FlowVLA: Thinking in Motion with a Visual Chain of Thought

Zhide Zhong[1], **Haodong Yan**[1], **Junfeng Li**[1], **Xiangchen Liu**[1], **Xin Gong**[1], **Wenxuan Song**[1], **Jiayi Chen**[1], **and Haoang Li** [1]

[1]HKUST(GZ)

Many Vision-Language-Action (VLA) models rely on an internal world model trained via next-frame prediction. This approach, however, struggles with physical reasoning as it entangles static appearance with dynamic motion, often resulting in implausible visual forecasts and inefficient policy learning. To address these limitations, we introduce the Visual Chain of Thought (Visual CoT): a pre-training framework that encourages a model to reason about how a scene evolves before predicting what it will look like. We instantiate this principle in FlowVLA, which predicts a future frame (v_{t+1}) only after generating an intermediate optical flow representation (f_t) that encodes motion dynamics. This "$v_t \rightarrow f_t \rightarrow v_{t+1}$" reasoning process is implemented within a single autoregressive Transformer, guiding the model to learn disentangled dynamics. As a result, FlowVLA produces coherent visual predictions and facilitates more efficient policy learning. Experiments on challenging robotics manipulation benchmarks demonstrate state-of-the-art performance with substantially improved sample efficiency, pointing toward a more principled foundation for world modeling. Project page: https://irpn-lab.github.io/FlowVLA/

Keywords: Visual Chain of Thought, World Models, Vision-Language-Action Models

1. Introduction

Recent advances in Vision-Language-Action (VLA) models Kim et al. (2024), Zitkovich et al. (2023), Black et al. (2024), Team et al. (2024), particularly those pre-trained as world models like UniVLA Wang et al. (2025) and WorldVLA Cen et al. (2025), have shown remarkable promise for creating generalist robots. The prevailing strategy involves training a large autoregressive transformer to predict the next visual frame given past observations, effectively learning the dynamics of the environment from vast amounts of video data. This learned world model then serves as a powerful foundation for fine-tuning downstream action policies.

Despite their success, these models suffer from a critical, foundational flaw: they conflate the task of physical reasoning with simple pixel prediction. This next-frame prediction paradigm is often a "pixel-copying trap", where the model learns to replicate static backgrounds without a deep understanding of spatiotemporal dynamics, leading to blurry, inconsistent, and physically implausible long-horizon forecasts Ming et al. (2024). Furthermore, this approach creates a significant domain gap between the passive, observational knowledge learned during pre-training and the active, control-oriented knowledge required for policy learning. This results in inefficient knowledge transfer and requires extensive fine-tuning, as evidenced by slow convergence on downstream tasks Zeng et al. (2024).

We argue that the key to learning **physically-grounded** dynamics is to force the model to reason explicitly. Drawing inspiration from the success of Chain of Thought (CoT) prompting in Large Language Models Wei et al. (2022), which enhances reasoning by generating intermediate steps, we propose a novel counterpart for world models: a **Visual Chain of Thought (Visual CoT)**. Instead of a single, opaque leap from the

