WorldDiT: A Unified Diffusion Architecture for World and Action Modeling

Sen Wang, R. Gnana Praveen, Bidhan Roy and Marcos Villagra
Bagel Labs (bagel.com) 🤗bageldotcom/worlddit

Many recent robot policies pursue stronger control by using large pretrained vision-language models (VLMs) as the action backbone. We introduce WorldDiT, a unified diffusion transformer architecture that couples action generation with visual world modeling and achieves strong performance without a large pretrained VLM action backbone. During training, a single diffusion transformer generates continuous action chunks and predicts normalized RGB patch targets from future camera frames. Across four LIBERO simulation suites, WorldDiT lies on the reported Pareto frontier for total model parameters and mean success among methods reporting all four suites. These results provide a strong sub-billion-parameter baseline for future scaling studies.

Keywords: Robot learning, robot manipulation, diffusion policies, world-action models, world models

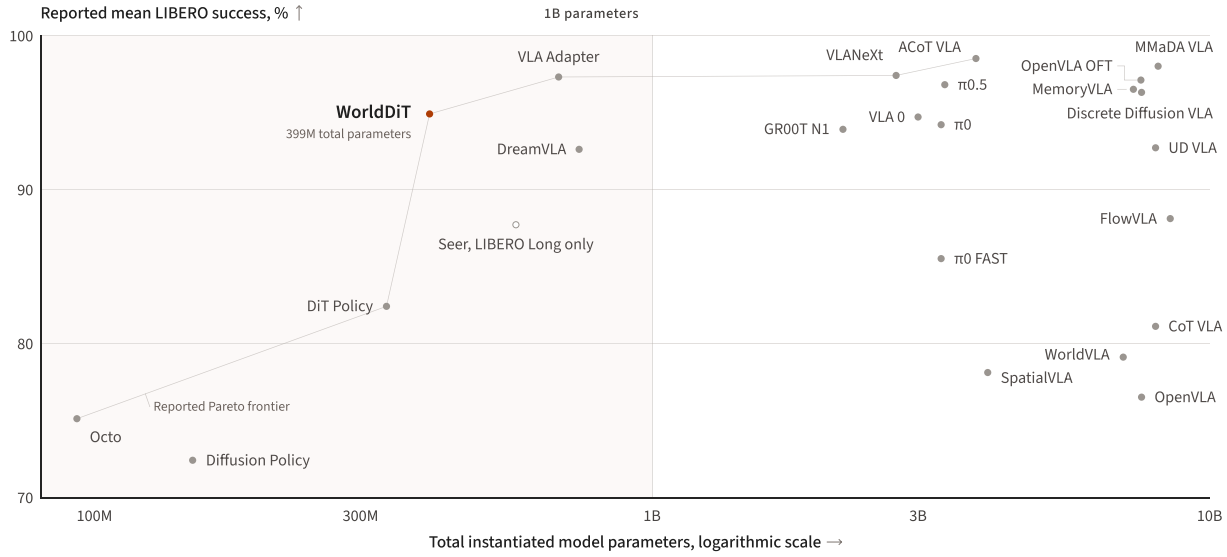

Figure 1 | Reported LIBERO success against total model parameters for 24 methods. The line connects methods with mean success computed across all four suites on the reported Pareto frontier.

1. Introduction

Many robot policies follow a familiar pattern from language modeling, attaching action generation to a large pretrained model. This pattern brings broad perception and language understanding, but it also makes the contributions of scale and architecture difficult to separate. When a policy contains several billion parameters, strong control may come from the action design, the pretrained backbone, or their combination. We therefore ask whether a unified diffusion transformer architecture can combine continuous action generation with an auxiliary future visual prediction objective and retain strong control performance without using a large pretrained vision-language model as the action backbone.

Corresponding email: research@bagel.com

We pursue this objective with WorldDiT, whose frozen visual and language encoders and trainable robot state encoder condition the shared DiT backbone on recent visual observations, robot state, and a language instruction. During training, the shared backbone learns a continuous flow for a seven-step action chunk and an auxiliary normalized RGB patch target selected from future primary-camera and wrist-camera frames. At inference, the shared backbone generates an action chunk. We execute the first three actions, observe the resulting state, and replan.

