

DepthVLA: Enhancing Vision-Language-Action Models with Depth-Aware Spatial Reasoning

Tianyuan Yuan^{1,2}, Yicheng Liu^{1,2}, Chenhao Lu^{1,2}, Zhuoguang Chen¹, Tao Jiang², Hang Zhao^{1,2}

¹IIS, Tsinghua University

²Galaxea AI

yuanty22@mails.tsinghua.edu.cn

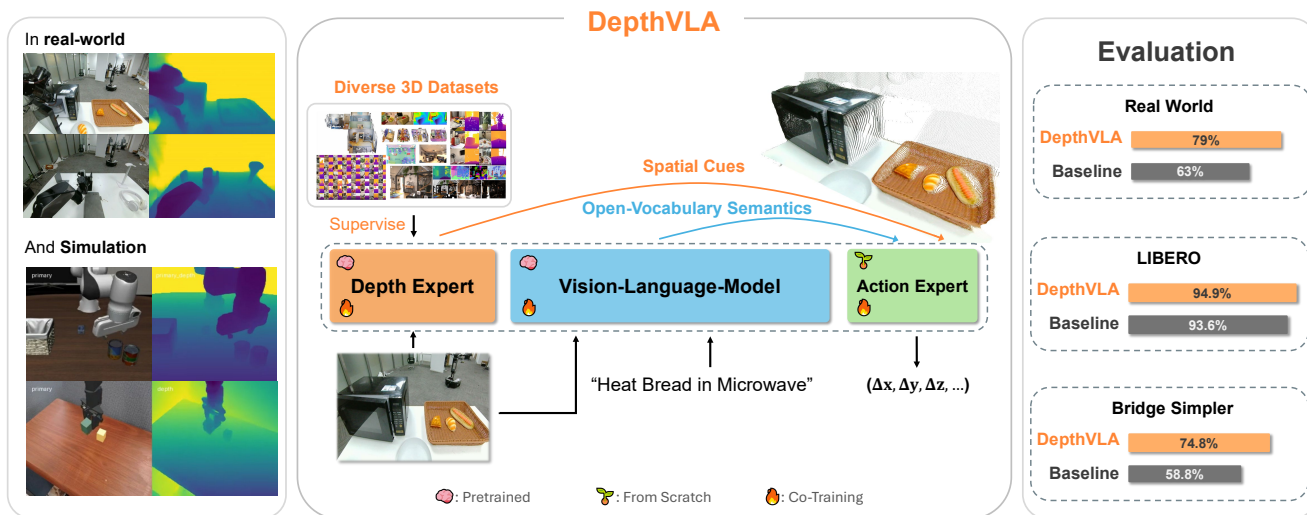


Fig. 1: We propose DepthVLA, a vision-language-action (VLA) model that explicitly incorporates spatial reasoning through a pretrained depth expert. Quantitative comparisons across Simpler, LIBERO, and real-world benchmarks show that DepthVLA consistently outperforms baselines, particularly in tasks requiring fine-grained 3D perception.

Abstract—Vision-Language-Action (VLA) models have recently shown impressive generalization and language-guided manipulation capabilities. However, their performance degrades on tasks requiring precise spatial reasoning due to limited spatial reasoning inherited from Vision-Language Models (VLMs). Existing VLAs rely on extensive action-data pretraining to ground VLMs in 3D space, which reduces training efficiency and is still insufficient for accurate spatial understanding. In this work, we present DepthVLA, a simple yet effective VLA architecture that explicitly incorporates spatial awareness through a pretrained depth prediction module. DepthVLA adopts a mixture-of-transformers design that unifies a VLM, a depth transformer, and an action expert with fully shared attentions, forming an end-to-end model with enhanced spatial reasoning. Extensive evaluations in both real-world and simulated environments show that DepthVLA outperforms state-of-the-art approaches, achieving 78.5% vs. 65.0% progress in real-world tasks, 94.9% vs. 93.6% in the LIBERO simulator, and 74.8% vs. 58.8% in the Simpler simulator. Our code will be made publicly available.

I. INTRODUCTION

Vision-Language-Action (VLA) models [1]–[6] have emerged as a pivotal paradigm in robotic manipulation

research. Built upon large-scale pretrained Vision-Language Models (VLMs), they inherit strong generalization capabilities from vast web data. VLMs provide robust language grounding and semantic visual perception, enabling VLAs to generalize across diverse tasks and embodiments. However, despite their strengths on semantics, VLMs exhibit limited spatial reasoning ability [7], [8], which in turn constrains the spatial perception abilities of VLAs, particularly in tasks requiring precise manipulation [9], [10]. Current VLAs often rely on extensive action-data pretraining to ground VLMs in 3D space [1]–[6], which limits scalability, and pretrained VLAs continue to struggle with precise spatial reasoning. In practice, VLAs often fail at grasping small objects, executing precise operations, or avoiding collisions, highlighting their weak spatial perception.

Recent works have attempted to address this limitation by employing generative world models to predict future states [11]–[16]. While promising, these methods lack explicit 3D knowledge, which we argue is essential for precise manipulation. Another line of work leverages Chain-of-Thought (CoT) reasoning [17] to autoregressively generate

spatial tokens. However, this approach introduces significant latency (over 2 seconds), as hundreds of spatial tokens must be generated before action prediction. To overcome these limitations, we ask: **how can recent advances in 3D perception [18]–[20] be leveraged to enhance VLAs without sacrificing inference speed?**

To address this, we introduce **DepthVLA** (Figure 1), a simple yet effective VLA architecture that explicitly incorporates spatial awareness through a pretrained depth prediction expert. Trained on diverse 3D datasets [21]–[24], this module provides robust geometric understanding. Inspired by π_0 [3], DepthVLA uses a mixture-of-transformers (MoT) [25] design that integrates the depth expert with a VLM and a flow-matching action expert via fully shared attentions, forming an end-to-end VLA model. Intuitively, the VLM provides language understanding and open-vocabulary semantic perception, the depth expert provides fine-grained geometric cues, and the action expert generates actions conditioned on representations from both modalities. The MoT design also enables separate pretraining of each component, allowing training on a more diverse set of data beyond embodied action datasets. Despite adding a depth expert, DepthVLA only increases inference latency marginally, making it practical for real-time deployment.