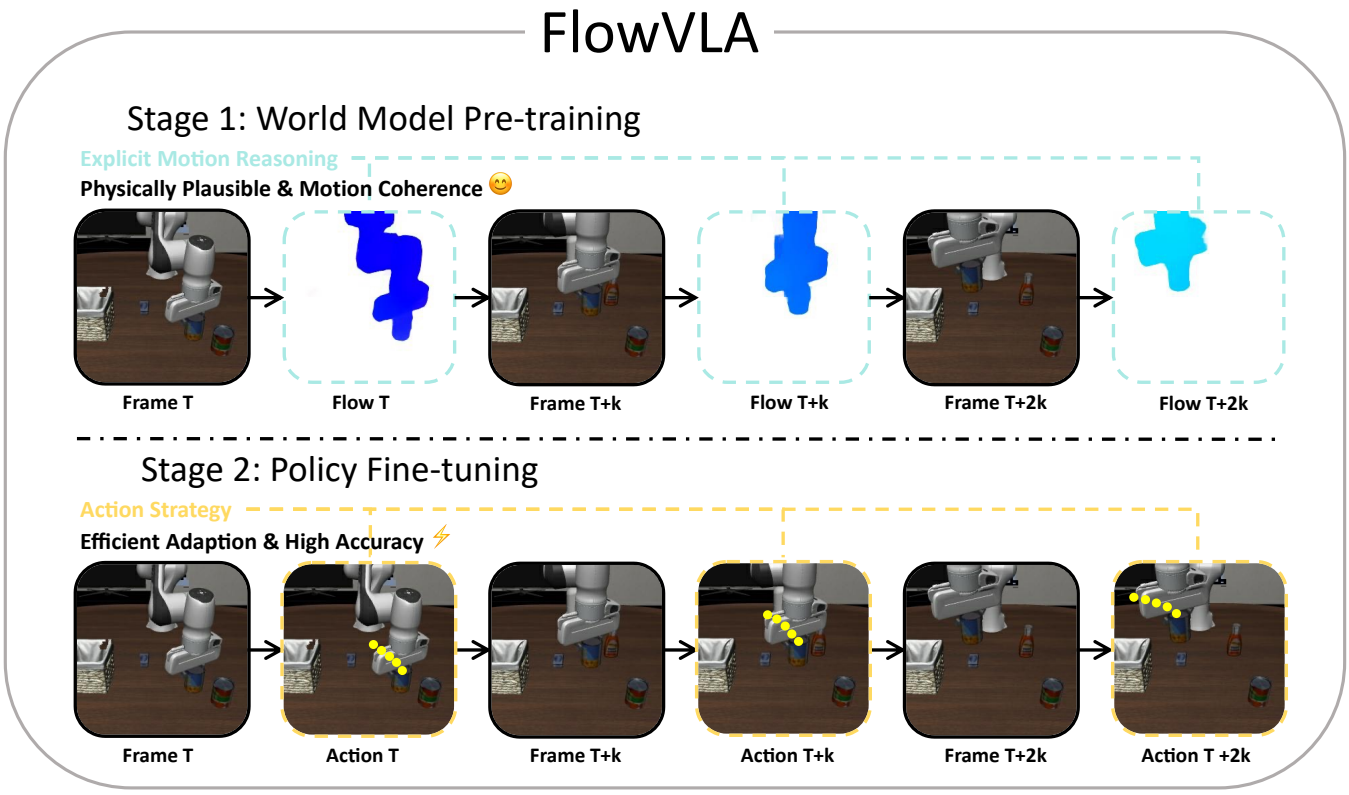


Figure 1: The FlowVLA Two-Stage Training Paradigm. (Top) Stage 1: World Model Pre-training with Visual CoT. The model learns to predict an intermediate motion representation (Flow T) from an initial frame (Frame T), and then uses both to forecast the subsequent frame (Frame T+k). This iterative process yields physically plausible, long-horizon video predictions. **(Bottom) Stage 2: Policy Fine-tuning.** The pre-trained world model is adapted for control, fine-tuned to generate precise robot actions (Action T) from visual observations. This paradigm leverages the learned dynamics for efficient and accurate policy learning.

current frame v_t to the next v_{t+1}, we decompose the prediction into a structured reasoning process: first, predict the intermediate physical dynamics—the optical flow f_t that describes *how* every pixel will move. Only then, conditioned on this explicit motion plan, predict the resulting future frame. This $v_t \rightarrow f_t \rightarrow v_{t+1}$ causal chain (see Figure 1 for an overview) transforms the learning objective from mere pattern recognition into a structured physical reasoning task.

We introduce FlowVLA as the first concrete realization of this principle. A key aspect of our design is to integrate motion without introducing dedicated architectural components. We encode optical flow fields as standard RGB-like images, allowing them to be processed by the exact same VQ tokenizer as regular camera observations. This enables a single, unmodified autoregressive transformer to seamlessly learn the interleaved sequence of appearance and motion tokens. This design makes FlowVLA a truly Unified Visual CoT, where the reasoning steps (flow) and states (frames) are expressed in a shared vocabulary and processed by a single, unified model.

Our work makes the following contributions:

- We identify a fundamental limitation of next-frame prediction for VLA world models and propose Visual Chain of Thought (Visual CoT) as a new principle for learning dynamics.
- We introduce FlowVLA, a simple yet effective instantiation that unifies appearance and motion reasoning

