HiF-VLA: Hindsight, Insight and Foresight through Motion Representation for Vision-Language-Action Models

Supplementary Material

1. More Implementation Details

Beyond the SigLIP [22] and DINOv2 [13] image encoders and the Prismatic VLM [8] backbone described in the main text, we provide further additional implementation details for the two core modules used in HiF-VLA: the Hindsight Encoder and the Hindsight-Modulated Joint Expert. These details complement the high-level description given in the main text.

1.1. Hindsight Encoder

We employ a 4-layer Vision Transformer (ViT) [5] as the hindsight motion encoder. The hindsight motion sequence of length h is first partitioned into $(h//2) \times (H//2) \times (W//2)$ spatiotemporal blocks using a 3D convolution, which simultaneously reduces temporal redundancy and preserves local motion continuity. These blocks, together with an additional $[CLS]$ token that aggregates global temporal context, are fed into the Transformer. The resulting historical features are subsequently projected via a linear layer into the joint expert embedding space, ensuring dimensional compatibility with the foresight and action pathways. This design allows the hindsight encoder to provide structured historical priors that can effectively regularize downstream decision-making.

1.2. Hindsight-Modulated Joint Expert

The hindsight-modulated joint expert serves as the fusion module that integrates hindsight, foresight, and action representations. All tokens are projected into a shared embedding dimension of 1024, including: hindsight motion tokens from the hindsight encoder, foresight motion tokens and latent action tokens produced by the VLM backbone. Within this space, the hindsight tokens serve as adaptive conditioning signals that modulate both foresight and action streams via AdaLN [21] scaling and shifting. This conditioning mechanism allows the model to dynamically adjust its predictions based on temporally grounded historical cues, enabling causal alignment between past observations and future behavior. Positional information is provided through Rotary Positional Embedding (RoPE) [15], allowing efficient encoding of both spatial and temporal ordering. The joint expert consists of 6 Transformer layers, each capable of cross-stream attention, enabling the model to capture complementary relationships between predicted dynamics and actions.

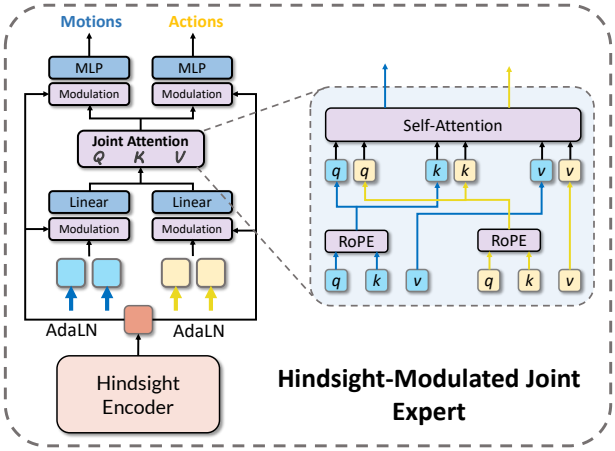


Figure 1. Architecture of the hindsight-modulated joint expert.

2. Comparison with Video-Generation VLAs

Compared to VLA approaches that rely on video generation [4, 6, 19, 20], our method differs fundamentally in how it models temporal dynamics. A large body of recent work [4, 6, 19, 20] employs general-purpose video generative models to predict future frames, using these predictions either for inverse dynamics computation or as conditional representations to assist action policy generation. While these approaches have made substantial progress, they face some limitations when contrasted with HiF-VLA, which leverages sparse motion representations.

Specifically, video-generation-based methods typically require multi-frame historical input and high-resolution future video synthesis, resulting in substantial computational overhead. Moreover, pixel-level prediction is prone to local artifacts and distortions, introducing uncontrollable noise. In contrast, HiF-VLA replaces video-level temporal modeling with low-dimensional, structured motion representations, enabling a more efficient and stable way to encode historical visual states and predict future dynamics. In addition, HiF-VLA, by jointly predicting foresight motion and action trajectories as two complementary information streams, allows the model to anticipate how the physical world may evolve while generating the corresponding actions—effectively enabling thinking while acting.