We validate DepthVLA through extensive experiments in both real-world and simulated environments. We validate DepthVLA through extensive experiments in both real-world and simulated environments. Our evaluations show notable gains in grasping accuracy and collision avoidance, underscoring DepthVLA’s enhanced spatial reasoning. For real-world evaluation, we pretrain on the Galaxea Open-World Dataset [26] and test on the Galaxea R1 Lite, a commercially available dual-arm mobile platform. In simulation, we evaluate on LIBERO [27] and Simpler [28]. Results show that DepthVLA outperforms existing approaches, achieving 78.5% vs. 65.0% success in real-world tasks, 94.9% vs. 93.6% in LIBERO, and 74.8% vs. 58.8% in Simpler, demonstrating the effectiveness of depth-aware representations for precise, generalizable manipulation.

Our contributions are summarized as follows:

- **DepthVLA architecture:** We propose DepthVLA, a novel VLA model that integrates a pretrained depth prediction expert into a mixture-of-transformers framework, enabling explicit spatial reasoning while preserving semantic grounding from VLMs.
- **Per-expert pretraining strategy:** Our MoT design allows each expert (VLM and depth) to be pretrained separately on diverse datasets, improving training efficiency and scalability beyond embodied action data.
- **Extensive real-world and simulated validation:** We demonstrate that DepthVLA significantly outperforms state-of-the-art VLAs in both real-world and simulated environments (LIBERO, Simpler), achieving notable gains in grasping accuracy, collision avoidance, and overall task success.

II. RELATED WORK

A. Generalist Robot Manipulation Policies

Robotic manipulation has evolved from single-task specialists to generalist models trained on broad, diverse datasets covering many tasks and embodiments. Fueled by advances in LLMs, VLMs [29], [30], and large-scale robot action datasets [31], [32], this evolution has given rise to Vision-Language-Action (VLA) models. Early VLAs [1], [2] typically fine-tuned VLMs to autoregressively generate action tokens, which facilitated knowledge transfer but incurred slow inference. More recent VLAs [6], [33] adopt diffusion-based action experts to generate continuous actions more efficiently. Despite differences in action generation, most existing VLAs still require large-scale action-data pretraining to adapt to embodied settings, which is inefficient and still insufficient for fine-grained spatial understanding.

B. VLAs with Spatial Awareness

Prior studies have shown that even state-of-the-art VLMs are insensitive to object shapes and fine geometry [7], [8], limiting their utility for precise manipulation. To enhance spatial perception, early efforts augmented VLAs with additional 3D inputs from sensors such as LiDAR or RGB-D cameras [10], [34], [35], but this reduced generalizability across platforms. SpatialVLA [9] proposes using an off-the-shelf depth estimator to generate pseudo point clouds as input. However, this approach is essentially a workaround, as the depth estimator is not optimized end-to-end with the VLA, limiting its performance upper bound.

More recent approaches incorporate generative world models that predict future frames, keypoints, or semantic states, and then condition action generation on these predictions [11]–[15]. While this improves planning by simulating futures, it does little to improve the encoding of the current scene. A concurrent line of work [17], inspired by methods in VLMs [36], uses Chain-of-Thought (CoT) reasoning to autoregressively generate depth tokens. However, this strategy introduces high latency (over 2 seconds on modern GPUs), as hundreds of tokens must be auto-regressively generated before action prediction.

C. 3D Geometry Perception

Recent advances in 3D perception [18]–[20], [37], [38] have demonstrated strong ability to infer geometry from monocular or multi-view images. By scaling both 3D datasets and model capacity, these vision foundation models achieve robust spatial estimation and support downstream applications such as SLAM [39], [40] and reconstruction [41], [42]. Their progress highlights the potential of integrating powerful 3D priors into VLAs for improved spatial reasoning without requiring additional sensors.

III. METHOD

In this section, we describe DepthVLA, its components, and the training framework.