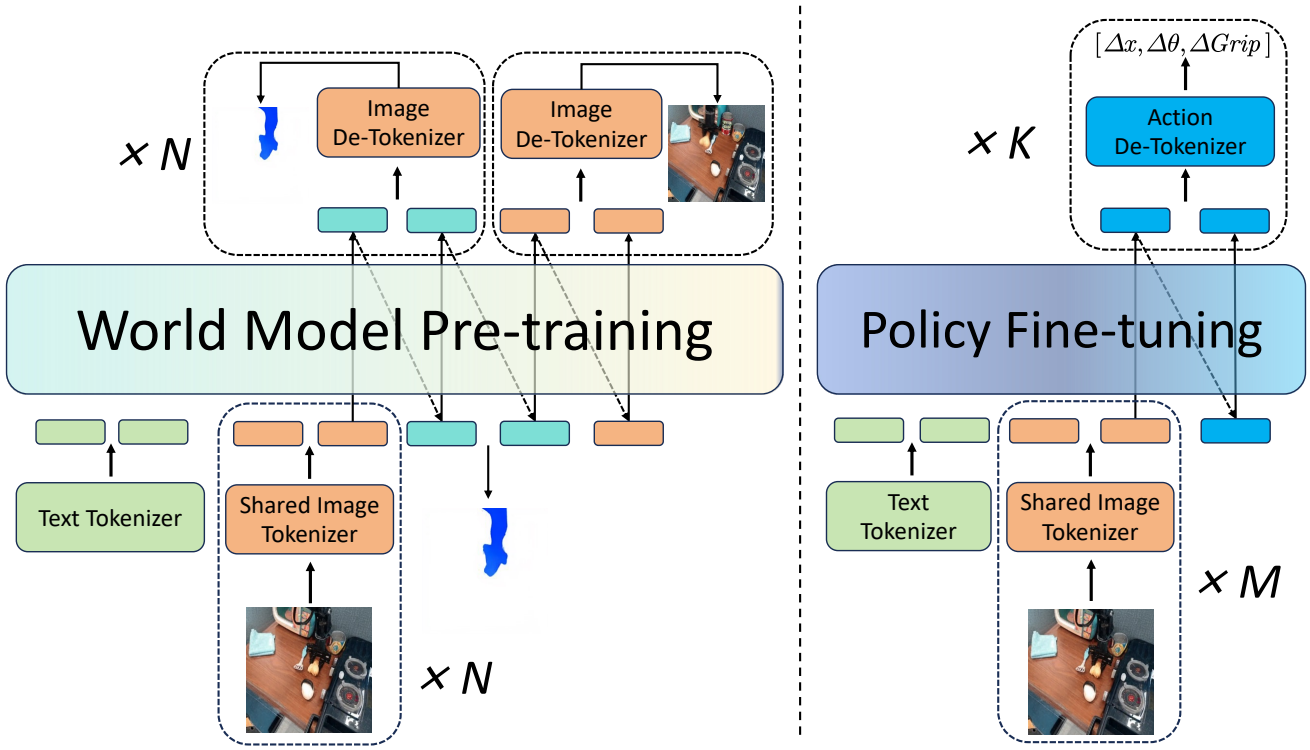


Figure 2: The FlowVLA Framework Architecture. Our model follows a two-stage paradigm. **(Left) Stage 1: World Model Pre-training with Visual CoT.** Input frames are encoded into appearance tokens (orange). The model then autoregressively predicts an interleaved sequence of motion tokens (cyan, representing flow) and future appearance tokens. This structured $v_t \rightarrow f_t \rightarrow v_{t+1}$ prediction forces the model to reason about dynamics before forecasting the future. **(Right) Stage 2: Policy Fine-tuning.** The pre-trained world model is adapted for control. Conditioned on a text instruction (green) and the current observation (orange), the model autoregressively predicts action tokens (blue) that are decoded into robot commands.

within a single autoregressive Transformer via shared tokenization.

- We demonstrate through extensive experiments that FlowVLA achieves state-of-the-art performance on challenging manipulation benchmarks, while offering superior sample efficiency, supporting our claim that explicit motion reasoning better bridges the gap between pre-training and policy fine-tuning.

2. FlowVLA

In this section, we introduce FlowVLA, a novel framework designed to instantiate our proposed Visual Chain of Thought (Visual CoT) principle for world model pre-training. We first provide a high-level overview of our two-stage training paradigm. We then detail the core of our contribution: the Visual CoT pre-training stage, including our unified tokenization scheme for appearance and motion. Finally, we describe how the learned world model is finetuned for downstream robotics tasks.

2.1. Framework Overview

FlowVLA follows a two-stage training paradigm, consistent with state-of-the-art methods like UniVLA Wang et al. (2025) and WorldVLA Cen et al. (2025) to ensure a fair basis for comparison.

1. Stage 1: World Model Pre-training: The model learns general physical dynamics from large-scale, action-free video data by executing our proposed Visual Chain of Thought.
2. Stage 2: Policy Finetuning: The pre-trained model weights are finetuned on downstream, action-annotated robotics datasets to learn specific control policies.

The primary contribution of this paper is concentrated in Stage 1, where we redefine the objective of world model pre-training from simple next-frame prediction to structured physical reasoning.

2.2. Stage 1: World Model Pre-training via Visual Chain of Thought

The goal of this stage is to learn a robust world model by compelling it to reason about dynamics before predicting future states. This is achieved through our Visual Chain of Thought (Visual CoT) pre-training task. Below, we detail the tokenization scheme that unifies appearance and motion, and then describe the autoregressive objective used to learn the reasoning chain.

Unified Motion and Appearance Tokenization A key challenge is to represent two distinct physical signals, appearance (images) and motion (optical flow), within a single model. Our solution is a unified tokenization scheme that preserves architectural simplicity. To process both appearance and motion, we first represent the 2-channel optical flow fields as standard RGB images. Following the technique from VideoJAM Chefer et al. (2025), for each pixel's flow vector (u, v), we convert it to a color representation based on its polar coordinates. The direction of motion is mapped to the color's Hue, calculated from the angle $\alpha = \arctan 2(v, u)$. The speed of motion is mapped to the color's Saturation and Value (Brightness), derived from the vector's magnitude $m = \sqrt{u^2 + v^2}$. To ensure that subtle movements are not lost while large motions do not saturate the representation, the magnitude is normalized to the range $[0, 1]$ using a scaling coefficient $\sigma = 0.15$:

$$m_{\text{norm}} = \min\left(1.0, \frac{m}{\sigma \cdot \sqrt{H^2 + W^2}}\right), \tag{1}$$

where H and W are the height and width of the frame.

Crucially, these resulting flow images, along with the original RGB frames, are processed by the **exact same pre-trained VQ-GAN tokenizer** Esser et al. (2021). This approach discretizes both modalities into token sequences from a shared vocabulary, offering three key advantages: 1) **Parameter Efficiency**, as no new motion-specific tokenizer is needed; 2) **Architectural Simplicity**, maintaining a single, end-to-end autoregressive pipeline; and 3) **Unified Representation**, enabling the model to learn correlations between appearance and motion in a shared latent space.