Figure 2 shows successful rollouts across four LIBERO suites.

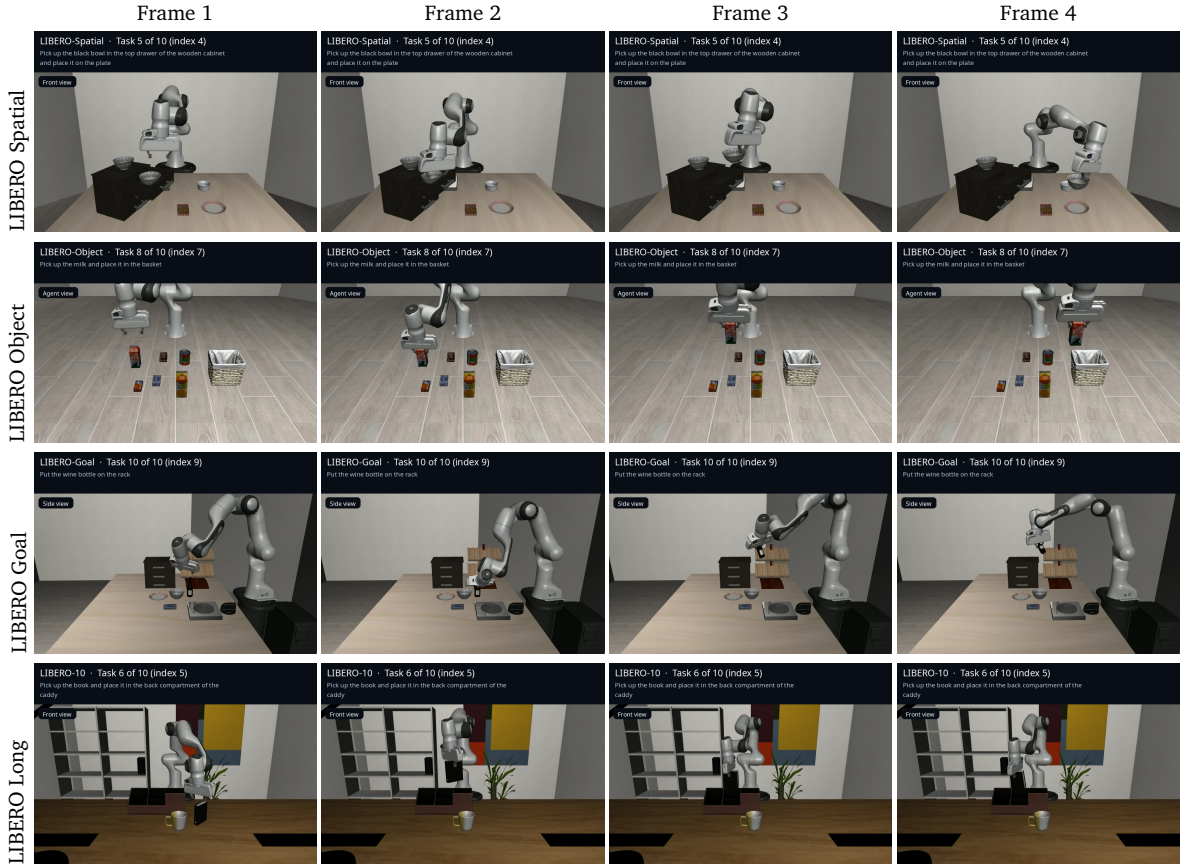

Figure 2 | Each row shows four temporally ordered frames from a successful WorldDiT rollout in one LIBERO suite.

Figure 1 summarizes the reported parameter and performance comparison across all 24 methods. We evaluate one WorldDiT configuration under the specified training and evaluation protocol, providing a baseline for future scaling studies rather than evidence of scaling behavior.

Within this scope, we contribute a unified diffusion transformer architecture for action generation and future normalized RGB patch prediction, a training and deployment design in which RGB patch prediction supplies supervision but is absent at inference, and reported results that characterize the tradeoff between parameter count and mean success below one billion parameters.