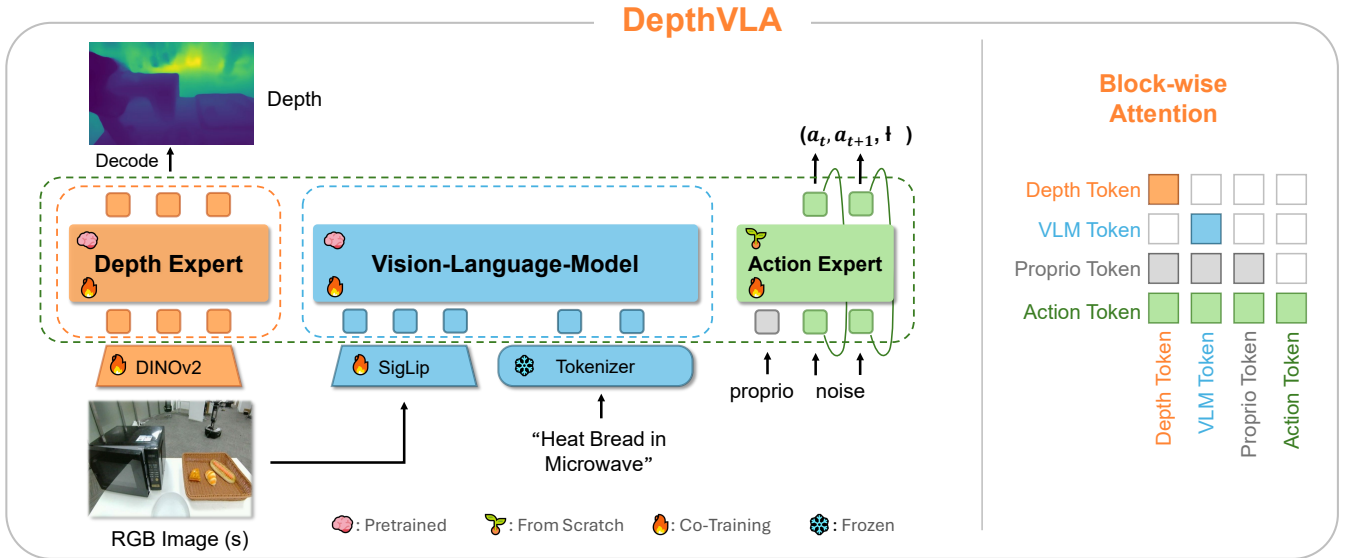


Fig. 2: The proposed mixture-of-transformers (MoT) framework integrates three experts: a vision-language model (VLM) for semantic and language understanding, a depth expert for geometric reasoning, and an action expert for continuous control. Attention layers are shared across experts, while block-wise masking ensures pretrained modules retain their learned abilities. The action expert attends to features from both the VLM and depth expert at every layer to generate actions conditioned on language, visual, and spatial cues.

A. Problem Formulation and Model Overview

We follow the standard end-to-end VLA setting, where a policy π_θ predicts a k -length action chunk $A_t = a_{t:t+k}$ given the current observation o_t (from one or multiple cameras), a language instruction l , and proprioceptive states s_t :

$$A_t = \pi_\theta(o_t, l, s_t).$$

DepthVLA adopts a mixture-of-transformers (MoT) architecture that integrates three experts: a VLM, a depth module, and a flow-matching action expert, as illustrated in Figure 2. This design extends π_0 [3], which uses a two-expert MoT (VLM + action expert), by adding an independent depth expert to provide explicit spatial information.

Specifically, the VLM expert encodes o_t and l to capture semantic and language-grounded features, while the depth expert processes o_t to infer geometric information. The action expert then generates continuous actions conditioned on the combined features from both semantic and geometric experts. All three experts share the same attention layers but maintain distinct weights and feature dimensions.

To preserve the pretrained capabilities of the VLM and depth modules, we apply a block-wise mask: tokens from the VLM and depth experts attend only to themselves, while action tokens can attend to all streams, as shown in right side of Figure 2. This design allows DepthVLA to leverage pretrained knowledge while fusing semantic and spatial cues for precise action generation.

B. Depth Expert

The depth expert serves as a dedicated spatial reasoner, providing geometric cues to guide the action expert. To

integrate seamlessly into the VLA, it adopts the same transformer backbone as the VLM, with separate weights and dimensions.

We design the depth expert as an encoder-decoder architecture. The encoder is based on DINOv2 [43], which captures fine-grained geometric features. We initialize from the pretrained checkpoint of Depth Anything V2 [19] to inherit strong spatial priors from large-scale 3D foundation models. The decoder mirrors the transformer structure of the VLM and outputs depth predictions through a linear head. Unlike approaches that only provide a final depth map [15], [17], we design the depth expert to perform spatial reasoning across all intermediate layers, which provides richer geometric cues for action prediction. The action expert attends to these intermediate features, leveraging rich geometric representations rather than low-dimensional depth outputs. This improves fine-grained spatial understanding, essential for tasks like precise grasping and collision avoidance.

Before integration to VLA, the depth expert is pretrained on diverse 3D datasets using a monocular depth prediction task to acquire robust spatial reasoning ability. We adopt the scale-invariant log loss [44]:

$$\mathcal{L}_{\text{si}}(\hat{d}, d) = \sqrt{\frac{1}{n} \sum_i y^2 - \lambda \left(\frac{1}{n} \sum_i y \right)^2},$$

$$\text{where } y = \log \hat{d} - \log d.$$

Here d is the ground-truth metric depth, \hat{d} is the predicted depth map, and λ controls the balance of the scale term (set to 0.5 by default). This simple loss suffices for learning robust spatial reasoning and distance estimation.

TABLE I: Success rates on the Simpler WidowX benchmark. Models are trained on BridgeData V2 and evaluated zero-shot in simulation. The "Pretrained" column indicates whether the model is pretrained with additional robot action data. DepthVLA achieves the highest average performance.

Model	Pretrained	Put Spoon	Put Carrot	Stack Block	Pick Eggplant	Average
Diffusion Policy [45]	×	4.2%	0%	0%	0%	1.0%
Octo-Base [4]	✓	12.5%	8.3%	0%	43.1%	16.0%
SpatialVLA [9]	✓	16.7%	25.0%	29.2%	100.0%	34.4%
π_0 (re-implemented) [3]	×	81.7%	64.2%	30.0%	59.2%	58.8%
DepthVLA (Ours)	×	75.8%	71.7%	62.5%	89.2%	74.8%

C. DepthVLA Policy Training

We train DepthVLA on embodied action data with an imitation learning objective, maximizing the log-likelihood of actions:

$$\max_{\theta} \mathbb{E}_{p(A_t, o_t, l, s_t)} [\log \pi_{\theta}(A_t | o_t, l, s_t)]$$

To better model continuous and diverse action trajectories, we adopt a flow-matching loss:

$$\mathcal{L}_{\text{flow}}(\theta) = \mathbb{E}_{p(A_t^{\tau} | o_t, l, s_t)} \left[\|v_{\theta}(A_t^{\tau}, \tau, o_t, l, s_t) - u(A_t^{\tau} | A_t)\|^2 \right]$$

Here, subscripts denote robot timesteps and superscripts denote flow matching timesteps, with $\tau \in [0, 1]$. A_t^{τ} is the interpolated noisy action $A_t^{\tau} = \tau A_t + (1 - \tau)\epsilon$. $v_{\theta}(\cdot)$ is the flow predicted by the model and $u(\cdot)$ is the target flow derived from the action trajectory.

To maintain the depth expert’s spatial reasoning, we retain the depth prediction loss during the VLA training. The final loss is therefore:

$$\mathcal{L} = \mathcal{L}_{\text{si}} + \mathcal{L}_{\text{flow}}.$$

This approach allows DepthVLA to jointly optimize spatial reasoning and action generation in an end-to-end manner.