Autoregressive Learning of the Visual CoT With a unified token representation for both frames (v_t) and flow (f_t), we construct a reasoning chain $v_t \to f_t \to v_{t+1}$. We employ a standard decoder-only Transformer, training it to predict an interleaved sequence of frames and optical flow fields given an optional language instruction L_{instr}:

$$S_{\text{wm}} = \{L_{\text{instr}}, v_0, f_0, v_1, f_1, \ldots, v_T, f_T\} \tag{2}$$

The model is trained using a standard next-token prediction objective, maximizing the log-likelihood of the sequence. The loss of the world model, $\mathcal{L}_{\text{WM}}$, is the sum of the cross-entropy losses in both the reasoning step (flow tokens) and the final state (next frame tokens). Formally, for each timestep t, the model first

predicts the flow f_t based on all preceding tokens, and then predicts the next frame v_{t+1} conditioned on both the history and the just-predicted flow:

$$\mathcal{L}_{\text{WM}} = \sum_{t=0}^{T-1} \left(\mathcal{L}_{\text{CE}}(f_t | S_{<v_{t+1}}) + \lambda \cdot \mathcal{L}_{\text{CE}}(v_{t+1} | S_{<v_{t+1}}, f_t) \right) \tag{3}$$

where $S_{<v_{t+1}}$ denotes all the tokens preceding v_{t+1}, and λ is a balancing hyperparameter (set to 1.0 in our experiments). This objective explicitly forces the model to perform a "reason → predict" process during both training and inference.

2.3. Stage 2: Finetuning for Robotic Control

Initialization and Task. The policy model is initialized with the weights from the pre-trained FlowVLA. During this stage, the input sequence is composed of interleaved observations and actions: $S_{\text{policy}} = \{L_{\text{instr}}, v_0, a_0, v_1, a_1, \dots\}$, where a_t represents the robot's action tokens.

Action Tokenization and Objective. Actions are discretized into tokens following the FAST Pertsch et al. (2025). Critically, the fine-tuning loss, $\mathcal{L}_{\text{policy}}$, is computed **only** over the action tokens. This objective directs the model to leverage all its learned visual and dynamical knowledge towards the singular goal of making correct action decisions.

3. Experiments

We conduct a comprehensive set of experiments to validate the effectiveness of our proposed Visual Chain of Thought framework. Our evaluation is designed to answer four key questions:

- **Q1:** Does FlowVLA achieve state-of-the-art performance on complex, long-horizon robotics tasks?
- **Q2:** Does explicit motion reasoning lead to superior world modeling capabilities?
- **Q3:** Is FlowVLA more sample-efficient during policy finetuning, validating our claim of bridging the pre-training/finetuning gap?
- **Q4:** Which components of our design are most critical to its success?

3.1. Experimental Setup

Benchmarks. To comprehensively evaluate FlowVLA's generalization capabilities, we test our method on two distinct and challenging benchmarks: LIBERO and SimplerEnv.

LIBERO Liu et al. (2023) serves as our primary benchmark for evaluating generalization across multiple axes. We follow the standard behavioral cloning setup and report performance on its four main suites, which test generalization to novel spatial layouts, objects, task goals, and long-horizon compositional challenges.

SimplerEnv Li et al. (2024) is used to assess the robustness of our model against significant domain shifts. This benchmark is specifically designed to evaluate policy transfer by introducing diverse variations in lighting, textures, and camera viewpoints, which are more representative of real-world complexity.

Implementation Details. Our FlowVLA model is built on the 8.5B parameter Emu3 Wang et al. (2024) and UniVLA Wang et al. (2025) architecture. Our key change is adding optical flow, pre-computed with

RAFT Teed and Deng (2020), as an additional input to represent motion. We follow the standard training setup for each benchmark: the model for LIBERO is trained only on the LIBERO dataset, while the model for SimplerEnv is trained only on the Bridge V2 dataset Walke et al. (2023). For LIBERO, we pre-train the world model for 5k steps with a batch size of 16, and then fine-tune the policy for 5k steps with a batch size of 96. For the SimplerEnv benchmark, pre-training runs for 12k steps with a batch size of 32 and policy fine-tuning for 20k steps with a batch size of 128.

3.2. Main Results on Robotics Benchmarks (Q1)

To answer Q1, we evaluate the final performance of FlowVLA after policy finetuning on two distinct and challenging benchmarks: LIBERO and SimplerEnv. Our method establishes a new state-of-the-art on both, demonstrating its effectiveness and robustness.

Results on LIBERO. As shown in Table 1, FlowVLA consistently outperforms all prior methods across the four evaluated suites. Notably, the performance gains are most significant on the *Long* horizon tasks. This directly highlights the benefit of learning a world model with a more robust understanding of physical dynamics, as our Visual CoT framework enables better long-term planning and reasoning.

Results on SimplerEnv. We further test our model's robustness on the SimplerEnv benchmark, which introduces significant visual domain shifts. Table 2 shows that FlowVLA achieves a substantial improvement over existing methods. The remarkable success on tasks that were previously challenging for other models (e.g., stacking blocks) validates that our explicit motion reasoning leads to policies that are more resilient to the visual and physical variations found in more realistic environments.

3.3. Analysis of World Modeling Capabilities (Q2)

To demonstrate the superiority of our Visual Chain of Thought (Visual CoT) framework, we conduct a detailed qualitative analysis on the challenging, real-world Bridge V2 dataset. Our investigation reveals two distinct and critical failure modes in the standard next-frame prediction baseline.

First, Figure 3 highlights failures in physical plausibility. In these examples, the baseline model generates physically incoherent rollouts, such as causing the robotic arm to suddenly vanish or producing inconsistent object motion. This indicates a fundamental inability to model the basic physical continuity of a scene. **Second, Figure 4 illustrates a more subtle but equally critical issue: semantic inconsistency.** Here, while the predicted frames from the baseline may appear visually coherent, the depicted action fails to follow the given language command. This reveals a disconnect between language understanding and visual forecasting.