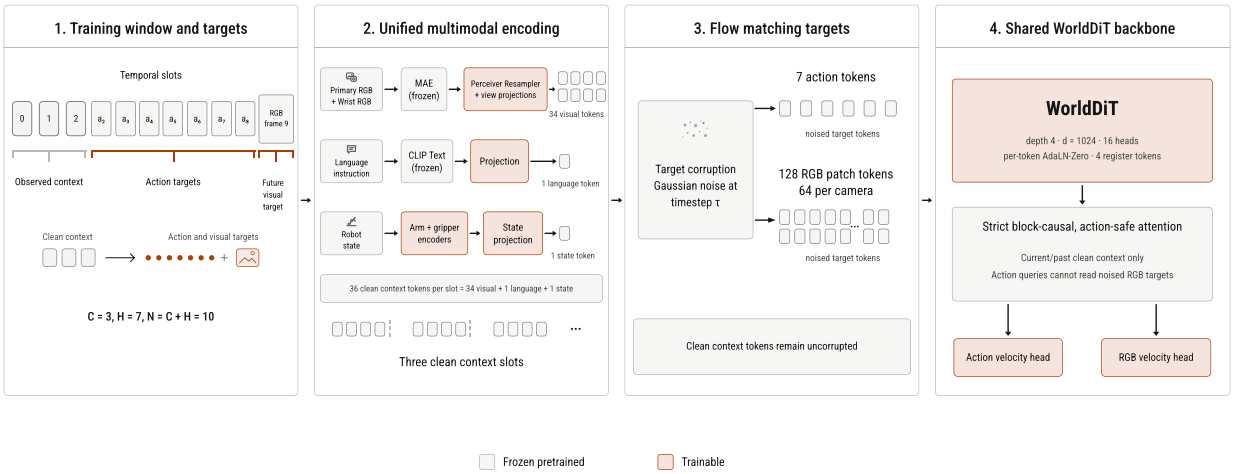

Figure 3 | A ten-step window provides three observed context steps, seven target actions, and one future normalized RGB patch target. Frozen encoders and trainable projections construct multimodal tokens, and one unified WorldDiT backbone predicts action and RGB patch flow velocities.

2. Method

2.1. Overview

We introduce **WorldDiT**, a unified diffusion transformer architecture that learns action generation together with an auxiliary future normalized RGB patch objective. Unlike recent world-action models that rely on a large vision-language model to generate action tokens autoregressively, WorldDiT uses one shared DiT backbone to model continuous robot actions and normalized RGB patches selected from future camera frames.

Given a language instruction and a temporally ordered context of observations from multiple views and robot states, WorldDiT encodes the observed context using frozen visual and language encoders. During training, the shared DiT backbone receives action tokens and normalized RGB patch tokens corrupted with Gaussian noise, selected from future primary-camera and wrist-camera frames, and predicts their flow velocities. Future RGB patch prediction supplies auxiliary supervision during training. Deployment follows the action path directly, keeping RGB patch tokens and the RGB prediction head outside the inference graph.

Figure 3 summarizes this path from the observation window and multimodal tokenization through the shared backbone to final slot supervision.

Let an N-step trajectory segment beginning at index u be

$$\mathcal{W}_u = \{(\mathbf{o}_i, \mathbf{s}_i, \mathbf{a}_i)\}_{i=u}^{u+N-1}, \tag{1}$$

where

$$\mathbf{o}_i = \left(I_i^{\mathrm{p}}, I_i^{\mathrm{w}}\right) \tag{2}$$

contains the primary-camera image I_i^{p} and wrist-camera image I_i^{w}, $\mathbf{s}_i$ is the robot state, and $\mathbf{a}_i \in \mathbb{R}^{d_a}$ is the robot action. A language instruction ℓ is shared across the trajectory segment.

