AR-VLA: True Autoregressive Action Expert for Vision–Language–Action Models

Yutong Hu*†, Jan-Nico Zaech*, Nikolay Nikolov*, Yuanqi Yao* Sombit Dey* Giuliano Albanese*
Renaud Detry† Luc Van Gool* Danda Paudel*
*INSAIT, Sofia University "St. Kliment Ohridski", Bulgaria
†KU Leuven, Belgium

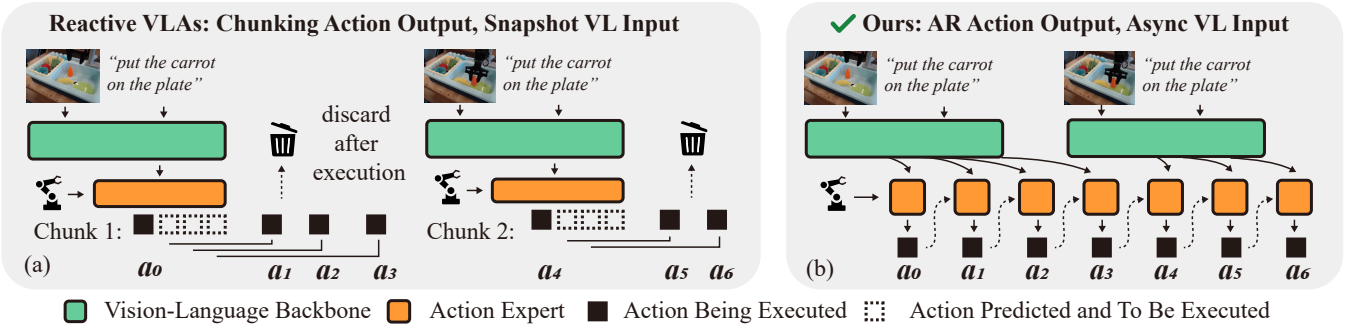


Fig. 1: (a) The prevalent approach in Vision-Language-Action models predicts action chunks based only on the current snapshot of information. It discards the temporal context and drops the state update within an action chunk execution, leading to reactive, memoryless prediction. (b) In contrast, we propose AR-VLA that leverages an autoregressive action expert that maintains its own history through a long-lived memory with inherent context-awareness. Visual-language conditions are updated asynchronously without interrupting the action stream.

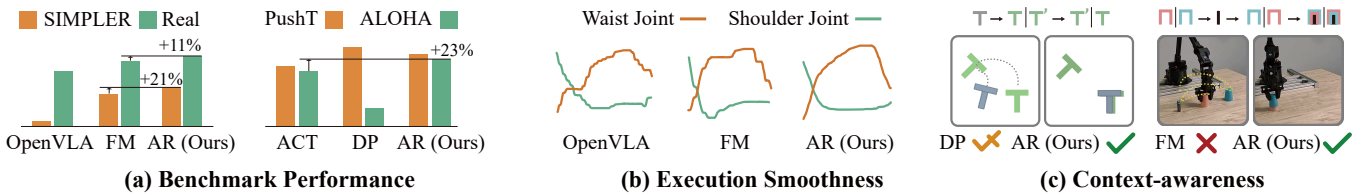
***Abstract*—We propose a standalone autoregressive (AR) Action Expert that generates actions as a continuous causal sequence while conditioning on refreshable vision-language prefixes. In contrast to existing Vision-Language-Action (VLA) models and diffusion policies that reset temporal context with each new observation and predict actions reactively, our Action Expert maintains its own history through a long-lived memory and is inherently context-aware. This structure addresses the frequency mismatch between fast control and slow reasoning, enabling efficient independent pretraining of kinematic syntax and modular integration with heavy perception backbones, naturally ensuring spatio-temporally consistent action generation across frames. To synchronize these asynchronous hybrid V-L-A modalities, we utilize a re-anchoring mechanism that mathematically accounts for perception staleness during both training and inference. Experiments on simulated and real-robot manipulation tasks demonstrate that the proposed method can effectively replace traditional chunk-based action heads for both specialist and generalist policies. AR-VLA exhibits superior history awareness and substantially smoother action trajectories while maintaining or exceeding the task success rates of state-of-the-art reactive VLAs. Overall, our work introduces a scalable, context-aware action generation schema that provides a robust structural foundation for training effective robotic policies.**