In stark contrast, FlowVLA successfully navigates both challenges across all scenarios. By first reasoning about motion dynamics via optical flow, our model generates predictions that are not only physically plausible but also semantically aligned with the task instructions, showcasing the robustness and generalizability of our approach.

3.4. Convergence Speed and Data Efficiency(Q3)

Figure 5 illustrates FlowVLA's dramatic advantage in training and sample efficiency. In the full-data setting (Figure 5a), FlowVLA proves vastly more sample-efficient, reaching the baseline's peak performance (0.64) with only **one-third of the training steps** (2k vs. 6k) while also achieving a higher final success rate of 0.73.

This efficiency advantage is amplified in the more challenging low-data regime (Figure 5b). Here, the

Table 1: Results on the LIBERO Benchmark. We report the final task success rate (%). We compare FlowVLA against state-of-the-art methods, grouped by their core methodology. Our key comparison is within the **w/ World Model** group, where our Visual CoT pre-training demonstrates superior performance. Notably, FlowVLA achieves this **without relying on massive external datasets**, highlighting the efficiency of our proposed framework.

Model	Large Scale Pretrain	Spatial	Object	Goal	Long	**Avg.**
w/o World Model						
Diffusion Policy Chi et al. (2023)	×	78.3	92.5	68.3	50.5	72.4
Octo Team et al. (2024)	✓	78.9	85.7	84.6	51.1	75.1
OpenVLA Kim et al. (2024)	✓	84.7	88.4	79.2	53.7	76.5
DiT Policy Hou et al. (2025)	✓	84.2	96.3	85.4	63.8	82.4
TraceVLA Zheng et al. (2024)	✓	84.6	85.2	75.1	54.1	74.8
SpatialVLA Qu et al. (2025)	✓	88.2	89.9	78.6	55.5	78.1
pi0-FAST Pertsch et al. (2025)	✓	96.4	96.8	88.6	60.2	85.5
ThinkAct Huang et al. (2025a)	✓	88.3	91.4	87.1	70.9	84.4
w/ World Model						
WorldVLA Cen et al. (2025)	×	85.6	89.0	82.6	59.0	79.1
UniVLA† Wang et al. (2025)	×	92.6	93.8	86.6	63.0	84.0
CoT-VLA Zhao et al. (2025)	✓	87.5	91.6	87.6	69.0	81.1
FlowVLA (ours)	×	**93.2**	**95.0**	**91.6**	**72.6**	**88.1**

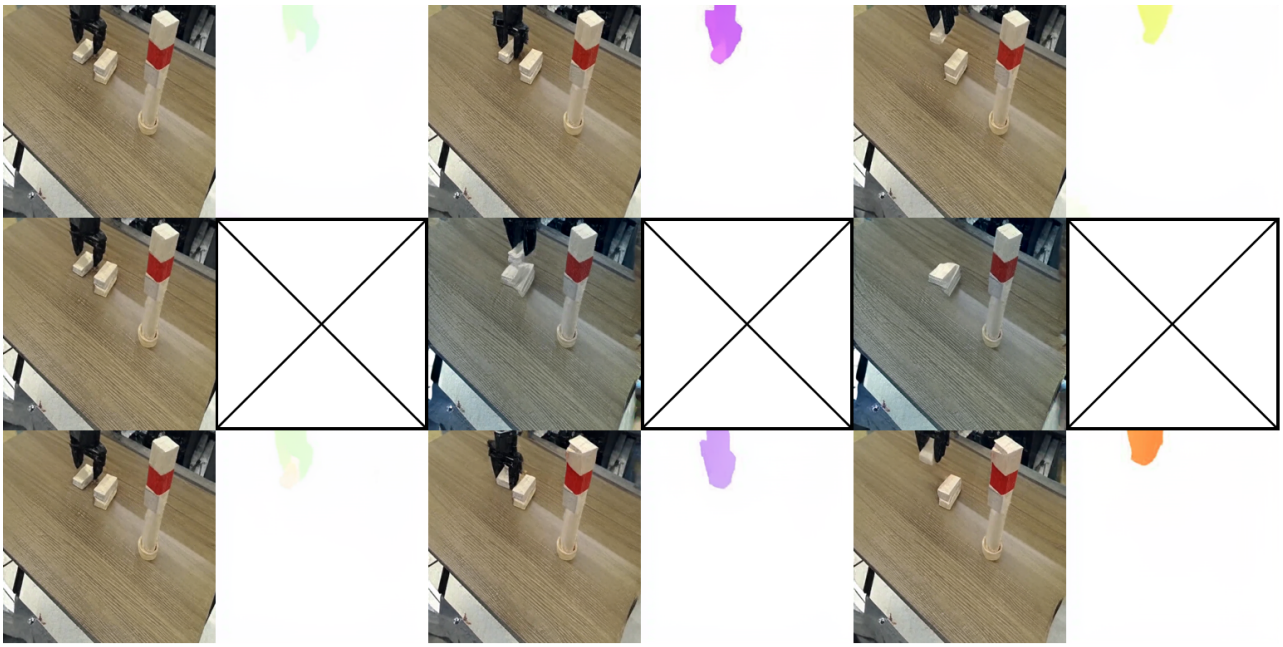
† Our reported UniVLA result is from our re-implementation, pre-trained only on LIBERO without wrist camera images for a fair comparison. The original paper utilized large-scale robotics data and an additional wrist camera.