IV. EXPERIMENTS

A. Implementation Details

Model Architecture. We implement all models in PyTorch. We use Paligemma-3B [29] as the VLM backbone, following prior VLA works [3], [9], [26] due to its strong generalization ability. The depth expert employs DINOv2-L as the encoder, initialized from Depth Anything V2 [19], while its decoder is matched in size to the action expert, with both modules containing approximately 300M parameters. As our closest baseline, we re-implement π_0 by strictly following the official JAX implementation. The only difference between our re-implemented π_0 and DepthVLA is the addition of the depth expert, allowing a fair comparison of the impact of explicit spatial reasoning.

Training Details. The depth expert is pretrained on large-scale 3D datasets, including WildRGB-D [22], Scannet [23], Scannet++ [24] and HyperSim [21]. Pretraining runs for 50k steps using a cosine learning rate schedule, with batch size 1024 and initial learning rate 5×10^{-5} . For VLA training, we use a batch size of 1024 for large-scale datasets (e.g., Galaxea

Open-World [26], BridgeData V2 [31]) and 64 for smaller-scale datasets (e.g., LIBERO [27], real-world benchmark tasks). For all models, we do not use any historical information for action generation. All models are trained on 32 NVIDIA H100 GPUs with using the AdamW optimizer [46] with learning rate 2.5×10^{-5} and weight decay 10^{-4} .

Inference Details. DepthVLA introduces 600M additional parameters compared with the baseline π_0 (300M from the DINOv2 encoder and 300M from the depth expert decoder). We run inference on an NVIDIA 4090 GPU with BF16 mixed precision. DepthVLA requires 8.0 GB of VRAM (vs. 6.7 GB for π_0) and has an inference latency of 210 ms per step (vs. 190 ms for π_0). Since actions are predicted in 1-second chunks (16 steps on a 15 Hz platform), the extra latency is negligible in practice.

B. Simulation Benchmarking

BridgeV2 & Simpler. BridgeData V2 [31] is a large-scale real-world robot manipulation dataset, containing over 60k trajectories collected across 24 environments using the WidowX robot. It provides diverse tasks and environment variations, making it a strong foundation for training generalist policies. To obtain depth supervision, we generate pseudo-labels using Depth Anything V2 [19] and UniDepth V2 [33].

Simpler WidowX [28] is a simulation environment designed to closely mirror BridgeData V2, providing a reproducible platform for policy evaluation. It includes four task suites with variations in environment, object configurations, and camera poses, effectively bridging the gap between real and simulated domains. We train DepthVLA on BridgeData V2 for 20k steps (approx. 12 epochs) and evaluate it zero-shot on Simpler WidowX. We report final success rate of each task suite, tested with 120 trials under different random seeds.

Results are shown in Table I. The "Pretrained" column indicates whether a model was pretrained on additional robot action data. DepthVLA achieves the highest average success rate on Simpler WidowX. Compared with the counterpart without a depth expert (*i.e.*, π_0 re-implemented), DepthVLA yields substantial gains on tasks such as block stacking and eggplant picking, which demand accurate spatial reasoning and collision avoidance. Remarkably, the depth expert improves 3D perception even when models are trained on real-world data but evaluated in simulation. Furthermore,

DepthVLA outperforms SpatialVLA [9], a spatial-aware VLA that leverages an external depth estimator, by a wide margin, highlighting the effectiveness of our mixture-of-transformers design.

LIBERO. LIBERO [27] is a simulated manipulation benchmark based on the Franka Panda arm, with demonstrations that include front-view and wrist-view camera images along with natural language instructions. It comprises four task suites: LIBERO-Spatial/-Object/-Goal/-Long, each containing 500 demonstrations across 10 tasks. Unlike prior works [1], [3], [15], [17], which typically train one model per suite, we train a single DepthVLA model jointly on all four suites for 30k steps (about 8 epochs). This creates a more challenging setting that requires stronger generalization across diverse task types. Success rates are reported per task suite, in total 2000 trials across 40 tasks with different random seeds.

Results are shown in Table II. The "Pretrained" column marks whether the model is pretrained on additional robot action datasets. DepthVLA achieves the highest average success rate, even surpassing all models with pretraining. This suggests that standard VLAs, even with large-scale action pretraining, still lack sufficient 3D grounding for precise manipulation. Moreover, DepthVLA surpasses both spatially enhanced baselines (*e.g.*, MolmoACT [17], SpatialVLA [9]) and world-model-based approaches (*e.g.*, DreamVLA [15], CoTVLA [11]), underscoring the strength of our depth expert design.

C. Real-World Benchmarking



Fig. 3: Real-robot experiment platform.

We evaluate DepthVLA on the Galaxea R1 Lite, a commercially available dual-arm mobile platform. The system consists of two 6-DoF arms, two wrist-mounted cameras, and a head camera, as shown in Figure 3. To assess the benefits of large-scale action pretraining on DepthVLA, we pretrain DepthVLA on the large-scale Galaxea Open-World Dataset [26], which contains 100k trajectories across 150 task categories and 50 real-world scenes. Depth labels are generated using VGGT [20] and UniDepth V2 [33]. Pretraining runs for 80k steps (about 4 epochs) for both DepthVLA and the re-implemented π_0 .

To evaluate spatial perception, fine-grained grasping, and collision avoidance, we design three benchmark tasks:

Table bussing: The robot organizes a cluttered desk by placing pens into a holder, hanging headphones, and moving a book onto a stand. This task measures small-object grasping and accurate position estimation.

Microwave operation: The robot opens a microwave door, places food on a plate, puts the plate inside, and closes the door. This task tests collision avoidance at each step.

Blocks stacking: The robot stacks blocks vertically, testing precise pick-and-place skills.

For each benchmark, we collect 100 trajectories and fine-tune the pretrained model for 4k steps. Performance is evaluated using progress scores, where each successful substep in a task contributes one point, and scores are averaged over 20 runs per task. Additionally, we also conduct few-shot experiments with only 20 fine-tuning trajectories to assess DepthVLA’s few-shot transferring ability.