The first C steps form the observation context. Let $i = u + C - 1$ denote the final observed step and

therefore the current control step. The history available at position i is

$$\mathcal{H}_i = \left(\ell, \left\{(\mathbf{o}_j, \mathbf{s}_j)\right\}_{j=u}^{i}\right). \tag{3}$$

The corresponding action target is an H-step action chunk:

$$\mathbf{A}_i = [\mathbf{a}_i, \mathbf{a}_{i+1}, \ldots, \mathbf{a}_{i+H-1}] \in \mathbb{R}^{H \times d_a}. \tag{4}$$

We construct the future normalized RGB patch target from the primary-camera and wrist-camera frames at temporal offset H. Each CLIP-preprocessed 224 by 224 RGB frame is divided into 16 by 16 patches. Each 768-dimensional patch vector is normalized using its own mean and variance, and 64 evenly spaced patches are retained from each camera. Concatenating both camera targets gives

$$\mathbf{Y}_i^{\text{rgb}} \in \mathbb{R}^{128 \times 768}.$$

The loss is applied only to the final temporal slot, whose RGB patch target comes from the frame at $i + H$.

Before entering WorldDiT, a trainable linear input projection maps each corrupted 768-dimensional RGB patch vector to the backbone hidden dimension, and a trainable RGB output projection maps each corresponding hidden representation back to a 768-dimensional patch velocity. These projections define the model interface and do not change the supervision target, which remains the normalized RGB patch vector. For the robot state, we encode the arm and gripper components separately, concatenate their embeddings, and project the result into one state token for each temporal slot.

2.2. Training

We train the model using flow matching [13]. The WorldDiT backbone receives the clean context tokens, action tokens and normalized RGB patch tokens corrupted with Gaussian noise, their timestep embeddings, and learned register tokens, and predicts the corresponding velocities.

The training objective regresses the velocity of a straight path from Gaussian noise to a clean target. For either an action target or a future RGB patch target, let $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{y}$ denote the clean target, sample $\tau \sim \mathcal{U}[0, 1]$, and define

$$\mathbf{x}_\tau = (1 - \tau)\boldsymbol{\epsilon} + \tau \mathbf{y}.$$

The flow-matching objective is

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{\boldsymbol{\epsilon}, \mathbf{y}, \tau}\left[\left\| v_\theta(\mathbf{x}_\tau, \tau, c) - (\mathbf{y} - \boldsymbol{\epsilon}) \right\|_2^2\right].$$

Here c is the context for the current step and $\mathbf{y} - \boldsymbol{\epsilon}$ is the target velocity.

WorldDiT predicts one velocity for each target. The total loss is the weighted sum of the action velocity loss and the RGB patch velocity loss.

$$\mathcal{L}_{\text{total}} = w_{\text{action}} \mathcal{L}_{\text{flow}}^{\text{action}} + w_{\text{rgb}} \mathcal{L}_{\text{flow}}^{\text{rgb}}. \tag{5}$$

The coefficients w_{action} and w_{rgb} are the corresponding loss weights.

WorldDiT uses the C observed steps for conditioning, then predicts one H-step action chunk and one future world target composed of normalized RGB patches from the primary-camera and wrist-camera frames at $i + H$. Only the final temporal slot contributes to the loss.