Table 2: Results on the SimplerEnv-WidowX benchmark. We report the final task success rate (%). FlowVLA significantly surpasses prior methods, demonstrating superior robustness to the visual domain shifts present in this benchmark. Best results are in **bold**.

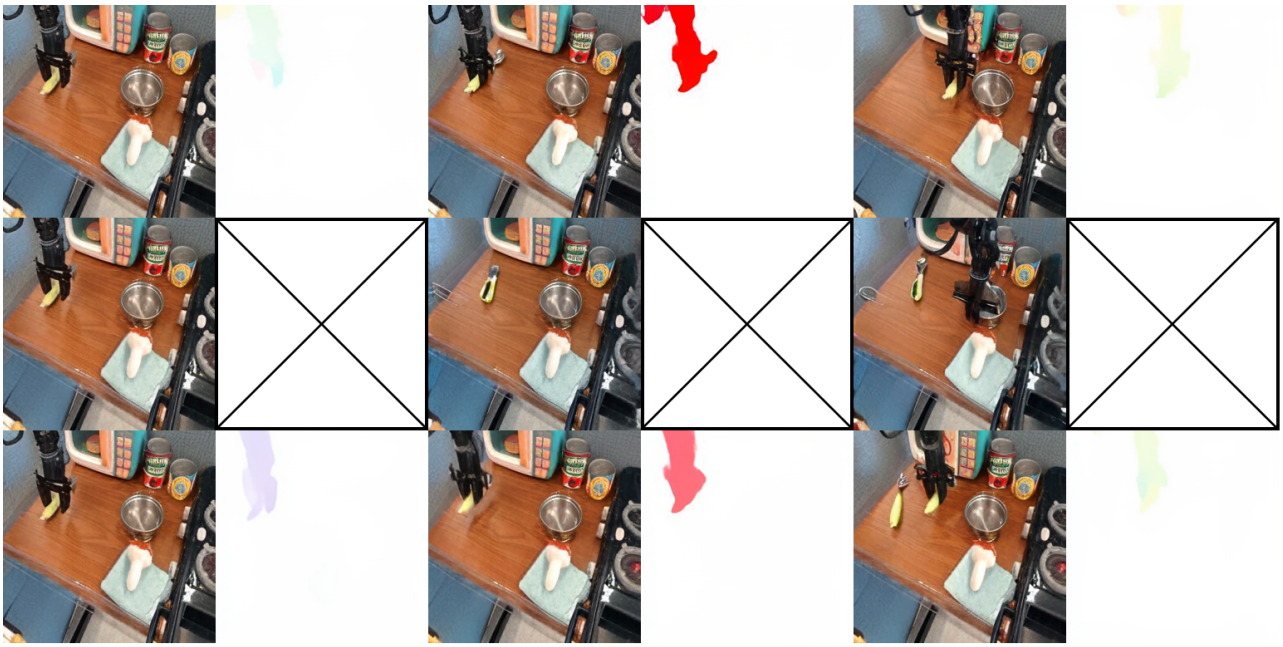
Model	Put Spoon	Put Carrot	Stack Block	Put Eggplant	**Avg.**
RT-1-X Team et al. (2024)	0.0	4.2	0.0	0.0	1.1
Octo-Base Team et al. (2024)	12.5	8.3	0.0	43.1	16.0
Octo-Small Team et al. (2024)	47.2	9.7	4.2	56.9	29.5
OpenVLA Team et al. (2024)	0.0	0.0	0.0	4.1	1.0
RoboVLMs Liu et al. (2025)	45.8	20.8	4.2	79.2	37.5
SpatialVLA Qu et al. (2025)	16.7	25.0	29.2	100	42.7
RoboPoint Yuan et al. (2024)	16.7	20.8	8.3	25.0	17.7
FSD Yuan et al. (2025a)	41.6	50.0	33.3	37.5	40.6
Embodied-R1 Yuan et al. (2025b)	62.5	68.0	36.1	58.3	56.2
UniVLA† Wang et al. (2025)	62.5	62.5	41.6	95.8	65.6
FlowVLA (Ours)	**70.8**	**62.5**	**62.5**	**100.0**	**74.0**

† Result obtained by evaluating the officially released checkpoint.

performance gap widens substantially. FlowVLA not only achieves a **55% higher peak success rate** relative to the baseline (0.48 vs. 0.31) but also surpasses the baseline's peak performance in just 1,000 steps. This



