enabling Fine-grained robotic manipulation via Semantic Keypoints Representation. *ArXiv* abs/2503.02748 (2025).

[48] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong T. Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. 2023. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA*, Vol. 229. 2165–2183. https://proceedings.mlr.press/v229/zitkovich23a.html