I. Introduction

The "next-token prediction" paradigm has emerged as one of the primary engines of modern artificial intelligence. Large-scale autoregressive models, such as LLMs [3] and VLMs [2], demonstrate that the synergy of causal sequence modeling, scalable attention, and massive computation is essential for the appearance of emergent reasoning and robust generalization. Naturally, this paradigm is now being extended from sequences of words to sequences of actions via Vision-Language-Action (VLA) models. However, while recent VLA architectures (e.g., OpenVLA [7], RT-2 [13], Pi-0-FAST [10]) are frequently labeled "autoregressive", this terminology is deceptive in the context of robotic control. These models utilize autoregression only to generate tokens within a single inference step. Effectively, they do not autoregress across time.

Current state-of-the-art robot learning methods, including Diffusion Policies [4] and existing VLAs, treat action generation not as a continuous stream, but as a series of isolated events. As shown in Fig. 1(a), these models typically employ "action chunking" [12]: predicting a static block of actions at once, directly or through iterative denoising. While effective for short-horizon smoothness, these approaches remain structurally reactive: at every perception step, the model acts as if it is "waking up" for the first time, re-encoding the visual context and generating a trajectory chunk without a persistent internal state of its own perception and action history. Consequently, they suffer from "Markovian amnesia", discarding temporal continuity and degrading fluid control to a series of disjointed, snapshot-conditioned responses.

We argue that manipulation is not merely a stack of separate visual-motor snapshots; it is a problem of streaming control. To act effectively, a policy requires two distinct forms of awareness: **situational** awareness (semantic understanding of "what" is in the workspace and "where" the robot is) and **temporal** awareness (kinematic understanding of "what" has already occurred and "how" the end-effector is accelerating).



(a) Benchmark Performance **(b) Execution Smoothness** **(c) Context-awareness**

Fig. 2: **Performance Overview.** (a) Quantitative Results: In both generalist (left) and specialist (right) benchmarks, AR-VLA achieves competitive or superior performance compared to state-of-the-art policies, including OpenVLA, Flow-Matching (FM), ACT, and Diffusion Policy (DP). (b) Trajectory Quality: Qualitative visualization of joint trajectories over time reveals that AR-VLA produces significantly smoother and more kinematically consistent motion compared to reactive baselines that reset context at each step. (c) Long-Horizon Capability: AR-VLA successfully completes long-horizon tasks where baselines like DP and FM fail due to a lack of temporal context awareness.

While VLMs provide the former, they are structurally ill-suited for the latter due to their high latency and episodic nature. The missing piece here is a truly Autoregressive Action Expert – just as an LLM predicts the next word based on the "flow" of a conversation, a robot policy should predict the next pose based on the "momentum" of its trajectory.

By treating action as a "language of motion", a true Autoregressive Action Expert provides three transformative benefits to the VLA paradigm, as in Fig. 2. **(1) It is naturally context-aware,** as its internal state captures the causal dependencies of the entire trajectory rather than reacting to a local snapshot. **(2) It is naturally decoupled from the VLM backbone,** allowing the motor thread to run at high frequencies with temporal consistency, regardless of perception latency. **(3) It facilitates independent pretraining using only the action labels**, enabling the model to master the syntax of movement (dynamics, joint constraints, and physical causality) on large-scale kinematic data before the visual alignment phase.