Results are shown in Figure 4. DepthVLA consistently outperforms the baseline, achieving an average progress score of 79% vs. 65% in the standard fine-tuning setting, and 63% vs. 45% in the few-shot setting. On **microwave operation**, it demonstrates improved collision avoidance when handling the door and plate. On **block stacking**, DepthVLA exhibits superior spatial perception, even with limited fine-tuning data, whereas the baseline struggles. On **table bussing**, DepthVLA performs comparably to the baseline, suggesting that both models handle relatively simple small-object grasping tasks effectively. Importantly, DepthVLA maintains language-following capabilities, indicating that the action expert effectively integrates the strengths of both the VLM and depth expert.

D. Ablation Studies

We conduct ablation studies to evaluate the design choices of the depth expert. Specifically, we investigate: (i) Is pretraining the depth expert necessary? (ii) Is the depth loss necessary during VLA training? (iii) What happens if the depth expert is frozen during VLA training? (iv) Is the block-wise mask between VLM and depth tokens necessary? (v) Does predicting depth outperform directly inputting ground-truth depth?

We test these questions under the following settings: (i) Depth expert randomly initialized without pretraining. (ii) Depth loss removed during VLA training. (iii) Depth expert frozen during VLA training. (iv) Depth and VLM tokens allowed to attend to each other. (v) Depth expert taking ground-truth depth as input.

Note that (ii) and (iii) differ, as the depth expert still receives gradients from the flow-matching loss in (ii). Settings (i)–(iv) are evaluated on BridgeData V2 & Simpler, while (v) is evaluated on LIBERO, which provides ground-truth depth maps during inference.

Results are summarized in Table III and Table IV. Each component proves essential for DepthVLA’s effectiveness. Notably, the performance is not greatly impacted when freezing the depth expert, which means the depth expert

TABLE II: Success rates on the LIBERO benchmark across four task suites. The "Pretrained" column indicates whether the model is pretrained with additional robot action data. DepthVLA outperforms all baselines, showing that explicit depth reasoning improves generalization across diverse manipulation tasks.

Model	Pretrained	Spatial	Object	Goal	Long	Average
Octo-Base [4]	✓	78.9%	85.7%	84.6%	51.1%	75.1%
OpenVLA [1]	✓	84.7%	88.4%	79.2%	53.7%	76.5%
SpatialVLA [9]	✓	88.2%	89.9%	78.6%	55.5%	78.1%
CoT-VLA [11]	✓	81.5%	91.6%	87.6%	69.0%	83.9%
MolmoACT [17]	✓	87.0%	95.4%	87.6%	77.2%	86.6%
DreamVLA [15]	✓	97.5%	94.0%	89.5%	89.5%	92.6%
π_0 (re-implemented) [3]	×	95.8%	96.4%	94.8%	87.4%	93.6%
π_0 (reported) [3]*	✓	96.8%	98.8%	95.8%	85.2%	94.2%
DepthVLA (Ours)	×	96.4%	98.0%	95.8%	89.2%	94.9%

* Reported in π_0 official JAX implementation.

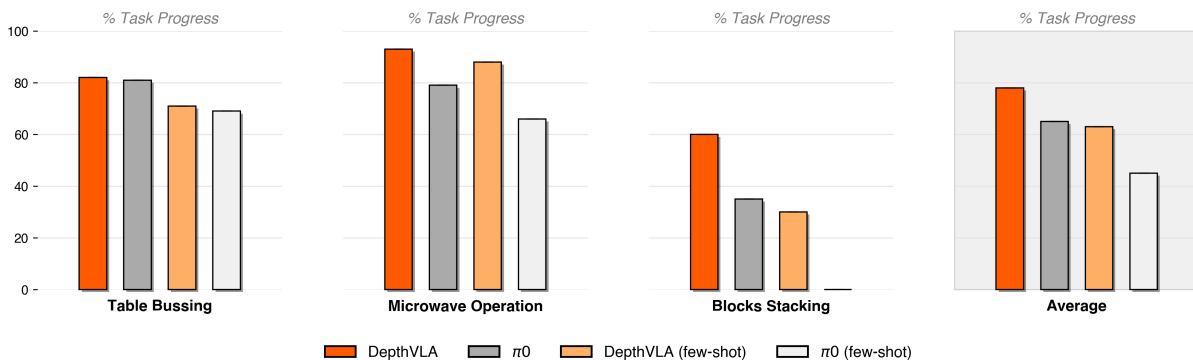


Fig. 4: Performance of DepthVLA and baseline on three bimanual tasks with standard fine-tuning and few-shot adaptation. DepthVLA shows improvements in tasks requiring precise spatial reasoning and collision avoidance while maintaining comparable performance on simpler small-object manipulation.

TABLE III: Ablation studies on different design of the depth expert.

Model	Spoon	Carrot	Block	Eggplant	Average
(i)	60.0%	60.8%	43.3%	40.0%	51.0%
(ii)	69.2%	60%	28.3%	70.0%	56.9%
(iii)	65.8%	69.2%	74.2%	78.3%	71.9%
(iv)	66.7%	65.0%	2.5%	88.3%	55.6%
DepthVLA	75.8%	71.7%	62.5%	89.2%	74.8%

TABLE IV: Comparison between predicted and ground-truth depth inputs. Predicting depth yields stronger performance.

Model	Spatial	Object	Goal	Long	Average
(v)	94.0%	97.6%	95.0%	86.4%	93.3%
DepthVLA	96.4%	98.0%	95.8%	89.2%	94.9%

learned robust and universal spatial representation. It allows DepthVLA to be easily deployed by fine-tuning on demonstrations without the need of depth ground-truth. Another interesting finding is that, the model performs better when predicting depth than when consuming ground-truth depth directly. We hypothesize this is due to modality compe-

tence [47], [48], where one modality can dominate others when jointly provided. By learning to predict depth internally, DepthVLA avoids over-reliance on external signals and instead integrates geometric reasoning more effectively into the shared representation space.