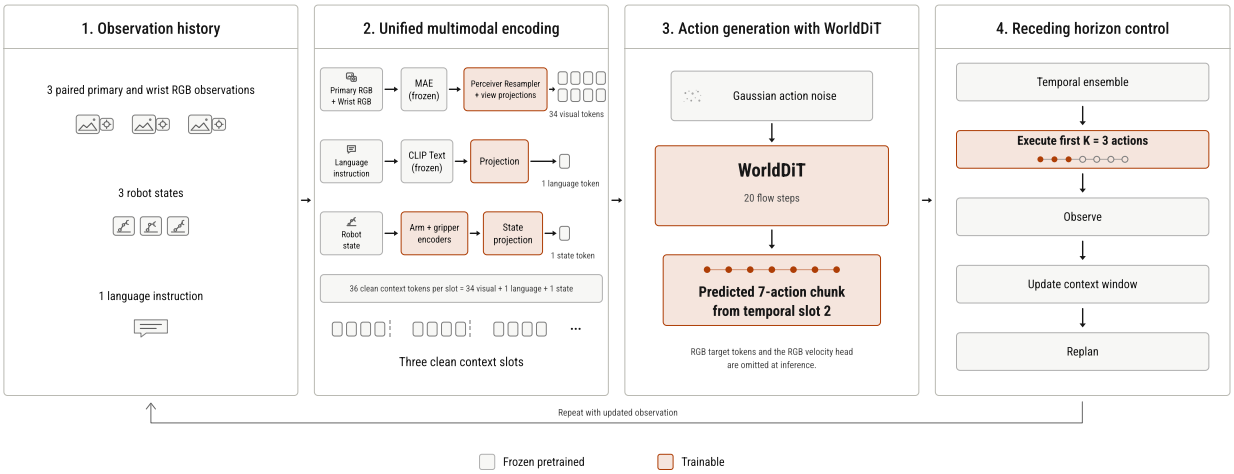

Figure 4 | The unified backbone encodes three observed context steps and samples a chunk of seven actions through twenty flow steps. We execute the first three actions, update the window, and replan.

2.3. Inference

At deployment, the unified backbone receives a C-step history ending at the current control step t:

$$\mathcal{H}_t = \left(\ell, \left\{(\mathbf{o}_j, \mathbf{s}_j)\right\}_{j=t-C+1}^{t}\right). \tag{6}$$

The same multimodal encoding path used during training maps $\mathcal{H}_t$ to the conditioning representation $\mathbf{G}_t$. No future RGB patch targets or future action labels are provided.

Figure 4 summarizes this deployment path from observed history through action generation and receding-horizon control.

Action generation begins from Gaussian noise:

$$\mathbf{x}_{t,0}^{\text{act}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{7}$$

The learned action velocity field is then numerically integrated:

$$\frac{d\mathbf{x}_{t,\tau}^{\text{act}}}{d\tau} = v_\theta^{\text{act}}\left(\mathbf{x}_{t,\tau}^{\text{act}}, \tau, \mathbf{G}_t\right), \qquad \tau \in [0,1]. \tag{8}$$

After T_{samp} integration steps, the final flow state defines the predicted action chunk:

$$\widehat{\mathbf{A}}_t = \mathbf{x}_{t,1}^{\text{act}} \in \mathbb{R}^{H \times d_a}. \tag{9}$$

We execute a prefix of $\widehat{\mathbf{A}}_t$, update the history, and replan.

3. Experiments

3.1. Setup

3.1.1. Data and model configuration

We evaluate WorldDiT on the LIBERO manipulation benchmark [14]. We use the four downstream suites `libero_spatial`, `libero_object`, `libero_goal`, and `libero_10`, reported as LIBERO Long, together with the large multi-task `libero_90` split used only for pretraining. We convert raw demonstrations to fixed-length windows containing multi-view RGB observations, robot states, language instructions, and action sequences. Future primary-camera and wrist-camera frames in each window provide the auxiliary normalized RGB patch targets.

Visual observations, language instructions, and robot states are encoded by a frozen MAE image encoder, a shared Perceiver Resampler, a frozen CLIP text encoder, and a trainable robot state encoder. The resulting tokens are concatenated and used to condition WorldDiT.

The WorldDiT backbone has depth 4, hidden size 1024, 16 attention heads, and 4 register tokens. Each input to the unified backbone contains a temporally ordered context of $C = 3$ observations. The unified backbone predicts an $H = 7$ action chunk, $\mathbf{A}_i \in \mathbb{R}^{7\times7}$, and uses 64 normalized RGB patch tokens from each future camera frame, for 128 target tokens in total. The complete training window contains $N = 10$ steps. We train this configuration solely with the unified flow-matching objective of Section 2.2.

3.1.2. Training

On LIBERO we first pretrain on the multi-task `libero_90` split for 30 epochs with a per-GPU batch size of 40 and gradient accumulation of 2, a learning rate of 1×10^{-4} with a cosine schedule and one warmup epoch, sequence length 10, observation window 10, a context length of 3 observations and a complete data-window length of 10 steps.

For LIBERO we fine-tune the pretrained `libero_90` checkpoint *independently on each downstream suite*. Fine-tuning uses an effective batch size of 512, with a per-GPU batch size of 32 and two gradient accumulation steps on eight GPUs. The unified head is trained on action and normalized RGB patch targets, with action and RGB patch loss weights of 0.1 and 0.001 respectively, with a learning rate of 1×10^{-4} in bf16 mixed precision.