Methods	LIBERO-Spatial	LIBERO-Object	LIBERO-Goal	LIBERO-Long	Average
TraceVLA [26]	84.6	85.2	75.1	54.1	74.8
Octo [16]	78.9	85.7	84.6	51.1	75.1
CoT-VLA [25]	81.1	87.5	91.6	87.6	69.0
SpatialVLA [10]	88.2	89.9	78.6	55.5	78.1
ThinkAct [7]	88.3	91.4	87.1	70.9	84.4
Seer [18]	-	-	-	87.7	87.7
FlowVLA [27]	93.2	95.0	91.6	72.6	88.1
4D-VLA [23]	88.9	95.2	90.9	79.1	88.6
DreamVLA [24]	97.5	94.0	89.5	89.5	92.6
CogACT [11]	97.2	08.0	90.2	88.8	93.2
π_0 [2]	96.8	98.8	95.8	85.2	94.2
GR00T N1 [1]	94.4	97.6	93.0	90.6	93.9
UniVLA [3]	96.5	96.8	95.6	92.0	95.2
MemoryVLA [14]	98.4	98.4	96.4	93.4	96.5
OpenVLA-OFT [9]	97.6	98.4	**97.9**	94.5	97.1
HiF-VLA(ours)	**98.8**	**99.4**	97.4	**96.4**	**98.0**

Table 1. Performance on the LIBERO benchmark. The table compares our method against a wide range of state-of-the-art approaches, with the best performance highlighted in **bold**.

	Ablation	Parameters	SR
(a)	Foresight Motion Loss Weight λ	0.1	94.4
		0.05	95.2
		0.01*	**96.4**
		0.001	95.6
(b)	Joint Expert Depth	2	95.2
		4	95.6
		6*	**96.4**
		8	95.2
(c)	Length: (Hindsight, Foresight)	(8,8)*	**96.4**
		(8,16)	94.6
		(16,16)	95.2

Table 2. Ablation on hyper-parameters. * denotes the submission setting. (multi-view)

3. More Experimental Results

3.1. Comprehensive Evaluation on the LIBERO Benchmark

We report detailed evaluation results on all four suites of the LIBERO benchmark [12] and compare our method against a broad set of baseline models, as summarized in Tab. 1. While achieving its greatest margin of superiority under the most challenging LIBERO-Long suite, HiF-VLA also delivers competitive or superior performance across the remaining three suites when compared to prior state-of-the-art approaches. As a result, HiF-VLA attains the best average performance over all four suites, demonstrating its robustness and general applicability across diverse task scenarios.

3.2. More Hyper-parameters Ablation

The Impact of Weight Factor λ. The hyperparameter λ balances the contributions between foresight motion prediction and action prediction. To systematically investigate this trade-off, we evaluate a range of λ values. Our analysis reveals that an appropriate weighting allows the motion prediction to effectively support the model's reasoning, thereby enhancing planning capabilities; however, an unsuitable value destabilizes the VLA architecture and impedes policy generation. Our results in Tab. 2(a) identify an optimal performance point at $\lambda = 0.01$. This key finding indicates that a modest, well-calibrated contribution from the motion-prediction branch is essential for achieving balanced and robust model behavior.

Depth of the Joint Expert. We investigated the optimal interaction depth within the Joint Expert by performing an ablation study over depths of $\{2, 4, 6, 8\}$ (see Tab. 2(b)). We observed that a depth of 6 yields the best performance. Shallower networks facilitate the interaction between motion and action modalities, whereas increasing the depth further yields negligible performance gains, indicating depth is not highly critical beyond moderate depth.

Foresight and Foresight Horizon. In Tab. 2 bottom, increasing foresight to $n = 16$ hurts performance due to error accumulation in long-term prediction. In contrast, extending hindsight to 16 (keeping foresight at 16) improves results by providing richer context, supporting our choice to decouple hindsight and foresight lengths.