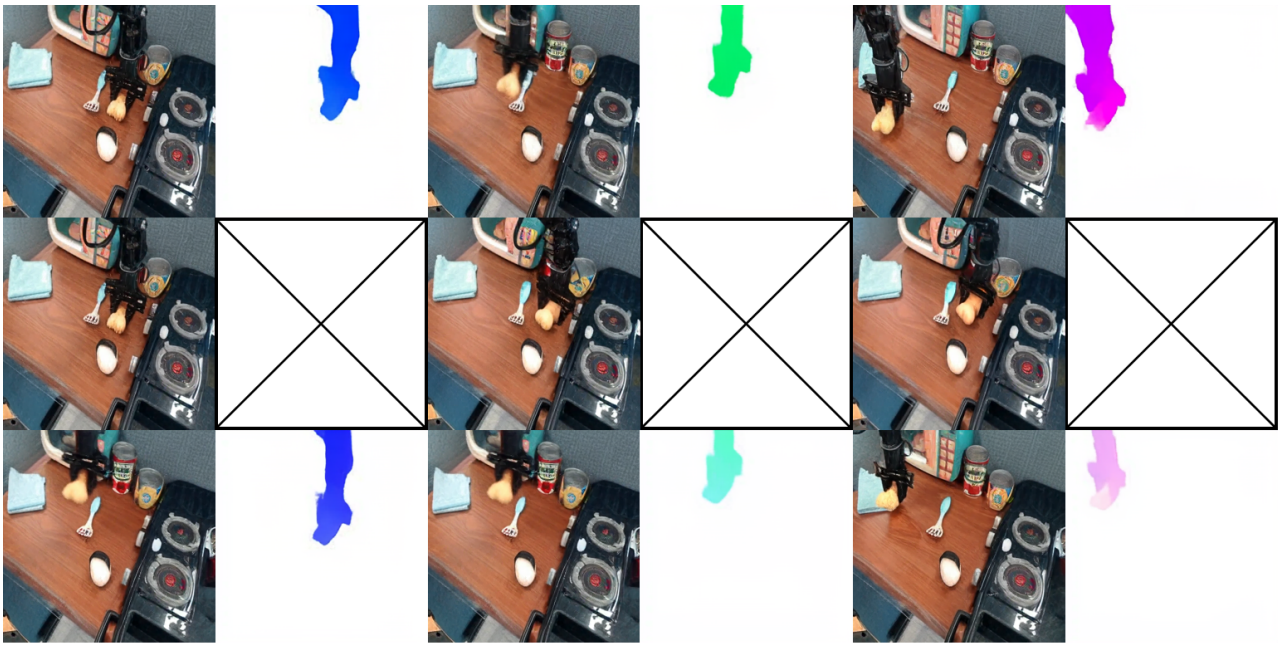
(a) Task: "Put the rectangular on top of the rectangular block next to it."



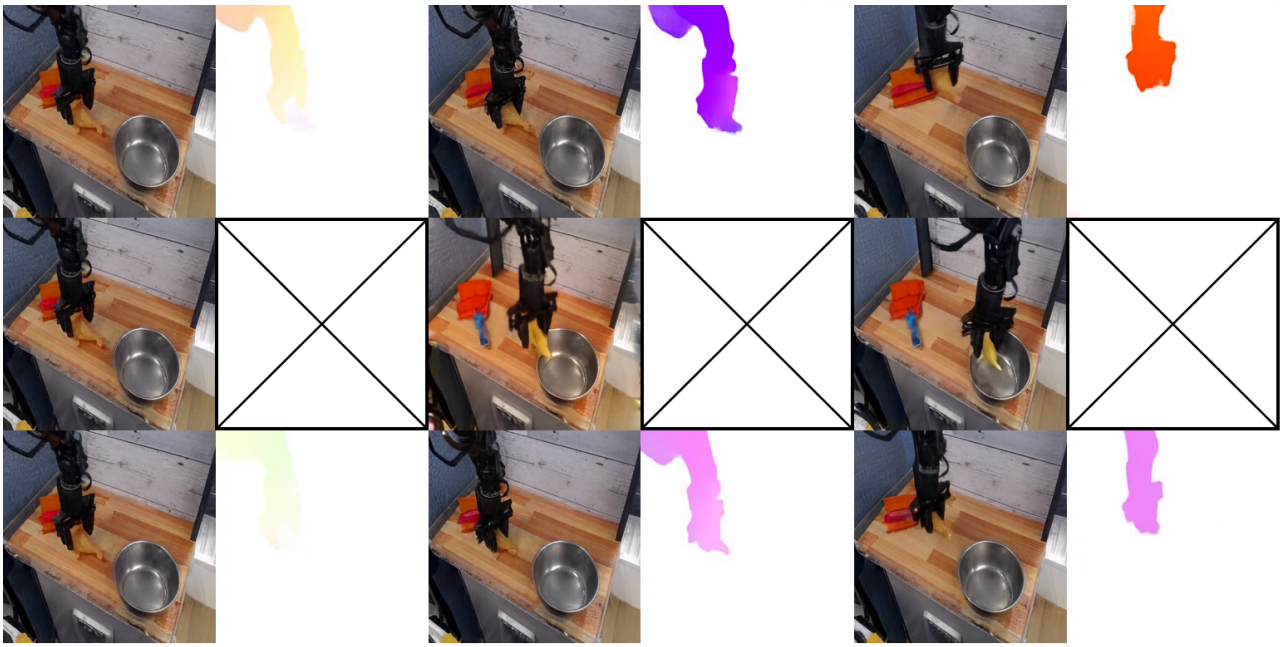
(b) Task: "Move the spoon so that it sits to the left of the metal pot."