E. Visualization of Depth Prediction

While DepthVLA primarily leverages intermediate features from the depth expert rather than its final outputs, we visualize predicted depth maps to better illustrate the model's spatial perception capabilities.

As shown in Figure 5, the predicted depth captures detailed 3D structure, including object boundaries, distances, and occlusions, which are critical for precise manipulation. Notably, in cluttered environments such as the microwave operation, DepthVLA accurately estimates the relative positions of objects, supporting reliable grasping and collision avoidance. Similarly, in Simpler block stacking and LIBERO-long tasks, the depth predictions provide the action expert with fine-grained geometric cues that improve object alignment and positioning accuracy.

These visualizations demonstrate that the depth expert effectively extracts 3D spatial information from monocular

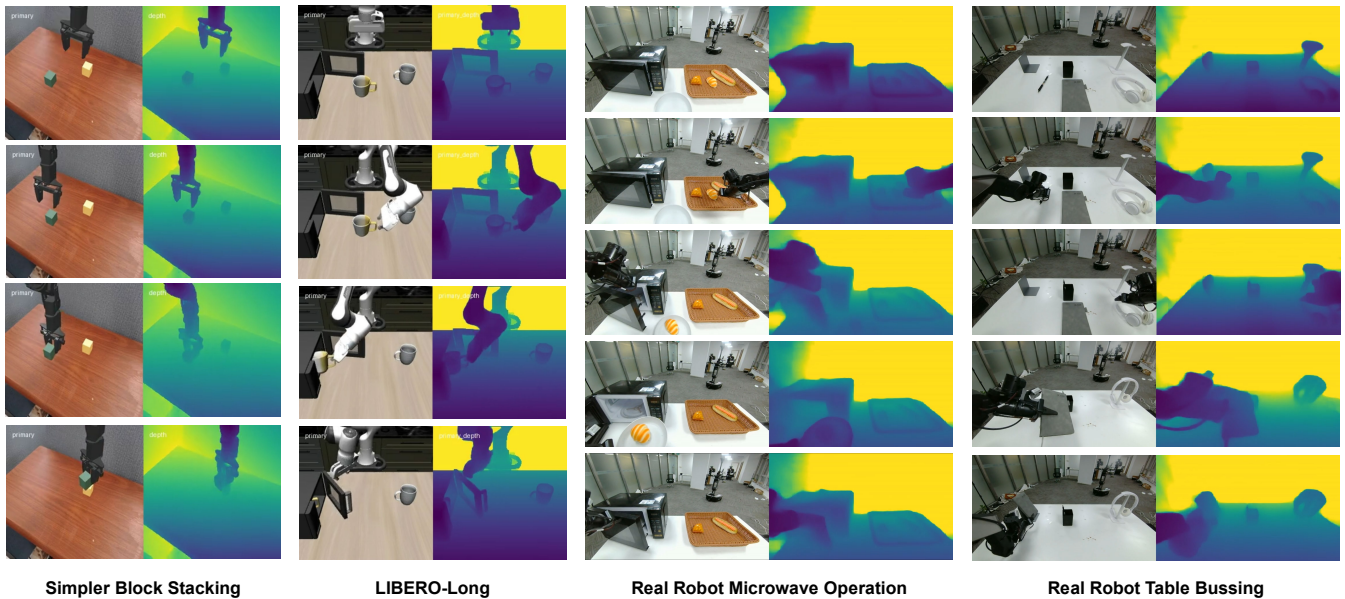


Fig. 5: Qualitative results of DepthVLA’s predicted depth maps across real-world and simulated environments. The predicted depth provides fine-grained geometric cues that guide accurate manipulation.

RGB input. This depth-aware representation complements the semantic grounding provided by the VLM and underpins the performance improvements observed across real and simulated environments.

V. CONCLUSION

We introduced DepthVLA, a VLA model that enhances spatial reasoning by integrating a pretrained depth expert with a VLM and action expert in a unified mixture-of-transformers framework. Experiments in both real-world and simulated environments show that DepthVLA improves performance on tasks requiring precise manipulation, collision avoidance, and fine-grained grasping, while preserving strong language-following capabilities. Ablations confirm the critical role of depth pretraining, depth loss, and attention design in achieving robust 3D perception.

Despite these improvements, DepthVLA has limitations: monocular depth prediction remains an ill-posed and challenging problem. Even when trained on diverse 3D datasets, the depth expert can struggle in difficult scenarios, such as tiny edges, reflective or transparent objects, or texture-less surfaces, which can impact action generation. Future work could explore multi-view depth or pointmap prediction [20], [37], [38] to further enhance spatial accuracy and robustness.