	Ablation	Variant	SR
(a)	Motion or State as (Hindsight+Foresight)	S+M	92.6
		S+S	92.0
		M+M*	**94.4**
(b)	Causal or Bidirectional	Causal-[M|A]	87.4
		Bi-[A|M]	94.0
		Bi-[M|A]*	**94.4**
(c)	Motion Representation	Flow	94.2
		MVs	**94.4**

Table 3. Ablation on different variants. M: Motion, S: State, A: Action. * denotes the submission setting. (third view)

3.3. More Variants Ablation

State replaces motion. To better understand the role of motion representations, we conduct an ablation study replacing codec motion vectors (**M**) with robot proprioceptive states (**S**). This experiment is motivated by the observation that motion vectors might appear to primarily reflect robot arm movement. If this were the case, robot states—which explicitly encode the robot configuration—should provide similar information and lead to comparable performance. However, motion vectors describe inter-frame visual changes and therefore capture not only robot motion but also the visual consequences of interactions with the environment, such as object displacement, contact-induced motion, or changes in scene structure. These effects are not fully observable from proprioceptive signals alone. As shown in Tab. 3(a), replacing motion vectors with robot states (**M+M** $\rightarrow$ **S+S**) leads to a clear performance drop. This result indicates that motion vectors provide complementary dynamic information beyond robot states and play an important role in capturing interaction-driven visual dynamics.

Bidirectional Interaction vs. Causal Separation. To evaluate the benefit of joint foresight–action modeling, we compare our design with a decoupled variant that removes the bidirectional interaction between motion and action tokens. In this variant, motion and action tokens are processed using causal attention and decoded with separate prediction heads. This modification preserves parallel computation but prevents motion and action representations from interacting during inference. In contrast, our Joint Expert employs bidirectional attention to model motion and action synchronously, allowing the two representations to influence each other during the forward pass. As shown in Tab. 3(b), the decoupled causal variant achieves a success rate of 87.4%, whereas our joint formulation improves performance to 94.4%. This result suggests that enabling bidirectional interaction between foresight motion and action representations is important for effective decision making.

We further investigate whether the ordering of motion and action tokens affects performance. Since bidirectional attention provides a global receptive field within each segment, the token order should not influence the learned representation. Consistent with this expectation, swapping the token order (M|A $\rightarrow$ A|M) results in less than 0.5% performance change.

Different Motion Representations. We further study the impact of different motion representations by replacing motion vectors (MVs) with optical flow, as reported in Tab. 3(c). Optical flow provides dense motion estimation and serves as a commonly used alternative motion cue. Specifically, optical flow is estimated using RAFT [17]. The results show that both representations achieve nearly identical success rates, indicating that the HiF architecture is robust to the specific choice of motion representation. This suggests that the performance gains primarily stem from the proposed modeling framework rather than a particular motion signal. However, the two representations differ significantly in computational cost. Computing optical flow introduces substantial preprocessing overhead. With a 4-frame history, flow-based inference requires 186.8 ms, whereas our MV-based approach takes only 121.6 ms. The efficiency gap becomes larger when the history length increases, resulting in up to a 78% reduction in latency overhead with MVs for 8-frame histories. These results indicate that motion vectors provide a more efficient motion representation while maintaining comparable performance.

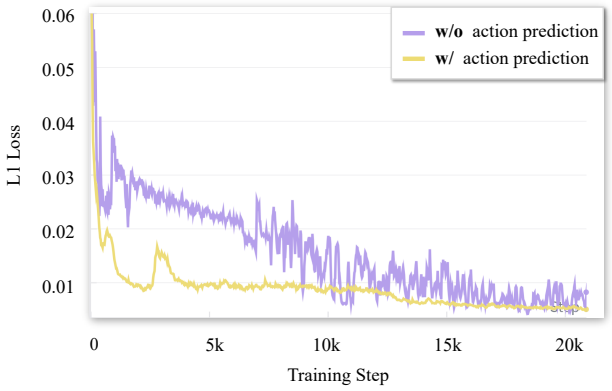


Figure 2. Convergence of the foresight-motion L1 loss during training. The curve "w/o action prediction" corresponds to a variant where the action-prediction branch is removed and only the foresight-motion pathway is trained. The curve "w/ action prediction" represents the full HiF-VLA architecture, where both streams in the joint expert (foresight motion and action) are retained.