To realize these potentials, we introduce AR-VLA, a unified framework that instantiates such an action expert within a single architecture for both robot specialists and generalists. As in Fig. 1(b), AR-VLA structurally decouples the high-level semantic reasoning of vision-language models from the high-frequency temporal consistency of robot control. Rather than treating the action head as a dependent appendage of a VLM, we formulate it as an independent expert that maintains a continuous, evolving memory of its own history. This design preserves long-horizon intent while allowing the model to asynchronously attend to the latest visual-language features provided by a VLM. This architecture bridges a fundamental frequency mismatch in robotics, providing a solution one step closer to the **system 1/2** [6, 1] dichotomy: the "brain" (semantic perception) updates slowly, while the "cerebellum" (motor control) streams high-frequency commands.

Our core contribution is **the formulation of an Autoregressive Action Expert**, which treats action generation as a causal sequence modeling problem across time. By maintaining a long-lived context of past actions, our model inherently resolves the temporal inconsistency and "jitter" prevalent in reactive policies, outperforming denoise/chunk-based baselines in trajectory smoothness and long-horizon stability. To instantiate this expert, we propose two technical pillars: **(1) Hybrid Key-Value (HKV) Cache:** A novel Transformer decoder architecture that manages two distinct memory streams: a rolling, token-wise FIFO for high-frequency actions and a block-wise, refreshable buffer for low-frequency visual semantics. This allows the action stream to function as an independent expert that is "guided" rather than "blocked" by perception. **(2) Dynamic Temporal Re-anchoring (DTR):** We solve the synchronization challenge of asynchronous streams via DTR, a mechanism that explicitly anchors visual keys based on their capture-time index. This ensures the model mathematically understands the "staleness" of a visual frame, bridging the gap between short-context training and long-horizon inference.

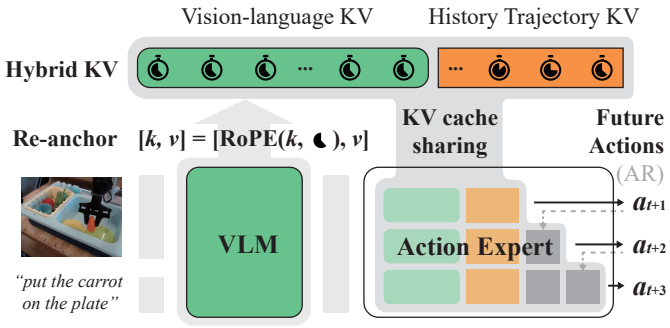


Fig. 3: **The AR-VLA Framework**. The system bridges an VLM backbone with a autoregressive Action Expert asynchronously. Atemporal features from the VLM are explicitly injected with temporal context via Dynamic Temporal Re-anchoring (DTR). Within the Hybrid KV Cache, re-anchored VL tokens (green) serve as a semantic prefix to the rolling kinematic history (orange). The Action Expert generates future action sequences by querying this shared cache using incrementally advancing step embeddings.

II. Methodology

We formulate robot trajectory generation not as a stateless, reactive mapping, but as an autoregressive (AR) sequence modeling problem. Standard Vision-Language-Action models (VLAs) suffer from a Markovian bottleneck—re-inferring intent at every step—which causes jittery control. Our proposed **AR-VLA** overcomes this by explicitly conditioning future actions on a continuous kinematic history alongside the most recent visual-language (VL) prefix, ensuring smooth control and robustness to visual latency.

Unified Decoder and Hybrid Cache: AR-VLA is instantiated as a unified Transformer decoder that decouples perception

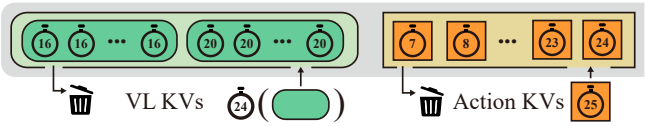


Fig. 4: **Heterogeneous FIFO Update Rules for the Hybrid KV Cache.** The framework manages memory through two distinct queueing strategies to ensure efficient context utilization. The VL Stream (green) operates as a short-lived, block-wise FIFO: In contrast, the Action Stream (orange) maintains a token-wise rolling FIFO, continuously appending the single latest action prediction while evicting the oldest kinematic state.

and control via a Hybrid Key-Value (HKV) Cache. This cache manages two distinct memory streams: a rolling, token-wise buffer for continuous proprioceptive history, and a single-slot, refreshable snapshot for atemporal VL embeddings provided by the VLM backbone.