We conduct all training runs on one node with eight RTX Pro 6000 GPUs.

3.1.3. Evaluation

Under this protocol, we evaluate one WorldDiT configuration, fine-tuned separately for each of four suites, over 500 episodes per suite.

At inference there is no clean target. We initialize the action tokens from Gaussian noise and integrate the predicted velocity field from τ=0 to τ=1 with 20 Euler steps, $x_{\tau+\Delta\tau} = x_\tau + \Delta\tau\, v_\theta(x_\tau, \tau, c)$, decoding the resulting action chunk and sliding the context window. All simulator rollouts use headless EGL rendering. On LIBERO we evaluate each suite with action ensembling and report success rates for each task and their mean over the suite's tasks. The unified backbone predicts seven actions. We execute three actions before replanning and apply temporal ensembling to overlapping absolute-time action predictions.

For each suite, the reported WorldDiT score aggregates 500 simulator episodes. The aggregate

includes 300 episodes per suite used during staged checkpoint selection. It is therefore not fully held out, and the reported 94.9% should not be interpreted as an unbiased test estimate. The incomplete public availability of model weights and training code prevents us from reproducing every compared method under a shared protocol. We therefore use scores from the cited reports. The graph and table use total instantiated parameter counts, including frozen modules required at inference, and label reconstructed counts as approximate.

3.2. Results

We evaluate a separately fine-tuned WorldDiT model on each of the four LIBERO suites over 500 simulator episodes. WorldDiT records 98.0% on Spatial, 97.0% on Object, 92.8% on Goal, and 91.8% on Long, producing a mean success rate of 94.9%. Long remains the hardest suite, which is consistent with the extended multistage behavior required by its tasks.

Figure 1 plots the same 24 rows against total instantiated parameter count. Seer is shown at its reported Long score but is excluded from the Pareto calculation because a mean over all four suites is unavailable.

Table 1 shows the reported performance of all 24 methods across the LIBERO Spatial, Object, Goal, and Long suites. Methods are grouped by whether they use a large pretrained vision-language model as the action backbone. The WorldDiT row reports an aggregate over 500 episodes per suite, including 300 episodes per suite used during staged checkpoint selection.

Category	Method	Spatial	Object	Goal	Long	Mean
Large pretrained VLM action backbone	ACoT VLA [26]	98.6	99.0	99.4	97.0	98.5
	MMaDA VLA [15]	98.8	99.8	98.0	95.2	98.0
	VLANeXt [23]	99.0	99.2	96.6	94.8	97.4
	VLA Adapter [22]	97.8	99.2	97.2	95.0	97.3
	OpenVLA OFT [10]	97.6	98.4	97.9	94.5	97.1
	$\pi_{0.5}$ [1]	98.8	98.2	98.0	92.4	96.9
	MemoryVLA [20]	98.4	98.4	96.4	93.4	96.7
	Discrete Diffusion VLA [12]	97.2	98.6	97.4	92.0	96.3
	VLA 0 [7]	97.0	97.8	96.2	87.6	94.7
	π_0 [2]	96.8	98.8	95.8	85.2	94.2
	GR00T N1 [16]	94.4	97.6	93.0	90.6	93.9
	UD VLA [4]	94.1	95.7	91.2	89.6	92.7
	π_0 FAST [17]	96.4	96.8	88.6	60.2	85.5
	CoT VLA [25]	87.5	91.6	87.6	69.0	83.9
	WorldVLA [3]	85.6	89.0	82.6	59.0	79.1
	SpatialVLA [18]	88.2	89.9	78.6	55.5	78.1
	OpenVLA [11]	84.7	88.4	79.2	53.7	76.5
No large pretrained VLM action backbone	**WorldDiT**	**98.0**	**97.0**	**92.8**	**91.8**	**94.9**
	DreamVLA [24]	97.5	94.0	89.5	89.5	92.6
	FlowVLA [27]	93.2	95.0	91.6	72.6	88.1
	DiT Policy [8]	84.2	96.3	85.4	63.8	82.4
	Octo [6]	78.9	85.7	84.6	51.1	75.1
	Diffusion Policy [5]	78.3	92.5	68.3	50.5	72.4
	Seer [21]	N/A	N/A	N/A	87.7	N/A

WorldDiT uses 399.084 million total parameters, including frozen modules required at inference, and 135.107 million trainable parameters. This point lies on the Pareto frontier in the reported comparison. Among methods with a reported mean over all four suites, every method with a higher

reported success value uses more total parameters, while every other method at or below the WorldDiT parameter count reports lower success. This characterizes the tradeoff between parameter count and reported mean success among the published results included in the comparison.