3.4. Analysis of "Think-while-Acting" Paradigm

Foresight Motion Prediction. To more comprehensively evaluate the interaction between foresight motion and action prediction, we additionally removed the action-prediction branch and trained the model using only the motion prediction pathway. We then compared the evolution of the motion L1 loss during training, as shown in Fig. 2. The results indicate that incorporating action prediction leads to faster convergence and a noticeably more stable training curve for motion prediction. This indicates that the action branch provides effective complementary information to the motion branch, enabling a synergistic interaction between the two. Such coupling supports the model's ability to reason about future dynamics while simultaneously making action decisions—thereby achieving a genuine "think-while-acting" capability.

Foresight and Action Prediction Visualization. We present qualitative visualizations of the foresight motion predictions and action predictions across several tasks in the LIBERO-Long dataset, as shown in Fig. 3. The visualizations demonstrate how our model's anticipated motion sequences closely align with its generated action plans. Across diverse long-horizon tasks, this close alignment ensures that the agent's actions are consistently grounded in a coherent forecast of future states. This visually corroborates that HiF-VLA maintains temporal coherence and successfully executes the proposed "think-while-acting" paradigm by dynamically adapting its policy based on foresighted reasoning.

4. Real-World Experiments

4.1. Real-World Experimental Setup

We evaluate our method on a series of long-horizon real-world tasks using an AgileX Piper robot, which is equipped with a 6-DoF manipulator and a 1-DoF gripper. A single Intel RealSense D435 camera provides third-person observations, while an additional USB wrist-mounted camera provides egocentric input. Data are collected at 20 Hz. All models are trained under the standard HiF-VLA configuration and deployed on a single NVIDIA RTX 4090 GPU.

4.2. Task Descriptions

Place blocks on the plates. The robot must place the white block onto the white tray and the pink block onto the pink tray. This task primarily evaluates the model's basic visual recognition ability and precise object placement. Its clear goal structure and low action complexity make it suitable for assessing whether the model can reliably establish accurate observation–action mappings in simple environments.

Cover block and stack bowls. The robot first covers a white block with a green bowl and then stacks a pink bowl on top of the green one. This task stresses the model's ability in multi-step sequential reasoning, spatial relation understanding, and long-horizon dependency modeling. The explicit hierarchical dependency structure (cover $\rightarrow$ stack) makes it an effective test for the model's capacity to handle layered object interactions and long-term constraints.

Press buttons in order. The robot must press three colored buttons in a specified order. This task assesses the model's ability to distinguish visually similar states, maintain correct action ordering, and perform time-sensitive decision making. Because pre-press and post-press observations can look visually similar, the task naturally introduces visual ambiguity and demands strong historical information integration and temporal reasoning.

4.3. Real-World Execution Visualization

Fig. 4 provides visualizations of HiF-VLA's rollout across the real-world tasks. The figure presents the model's generated actions at key time steps. As shown, HiF-VLA demonstrates stable behavior planning capabilities in real operation: in the block-grasping task, the robot consistently maintains reliable object recognition and performs precise grasp-and-place actions; in the multi-step bowl covering and stacking task, the system robustly executes cross-object sequential operations and preserves coherent action structure throughout stages with clear hierarchical dependencies; in the button-pressing task, the model correctly distinguishes between visually similar states and adheres to the required action order. These visualizations illustrate that HiF-VLA exhibits robust visual understanding, coherent action organization, and strong execution of long-horizon task structures in real scenarios, thereby validating its reliability in performing complex manipulation tasks.