Dynamic Temporal Reanchoring (DTR): To temporally align the high-frequency motor thread with asynchronous semantic updates, we introduce DTR. DTR leverages Rotary Positional Embeddings (RoPE) to mathematically encode the relative data staleness between an action query and the previously captured VL context. This shift-invariant property ensures the attention mechanism generalizes perfectly to varying temporal offsets, resolving the discrepancy between short training horizons and extended real-world deployment.

Training and Inference: The framework is trained in two phases: an action-only pretraining phase to master kinematic syntax, followed by a VL-action alignment phase utilizing stochastic historical dropout to force visual grounding. During deployment, the decoupled HKV cache enables an asynchronous dual-thread execution mode. The action thread autoregressively generates high-frequency control while the perception thread asynchronously pushes updates to the semantic cache, guaranteeing uninterrupted actuation independent of VLM inference delays.

III. Experiments

Our evaluation demonstrates that AR-VLA provides a competitive, highly efficient alternative to existing action experts, offering superior history awareness and high-frequency control. We assess its performance across generalist and specialist policies, inference efficiency, long-horizon temporal grounding, and architectural ablations.

SimplerEnv & Real-World VLA: As detailed in Tab. I and Fig. 6, AR-VLA achieves state-of-the-art zero-shot transfer. In SimplerEnv, it reaches a 61.5% average success rate, outperforming the next-best baseline (CogACT, 52.1%) and identical-scale chunking models (Pi-0.5, 51.0%). In real-world WidowX deployments, AR-VLA achieves an 89% average success rate. Crucially, its inherent temporal awareness allows for graceful recovery and re-grasping upon initial failures, unlike reactive baselines that exhibit erratic, irrecoverable motions.

System Efficiency: By structurally decoupling the high-frequency control thread from the VLM, AR-VLA maintains a stable 29ms per-action latency. As shown in Fig. 5, this eliminates inter-chunk latency gaps, resulting in the lowest maximum and average jerk during real-world execution.

History-Awareness: Reactive policies inherently suffer from "temporal amnesia" in long-horizon tasks. We evaluate AR-VLA on two unobservable-state tasks: *PushT2* (Sim) and *Stack3* (Real) (Fig. 7). AR-VLA successfully leverages its long-lived action KV cache to maintain task intent and memory of completed sub-goals, succeeding where stateless models become trapped in oscillatory loops.

TABLE I: **BridgeV2 Pretraining to SIMPLER Simulation Success Rate (%).** Evaluated on the Visual Matching setting. AR-VLA significantly outperforms standard predictive models (OpenVLA, Octo) and identical-backbone chunking models (Pi-0-Fast, Pi-0.5), proving the superiority of AR sequence generation using historical KV caches.

Model	Spoon	Carrot	Block	Eggplant	Average
OpenVLA[7]	0	0	0	4.1	1.0
Octo-Base[9]	12.5	8.3	0	43.1	16.0
Octo-Small[9]	47.2	9.7	4.2	56.9	30.0
SpatialVLA[11]	16.7	25.0	**29.2**	**100.0**	42.7
CogACT [8]	58.3	37.5	20.8	91.7	52.1
Pi-0-Fast*[10]	62.5	29.2	20.8	83.3	49.0
Pi-0.5*[5]	58.3	33.3	16.7	95.8	51.0
AR-VLA (Ours)	**75.0**	**54.2**	20.8	95.8	**61.5**