4. Discussion

WorldDiT shows that a single diffusion transformer can couple continuous action generation with future visual prediction while retaining an action-only deployment path. Its reported LIBERO performance places the 399M-parameter system on the parameter–success Pareto frontier, indicating that strong benchmark performance is attainable without placing a large pretrained vision-language model in the action backbone. More broadly, this result supports unified world-and-action modeling as a compact basis for scaling across model capacity, data diversity, and deployment settings. Future work can characterize how the normalized future RGB objective shapes control and whether the same tradeoff persists at larger scales. Since action and visual targets share a flow-matching formulation, WorldDiT may also support decomposition into independently trained experts, following the direction explored in Paris and Paris 2.0, including configurations suited to heterogeneous hardware and compute-constrained robot platforms [9, 19].

References

[1] Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M. R., Finn, C., Fusai, N., Galliker, M. Y., Ghosh, D., Groom, L., Hausman, K., ichter, b., Jakubczak, S., Jones, T., Ke, L., LeBlanc, D., Levine, S., Li-Bell, A., Mothukuri, M., Nair, S., Pertsch, K., Ren, A. Z., Shi, L. X., Smith, L., Springenberg, J. T., Stachowicz, K., Tanner, J., Vuong, Q., Walke, H., Walling, A., Wang, H., Yu, L., and Zhilinsky, U. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In Lim, J., Song, S., and Park, H.-W. (eds.), *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 17–40. PMLR, 27–30 Sep 2025.

[2] Black, K., Brown, N., Driess, D., Esmail, A., Equi, M. R., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., Jakubczak, S., Jones, T., Ke, L., Levine, S., Li-Bell, A., Mothukuri, M., Nair, S., Pertsch, K., Shi, L. X., Smith, L., Tanner, J., Vuong, Q., Walling, A., Wang, H., and Zhilinsky, U. π_0: A Vision-Language-Action Flow Model for General Robot Control. In *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.010.

[3] Cen, J., Yu, C., Yuan, H., Jiang, Y., Huang, S., Guo, J., Li, X., Song, Y., Luo, H., Wang, F., et al. WorldVLA: Towards autoregressive action world model. *arXiv preprint*, 2025.

[4] Chen, J., Song, W., Ding, P., Zhou, Z., Zhao, H., Tang, F., Wang, D., and Li, H. Unified diffusion VLA: Vision-language-action model via joint discrete denoising diffusion process. *International Conference on Learning Representations*, 2026.

[5] Chi, C., Feng, S., Du, Y., Xu, Z., Cousineau, E., Burchfiel, B. C., and Song, S. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. doi: 10.15607/RSS.2023.XIX.026.

[6] Ghosh, D., Walke, H. R., Pertsch, K., Black, K., Mees, O., Dasari, S., Hejna, J., Kreiman, T., Xu, C., Luo, J., Tan, Y. L., Chen, L. Y., Vuong, Q., Xiao, T., Sanketi, P. R., Sadigh, D., Finn, C., and Levine, S. Octo: An Open-Source Generalist Robot Policy. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. doi: 10.15607/RSS.2024.XX.090.

[7] Goyal, A., Hadfield, H., Yang, X., Blukis, V., and Ramos, F. VLA-0: Building state-of-the-art vlas with zero modification. *arXiv preprint*, 2025.

[8] Hou, Z., Zhang, T., Xiong, Y., Pu, H., Zhao, C., Tong, R., Qiao, Y., Dai, J., and Chen, Y. Diffusion transformer policy. *arXiv preprint arXiv:2410.15959*, 2024.

[9] Jiang, Z., Seraj, R., Villagra, M., and Roy, B. Paris: A decentralized trained open-weight diffusion model. *arXiv preprint arXiv:2510.03434*, 2025.