4.4. Failure Cases

Despite HiF-VLA outperforming in both simulation benchmarks and real-world experiments, several failure cases still occur, as shown in Fig. 5. In the first task, the model prematurely opened the gripper based on an incorrect spatial judgment, resulting in a placement failure. In the second task, the stacking failed because the robotic arm did not lift the bowl to an appropriate height. In the third task, an erroneous depth estimation caused the gripper to descend insufficiently, leading to an incomplete button press.

These failure modes highlight the critical role of spatial geometry and 3D perception. They suggest that future work may benefit from integrating richer 3D representations into our framework to further enhance robustness in real-world manipulation.

References

[1] Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open

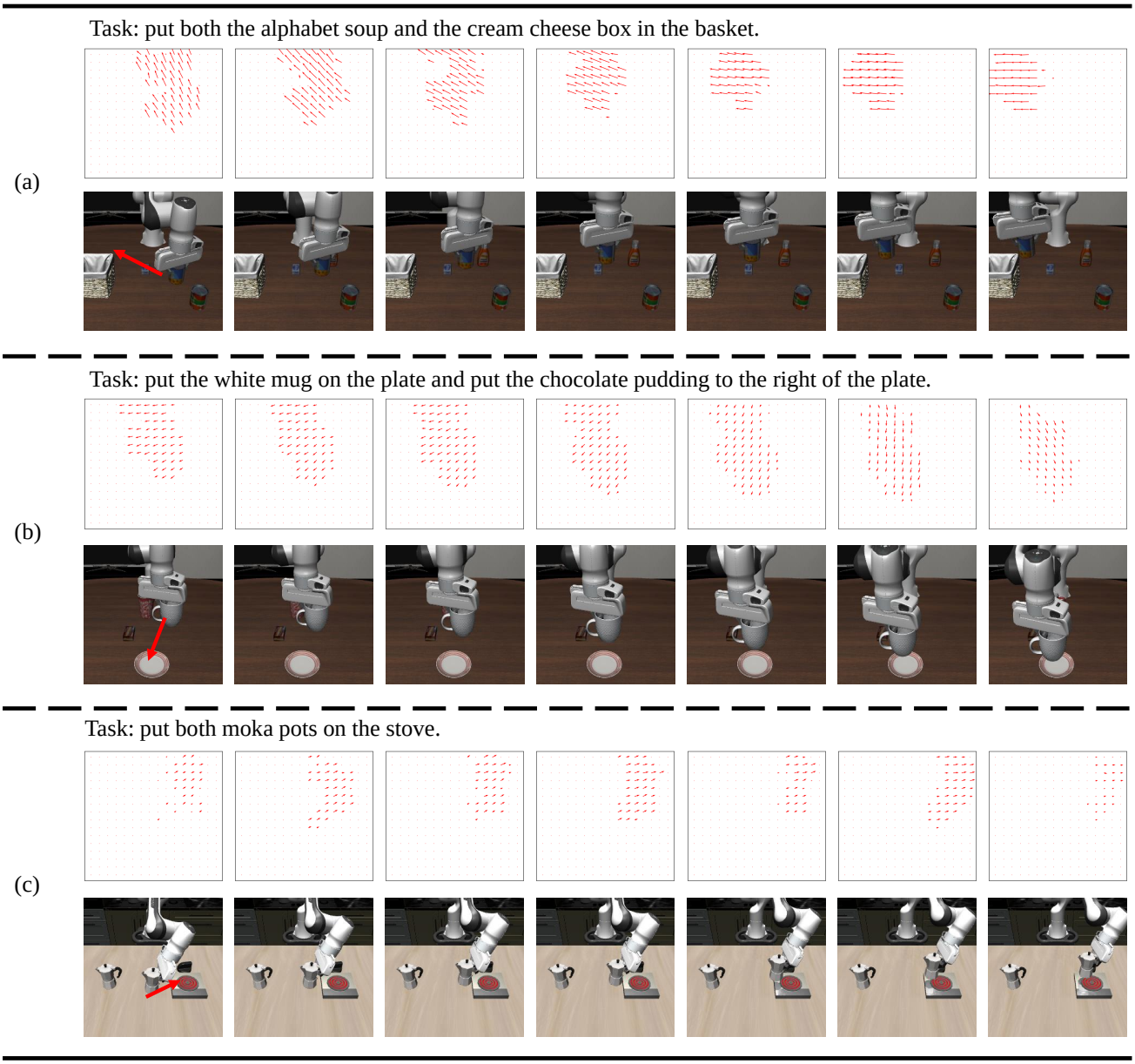


Figure 3. Example rollouts of three tasks in LIBERO-Long, illustrating the close alignment between the predicted foresight motion and the observed action execution over a short inference window.