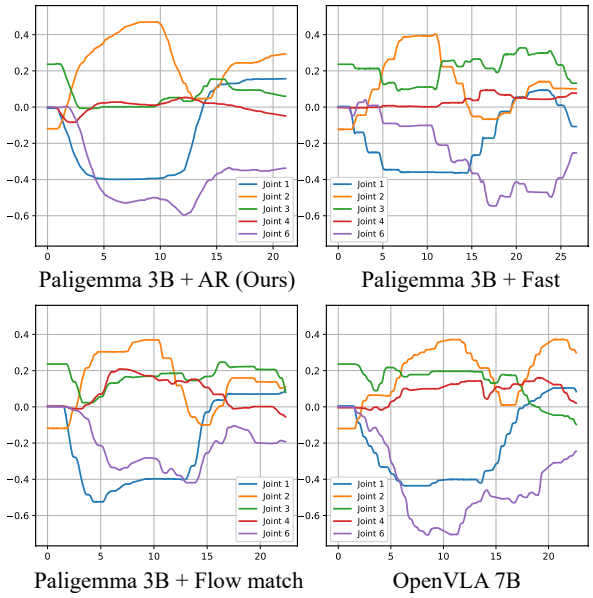


Fig. 5: **Inference Smoothness** The decoupled architecture of AR-VLA achieves a highly efficient 46.25ms total effective latency per action, eliminating the inter-chunk blocking seen in FM and Fast baselines. Jerk is also the smallest, as showed both qualitatively and qualitatively showed as joint angle plot during same task execution.

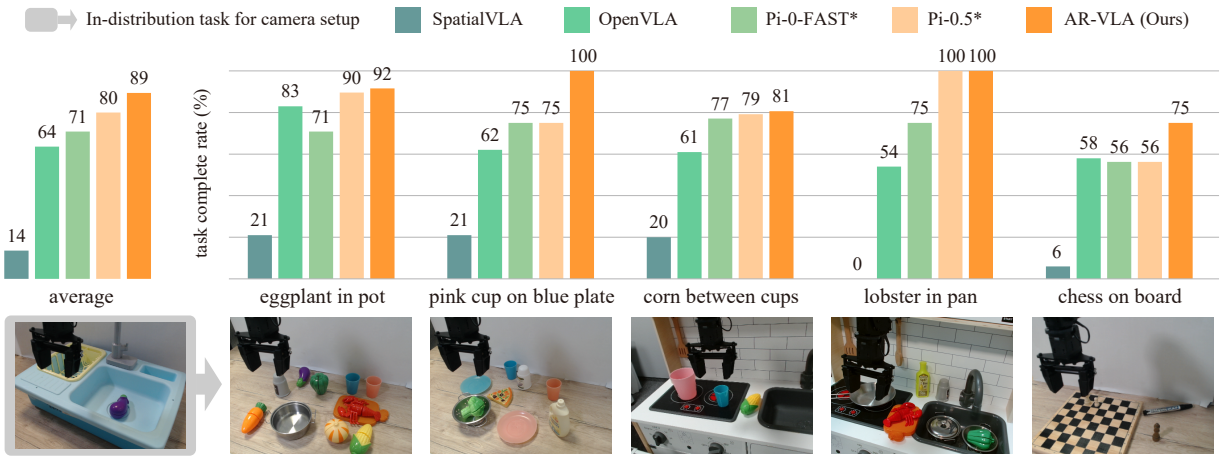


Fig. 6: **Zero-Shot Real-World WidowX Performance.** Models execute at 5Hz with the VLM refreshed every 4 actions. Camera poses were calibrated to ensure 100% baseline success on an in-distribution task (eggplant in sink), followed by testing on challenging layout variants (4 layouts, 3 repeats). AR-VLA achieves superior zero-shot transfer and unique error-recovery capabilities.