[10] Kim, M. J., Finn, C., and Liang, P. Fine-Tuning Vision-Language-Action Models: Optimizing Speed and Success. In *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.017.

[11] Kim, M. J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E. P., Sanketi, P. R., Vuong, Q., et al. OpenVLA: An open-source vision-language-action model. In *Conference on Robot Learning*, pp. 2679–2713. PMLR, 2025.

[12] Liang, Z., Li, Y., Yang, T., Wu, C., Mao, S., Pei, L., Nian, T., Zhou, S., Yang, X., Pang, J., Mu, Y., and Luo, P. Discrete diffusion VLA: Bringing discrete diffusion to action decoding in vision-language-action policies. *arXiv preprint arXiv:2508.20072*, 2025.

[13] Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[14] Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., and Stone, P. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023.

[15] Liu, Y., Ding, P., Jiang, T., Wang, X., Song, W., Lin, M., Zhao, H., Zhang, H., Zhuang, Z., Zhao, W., et al. Mmada-vla: Large diffusion vision-language-action model with unified multi-modal instruction and generation. *arXiv preprint arXiv:2603.25406*, 2026.

[16] NVIDIA. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

[17] Pertsch, K., Stachowicz, K., Ichter, B., Driess, D., Nair, S., Vuong, Q., Mees, O., Finn, C., and Levine, S. FAST: Efficient Action Tokenization for Vision-Language-Action Models. In *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.012.

[18] Qu, D., Song, H., Chen, Q., Yao, Y., Ye, X., Gu, J., Wang, Z., Ding, Y., Zhao, B., Wang, D., and Li, X. SpatialVLA: Exploring Spatial Representations for Visual-Language-Action Models. In *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.011.

[19] Rouzbayani, A., Roy, B., Villagra, M., and Jiang, Z. Paris 2.0: A decentralized diffusion model for video generation. *arXiv preprint arXiv:2605.26064*, 2026.

[20] Shi, H., Xie, B., Liu, Y., Sun, L., Liu, F., Wang, T., Zhou, E., Fan, H., Zhang, X., and Huang, G. MemoryVLA: Perceptual-cognitive memory in vision-language-action models for robotic manipulation. *International Conference on Learning Representations*, 2026.

[21] Tian, Y., Yang, S., Zeng, J., Wang, P., Lin, D., Dong, H., and Pang, J. Predictive inverse dynamics models are scalable learners for robotic manipulation. In *International Conference on Learning Representations*, 2025.

[22] Wang, Y., Ding, P., Li, L., Cui, C., Ge, Z., Tong, X., Song, W., Zhao, H., Zhao, W., Hou, P., et al. Vla-adapter: An effective paradigm for tiny-scale vision-language-action model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

[23] Wu, X.-M., Fan, B., Liao, K., Jiang, J.-J., Yang, R., Luo, Y., Wu, Z., Zheng, W.-S., and Loy, C. C. VLANeXt: Recipes for building strong vla models. In *Forty-third International Conference on Machine Learning*, 2026.

[24] Zhang, W., Liu, H., Qi, Z., Wang, Y., Yu, X., Zhang, J., Dong, R., He, J., Wang, H., Zhang, Z., et al. Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge. *Advances in Neural Information Processing Systems*, 38:24195–24228, 2025.

[25] Zhao, Q., Lu, Y., Kim, M. J., Fu, Z., Zhang, Z., Wu, Y., Li, Z., Ma, Q., Han, S., Finn, C., Handa, A., Lin, T.-Y., Wetzstein, G., Liu, M.-Y., and Xiang, D. 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)*, pp. 1702–1713, June 2025.

[26] Zhong, L., Liu, Y., Wei, Y., Xiong, Z., Liu, S., and Ren, G. Acot-vla: Action chain-of-thought for vision-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8152–8162, 2026.

[27] Zhong, Z., Yan, H., Li, J., Liu, X., Gong, X., Zhang, T., Song, W., Chen, J., Zheng, X., Wang, H., and Li, H. FlowVLA: Visual chain of thought-based motion reasoning for vision-language-action models. *arXiv preprint arXiv:2508.18269*, 2025.