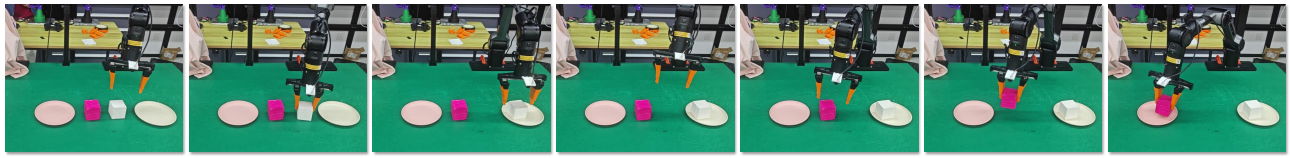
foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025. 2

[2] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. 2

[3] Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. UniVLA: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025. 2

[4] Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, et al. GR-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024. 1

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on*

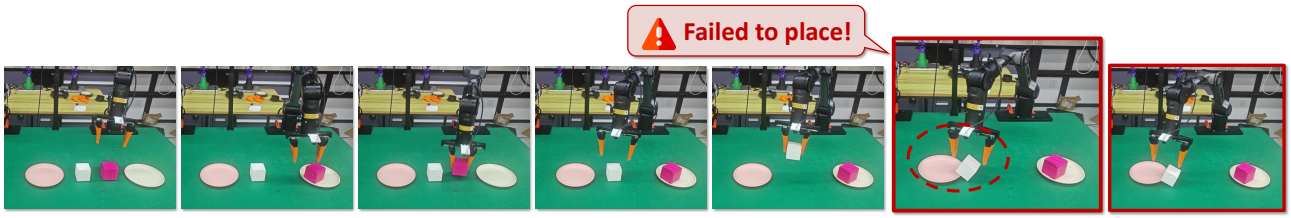


(a) Place blocks on the plates.

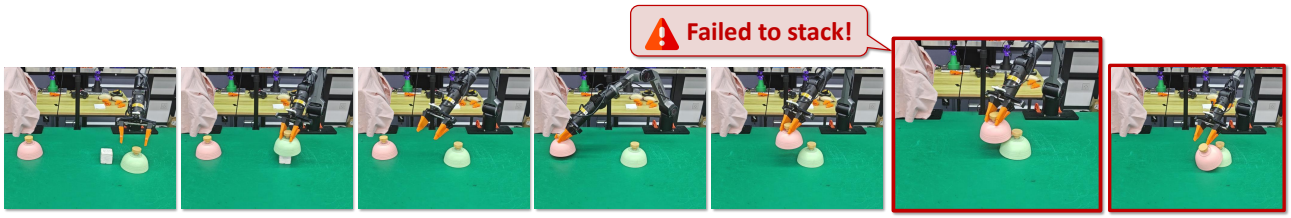
(b) Cover block and stack bowls.

(c) Press buttons in order.

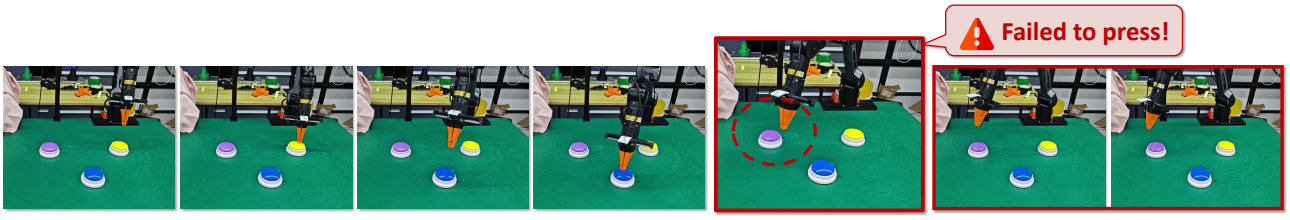
Figure 4. Example rollouts of real-world tasks.