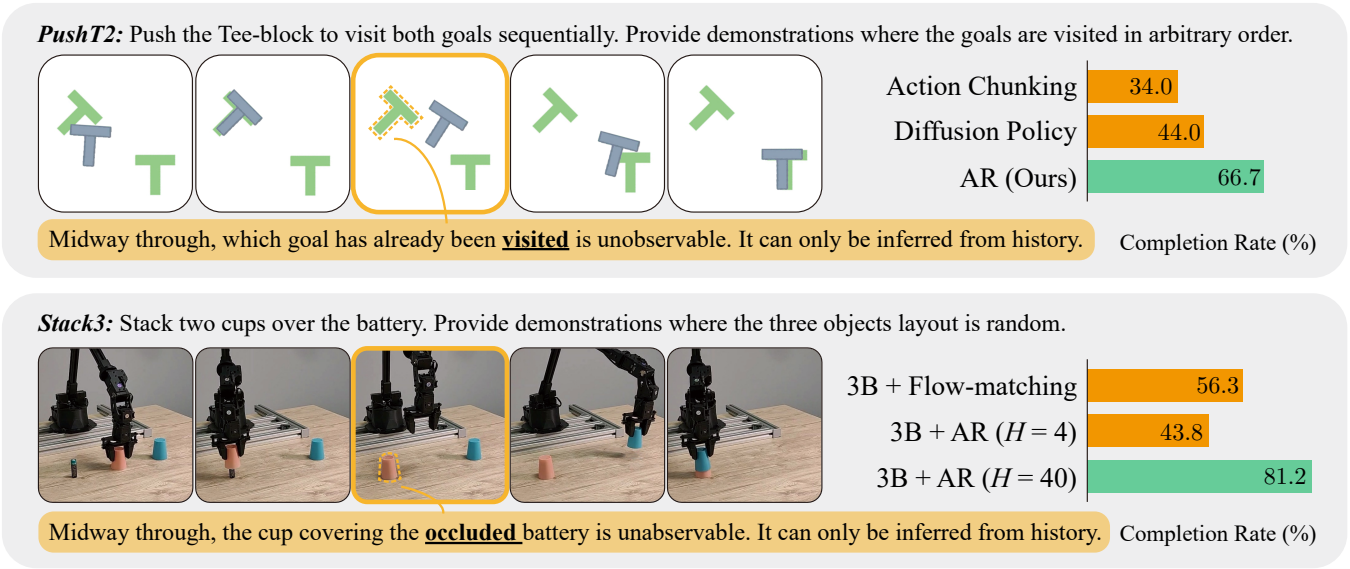


Fig. 7: **History-Awareness Evaluation.** PushT2 requires visiting both goals, but which goal has been visited is unobservable midway. Stack3 requires stacking cups over a battery that becomes occluded. Both task require memory of unobservable past states. H donates the context window length of AR-VLA. Details about task defination, data collection, training and execution in Appendix.

Acknowledgments

This research was partially funded by the Ministry of Education and Science of Bulgaria (support for INSAIT, part of the Bulgarian National Roadmap for Research Infrastructure), and Interne Fondsen KU Leuven/Internal Funds KU Leuven (C2E/24/034).

References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances, August 2022.

[2] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie

Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165.

[4] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023.

[5] 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. $\pi_{0.5}$: A Vision-Language-Action Model with Open-World Generalization, April 2025.

[6] Daniel Kahneman. Thinking, fast and slow. *Farrar, Straus and Giroux*, 2011.

[7] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. In *Conference on Robot Learning*, 2024.

[8] Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, Xiaofan Wang, Bei Liu, Jianlong Fu, Jianmin Bao, Dong Chen, Yuanchun Shi, Jiaolong Yang, and Baining Guo. CogACT: A Foundational Vision-Language-Action Model for Synergizing Cognition and Action in Robotic Manipulation, November 2024.

[9] Octo Model Team, Dibya Ghosh, Homer Walke bit, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. In *Conference on Robot Learning*, 2024.

[10] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025.

[11] Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, and Xuelong Li. SpatialVLA: Exploring Spatial Representations for Visual-Language-Action Model, May 2025.

[12] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware, April 2023.

[13] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.