REFERENCES

- [1] M. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn, “Openvla: An open-source vision-language-action model,” *arXiv preprint arXiv:2406.09246*, 2024.
- [2] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choremanski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. G. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, L. Lee, T.-W. E. Lee, S. Levine, Y. Lu, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. Ryoo, G. Salazar, P. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, J. Wu, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich, “Rt-2: Vision-language-action models transfer web knowledge to robotic control,” in *arXiv preprint arXiv:2307.15818*, 2023.
- [3] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, *et al.*, “ π_0 : A vision-language-action flow model for general robot control,” *arXiv preprint arXiv:2410.24164*, 2024.
- [4] Octo Model Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, C. Xu, J. Luo, T. Kreiman, Y. Tan, L. Y. Chen, P. Sanketi, Q. Vuong, T. Xiao, D. Sadigh, C. Finn, and S. Levine, “Octo: An open-source generalist robot policy,” in *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.
- [5] Q. Li, Y. Liang, Z. Wang, L. Luo, X. Chen, M. Liao, F. Wei, Y. Deng, S. Xu, Y. Zhang, *et al.*, “Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation,” *arXiv preprint arXiv:2411.19650*, 2024.
- [6] C. Cheang, S. Chen, Z. Cui, Y. Hu, L. Huang, T. Kong, H. Li, Y. Li, Y. Liu, X. Ma, H. Niu, W. Ou, W. Peng, Z. Ren, H. Shi, J. Tian, H. Wu, X. Xiao, Y. Xiao, J. Xu, and Y. Yang, “Gr-3 technical report,” 2025. [Online]. Available: <https://arxiv.org/abs/2507.15493>
- [7] S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie, “Eyes wide shut? exploring the visual shortcomings of multimodal llms,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 9568–9578.
- [8] P. Rahmzadehgervi, L. Bolton, M. R. Taesiri, and A. T. Nguyen, “Vision language models are blind,” in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, December 2024, pp. 18–34.
- [9] D. Qu, H. Song, Q. Chen, Y. Yao, X. Ye, Y. Ding, Z. Wang, J. Gu, B. Zhao, D. Wang, *et al.*, “Spatialvla: Exploring spatial representations for visual-language-action model,” *arXiv preprint arXiv:2501.15830*, 2025.
- [10] C. Li, J. Wen, Y. Peng, Y. Peng, F. Feng, and Y. Zhu, “Pointvla: Injecting the 3d world into vision-language-action models,” 2025. [Online]. Available: <https://arxiv.org/abs/2503.07511>
- [11] Q. Zhao, Y. Lu, M. J. Kim, Z. Fu, Z. Zhang, Y. Wu, Z. Li, Q. Ma, S. Han, C. Finn, A. Handa, T.-Y. Lin, G. Wetzstein, M.-Y. Liu, and D. Xiang, “Cot-vla: Visual chain-of-thought reasoning for vision-language-action models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025, pp. 1702–1713.
- [12] H. Zhen, X. Qiu, P. Chen, J. Yang, X. Yan, Y. Du, Y. Hong, and C. Gan, “3d-vla: 3d vision-language-action generative world model,” *arXiv preprint arXiv:2403.09631*, 2024.
- [13] J. Zhang, Y. Guo, Y. Hu, X. Chen, X. Zhu, and J. Chen, “Up-vla:

- A unified understanding and prediction model for embodied agent,” *arXiv preprint arXiv:2501.18867*, 2025.
- [14] C. Zhu, R. Yu, S. Feng, B. Burchfiel, P. Shah, and A. Gupta, “Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets,” 2025. [Online]. Available: <https://arxiv.org/abs/2504.02792>
- [15] W. Zhang, H. Liu, Z. Qi, Y. Wang, X. Yu, J. Zhang, R. Dong, J. He, H. Wang, Z. Zhang, L. Yi, W. Zeng, and X. Jin, “Dreamvla: A vision-language-action model dreamed with comprehensive world knowledge,” *CoRR*, vol. abs/2507.04447, 2025. [Online]. Available: <https://doi.org/10.48550/arXiv.2507.04447>
- [16] Y. Tian, S. Yang, J. Zeng, P. Wang, D. Lin, H. Dong, and J. Pang, “Predictive inverse dynamics models are scalable learners for robotic manipulation,” *arXiv preprint arXiv:2412.15109*, 2024.
- [17] J. Lee, J. Duan, H. Fang, Y. Deng, S. Liu, B. Li, B. Fang, J. Zhang, Y. R. Wang, S. Lee, W. Han, W. Pumacay, A. Wu, R. Hendrix, K. Farley, E. VanderBilt, A. Farhadi, D. Fox, and R. Krishna, “Molmoact: Action reasoning models that can reason in space,” 2025. [Online]. Available: <https://arxiv.org/abs/2508.07917>
- [18] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, “Depth anything: Unleashing the power of large-scale unlabeled data,” in *CVPR*, 2024.
- [19] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, “Depth anything v2,” *arXiv:2406.09414*, 2024.
- [20] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, “Vggt: Visual geometry grounded transformer,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- [21] M. Roberts, J. Ramapuram, A. Ranjan, A. Kumar, M. A. Bautista, N. Paczan, R. Webb, and J. M. Susskind, “Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding,” in *International Conference on Computer Vision (ICCV) 2021*, 2021.
- [22] H. Xia, Y. Fu, S. Liu, and X. Wang, “Rgb-d objects in the wild: Scaling real-world 3d object learning from rgb-d videos,” 2024.
- [23] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, “ScanNet: Richly-annotated 3d reconstructions of indoor scenes,” in *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.
- [24] C. Yeshwanth, Y.-C. Liu, M. Nießner, and A. Dai, “ScanNet++: A high-fidelity dataset of 3d indoor scenes,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12–22.
- [25] W. Liang, L. YU, L. Luo, S. Iyer, N. Dong, C. Zhou, G. Ghosh, M. Lewis, W. tau Yih, L. Zettlemoyer, and X. V. Lin, “Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models,” *Transactions on Machine Learning Research*, 2025. [Online]. Available: <https://openreview.net/forum?id=Nu6N69i8SB>
- [26] T. Jiang, T. Yuan, Y. Liu, C. Lu, J. Cui, X. Liu, S. Cheng, J. Gao, H. Xu, and H. Zhao, “Galaxea open-world dataset and g0 dual-system vla model,” 2025. [Online]. Available: <https://arxiv.org/abs/2509.00576>
- [27] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone, “Liberor: Benchmarking knowledge transfer for lifelong robot learning,” *arXiv preprint arXiv:2306.03310*, 2023.
- [28] X. Li, K. Hsu, J. Gu, K. Pertsch, O. Mees, H. R. Walke, C. Fu, I. Lunawat, I. Sieh, S. Kirmani, S. Levine, J. Wu, C. Finn, H. Su, Q. Vuong, and T. Xiao, “Evaluating real-world robot manipulation policies in simulation,” *arXiv preprint arXiv:2405.05941*, 2024.
- [29] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschanen, E. Bugliarello, T. Unterthiner, D. Keysers, S. Koppula, F. Liu, A. Grycner, A. Gritsenko, N. Houlsby, M. Kumar, K. Rong, J. Eisenschlos, R. Kabra, M. Bauer, M. Bošnjak, X. Chen, M. Minderer, P. Voigtlaender, I. Bica, I. Balazevic, J. Puigcerver, P. Papalampidi, O. Henaff, X. Xiong, R. Soricut, J. Harmsen, and X. Zhai, “PaliGemma: A versatile 3B VLM for transfer,” *arXiv preprint arXiv:2407.07726*, 2024.
- [30] S. Karamcheti, S. Nair, A. Balakrishna, P. Liang, T. Kollar, and D. Sadigh, “Prismatic vlms: Investigating the design space of visually-conditioned language models,” 2024. [Online]. Available: <https://arxiv.org/abs/2402.07865>
- [31] H. Walke, K. Black, A. Lee, M. J. Kim, M. Du, C. Zheng, T. Zhao, P. Hansen-Estruch, Q. Vuong, A. He, V. Myers, K. Fang, C. Finn, and S. Levine, “Bridgedata v2: A dataset for robot learning at scale,” in *Conference on Robot Learning (CoRL)*, 2023.
- [32] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, P. D. Fagan, J. Hejna, M. Itkina, M. Lepert, Y. J. Ma, P. T. Miller, J. Wu, S. Belkhale, S. Dass, H. Ha, A. Jain, A. Lee, Y. Lee, M. Memmel, S. Park, I. Radosavovic, K. Wang, A. Zhan, K. Black, C. Chi, K. B. Hatch, S. Lin, J. Lu, J. Mercat, A. Rehman, P. R. Sanketi, A. Sharma, C. Simpson, Q. Vuong, H. R. Walke, B. Wulfe, T. Xiao, J. H. Yang, A. Yavary, T. Z. Zhao, C. Agia, R. Baijal, M. G. Castro, D. Chen, Q. Chen, T. Chung, J. Drake, E. P. Foster, J. Gao, V. Guizilini, D. A. Herrera, M. Heo, K. Hsu, J. Hu, M. Z. Irshad, D. Jackson, C. Le, Y. Li, K. Lin, R. Lin, Z. Ma, A. Maddukuri, S. Mirchandani, D. Morton, T. Nguyen, A. O’Neill, R. Scalise, D. Seale, V. Son, S. Tian, E. Tran, A. E. Wang, Y. Wu, A. Xie, J. Yang, P. Yin, Y. Zhang, O. Bastani, G. Berseth, J. Bohg, K. Goldberg, A. Gupta, A. Gupta, D. Jayaraman, J. J. Lim, J. Malik, R. Martín-Martín, S. Ramamoorthy, D. Sadigh, S. Song, J. Wu, M. C. Yip, Y. Zhu, T. Kollar, S. Levine, and C. Finn, “Droid: A large-scale in-the-wild robot manipulation dataset,” 2025. [Online]. Available: <https://arxiv.org/abs/2403.12945>
- [33] L. Piccinelli, C. Sakaridis, Y.-H. Yang, M. Segu, S. Li, W. Abbeels, and L. V. Gool, “UniDepthV2: Universal monocular metric depth estimation made simpler,” 2025. [Online]. Available: <https://arxiv.org/abs/2502.20110>
- [34] P. Li, Y. Chen, H. Wu, X. Ma, X. Wu, Y. Huang, L. Wang, T. Kong, and T. Tan, “Bridgevla: Input-output alignment for efficient 3d manipulation learning with vision-language models,” 2025. [Online]. Available: <https://arxiv.org/abs/2506.07961>
- [35] Y. Jia, J. Liu, S. Chen, C. Gu, Z. Wang, L. Luo, L. Lee, P. Wang, Z. Wang, R. Zhang, and S. Zhang, “Lift3d foundation policy: Lifting 2d large-scale pretrained models for robust 3d robotic manipulation,” 2024. [Online]. Available: <https://arxiv.org/abs/2411.18623>
- [36] M. Bigverdi, Z. Luo, C.-Y. Hsieh, E. Shen, D. Chen, L. G. Shapiro, and R. Krishna, “Perception tokens enhance visual reasoning in multi-modal language models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025, pp. 3836–3845.
- [37] V. Leroy, Y. Cabon, and J. Revaud, “Grounding image matching in 3d with mast3r,” 2024.
- [38] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, “Dust3r: Geometric 3d vision made easy,” in *CVPR*, 2024.
- [39] J. Zhang, C. Herrmann, J. Hur, V. Jampani, T. Darrell, F. Cole, D. Sun, and M.-H. Yang, “Monst3r: A simple approach for estimating geometry in the presence of motion,” *arXiv preprint arXiv:2410.03825*, 2024.
- [40] R. Murai, E. Dexheimer, and A. J. Davison, “MASt3R-SLAM: Real-time dense SLAM with 3D reconstruction priors,” *arXiv preprint*, 2024.
- [41] H. Wang and L. Agapito, “3d reconstruction with spatial memory,” *arXiv preprint arXiv:2408.16061*, 2024.
- [42] Z. Chen, M. Qin, T. Yuan, Z. Liu, and H. Zhao, “Long3r: Long sequence streaming 3d reconstruction,” *arXiv preprint arXiv:2507.18255*, 2025.
- [43] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, “Dinov2: Learning robust visual features without supervision,” 2023.
- [44] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” 2014. [Online]. Available: <https://arxiv.org/abs/1406.2283>
- [45] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, “Diffusion policy: Visuomotor policy learning via action diffusion,” *The International Journal of Robotics Research*, 2024.
- [46] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” 2019. [Online]. Available: <https://arxiv.org/abs/1711.05101>
- [47] C. Du, J. Teng, T. Li, Y. Liu, T. Yuan, Y. Wang, Y. Yuan, and H. Zhao, “On uni-modal feature learning in supervised multi-modal learning,” 2023. [Online]. Available: <https://arxiv.org/abs/2305.01233>
- [48] X. Zhang, J. Yoon, M. Bansal, and H. Yao, “Multimodal representation learning by alternating unimodal adaptation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 27 456–27 466.