(a) Place blocks on the plates.



(b) Cover block and stack bowls.



(c) Press buttons in order.

Figure 5. Failure cases of real-world tasks.

Learning Representations, 2021. 1

[6] Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. In *Proceedings of the International Conference on Machine Learn-*

ing, 2025. 1

[7] Chi-Pin Huang, Yueh-Hua Wu, Min-Hung Chen, Yu-Chiang Frank Wang, and Fu-En Yang. ThinkAct: Vision-language-action reasoning via reinforced visual latent planning. In *Proceedings of the Advances in Neural Information Processing Systems*, 2025. 2

[8] Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic VLMs: Investigating the design space of visually-conditioned language models. In *Proceedings of the International Conference on Machine Learning*, 2024. 1

[9] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025. 2

[10] Fuhao Li, Wenxuan Song, Han Zhao, Jingbo Wang, Pengxiang Ding, Donglin Wang, Long Zeng, and Haoang Li. Spatial Forcing: Implicit spatial representation alignment for vision-language-action model. *arXiv preprint arXiv:2510.12276*, 2025. 2

[11] Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, et al. CogACT: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. *arXiv preprint arXiv:2411.19650*, 2024. 2

[12] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 44776–44791, 2023. 2

[13] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 1

[14] Hao Shi, Bin Xie, Yingfei Liu, Lin Sun, Fengrong Liu, Tiancai Wang, Erjin Zhou, Haoqiang Fan, Xiangyu Zhang, and Gao Huang. MemoryVLA: Perceptual-cognitive memory in vision-language-action models for robotic manipulation. *arXiv preprint arXiv:2508.19236*, 2025. 2

[15] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 1

[16] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024. 2

[17] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Proceedings of the European Conference on Computer Vision*, pages 402–419. Springer, 2020. 3

[18] Yang Tian, Sizhe Yang, Jia Zeng, Ping Wang, Dahua Lin, Hao Dong, and Jiangmiao Pang. Predictive inverse dynamics models are scalable learners for robotic manipulation. In *Proceedings of the International Conference on Learning Representations*, 2025. 2

[19] Youpeng Wen, Junfan Lin, Yi Zhu, Jianhua Han, Hang Xu, Shen Zhao, and Xiaodan Liang. VidMan: Exploiting implicit dynamics from video diffusion model for effective robot manipulation. *Proceedings of the Advances in Neural Information Processing Systems*, 37:41051–41075, 2024. 1

[20] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. *arXiv preprint arXiv:2312.13139*, 2023. 1

[21] Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. Understanding and improving layer normalization. *Advances in neural information processing systems*, 32, 2019. 1

[22] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 1

[23] Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Yanpeng Zhou, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, et al. 4d-vla: Spatiotemporal vision-language-action pretraining with cross-scene calibration. In *Proceedings of the Advances in Neural Information Processing Systems*. 2

[24] Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, He Wang, Zhizheng Zhang, et al. DreamVLA: a vision-language-action model dreamed with comprehensive world knowledge. *arXiv preprint arXiv:2507.04447*, 2025. 2

[25] 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. 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*, pages 1702–1713, 2025. 2

[26] Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé III, Andrey Kolobov, Furong Huang, and Jianwei Yang. TraceVLA: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. In *Proceedings of the International Conference on Learning Representations*, 2025. 2

[27] Zhide Zhong, Haodong Yan, Junfeng Li, Xiangchen Liu, Xin Gong, Wenxuan Song, Jiayi Chen, and Haoang Li. FlowVLA: Thinking in motion with a visual chain of thought. *arXiv preprint arXiv:2508.18269*, 2025. 2