Figure 3: Analysis of Physical Plausibility on the Bridge V2 Benchmark. This figure highlights common physical failures in the baseline model. In both examples, the baseline model (middle row) struggles to maintain physical consistency, leading to implausible outcomes such as a disappearing manipulator or erratic object behavior. In contrast, **FlowVLA** (bottom row), guided by its motion-first reasoning, produces stable and physically coherent predictions that accurately reflect the scene's dynamics.

substantial improvement in sample efficiency validates our core hypothesis: by requiring the model to

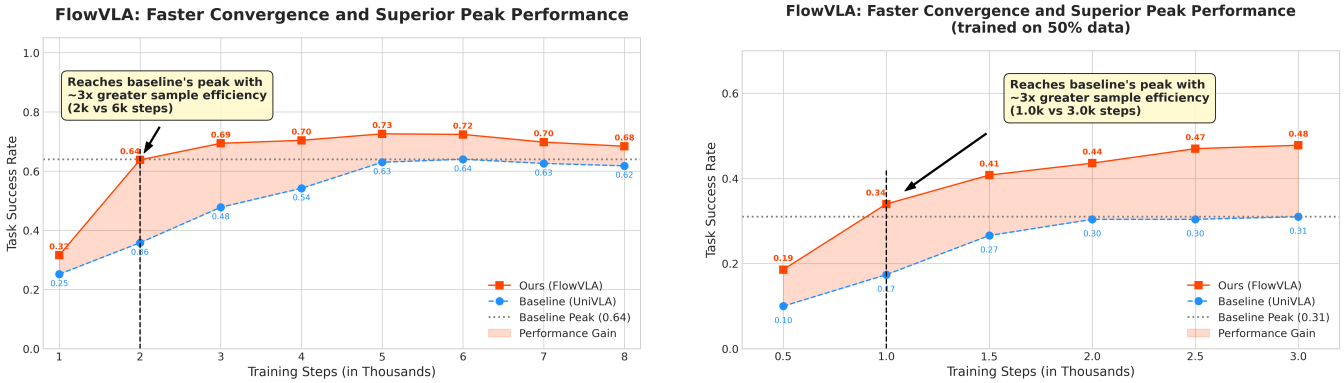


(a) Task: "Put the toy into left of table."



(b) Task: "Move toy diagonally little bit top on the right side."

Figure 4: Analysis of Semantic Alignment on the Bridge V2 Benchmark. This figure illustrates the baseline's failure to align predictions with language instructions. While the predicted frames from baseline model (middle row) might appear visually plausible at a glance, the resulting motion does not correspond to the specified task (e.g., moving an object in the wrong direction). **FlowVLA** (bottom row) again demonstrates superior performance, correctly interpreting the command and generating a corresponding visual trajectory. This underscores that our Visual CoT not only improves physical realism but also enhances the model's ability to ground language in action.



(a) Training on 100% of data

(b) Training on 50% of data

Figure 5: Training Efficiency Comparison in Full and Low-Data Regimes. Success rate versus training steps for FlowVLA and the baseline. Our method converges dramatically faster and reaches a higher peak performance across both the full dataset **(a)** and a data-scarce setting **(b)**. The performance gap widens significantly with limited data, highlighting the superior sample efficiency of our approach.

explicitly reason about motion via a visual chain-of-thought, FlowVLA benefits from a powerful inductive bias. This simplifies the learning of physical dynamics from raw pixels, leading to a more effective and robust learning process, particularly when data is limited.

3.5. Ablation Studies (Q4)

Finally, we conduct a series of ablation studies to understand the contribution of each key component in our framework. The results, summarized in Table 3, are evaluated on the LIBERO-10 benchmark.

We first remove the entire Visual Chain-of-Thought (CoT) structure, which causes our model to degenerate into the UniVLA baseline. As shown in Table 3, the success rate drops sharply from 73.0% to 64.0%. This significant 9-point drop confirms that the explicit, step-by-step reasoning process, where the model first thinks about "how to move" before predicting the outcome, is the primary driver of our model's performance gain.

Next, we investigate the importance of direct supervision for the intermediate reasoning step. In this variant, we retain the interleaved visual-flow sequence structure but remove the optical flow loss during training, meaning the model is not explicitly guided to generate physically correct flows. The performance degrades to 69.5%. This result indicates that while the interleaved architecture provides a useful structural prior, direct supervision is crucial to prevent the model from generating arbitrary or collapsed representations for the intermediate step (f_t). The supervision ensures the "thought" is physically grounded.

Finally, we challenge the core design of interleaving information. We restructure the input sequence by grouping all visual frames first, followed by all corresponding flow frames (i.e., $v_0, v_1, ..., f_0, f_1, ...$). This configuration leads to a severe performance collapse, with the success rate plummeting to 49.4%. This is because the model can no longer leverage the predicted motion f_t to inform the generation of the next state v_{t+1} in a causal, forward-looking manner. This result provides strong evidence that the "interleaved, step-by-step causal chain" ($v_t \rightarrow f_t \rightarrow v_{t+1}$) is essential for effective planning and action generation.

Table 3: Ablation studies on the LIBERO-10 benchmark. We evaluate the importance of our key design choices: the Visual CoT structure, the flow supervision loss, and the interleaved sequence format. The full FlowVLA model is shown for comparison.

Configuration	Success Rate (%)
FlowVLA (Ours, Full Model)	**73.0**
Ablations:	
1. w/o CoT (degenerates to baseline)	64.0
2. w/o Flow Loss	69.5
3. Grouped Sequence	49.4

4. Related Work

4.1. Vision-Language-Action (VLA) Models

The dominant paradigm for creating generalist robot agents is the Vision-Language-Action (VLA) model Zitkovich et al. (2023), Kim et al. (2024), Black et al. (2024), Song et al. (2025a,b). These models extend large, pre-trained Vision-Language Models (VLMs) by fine-tuning them on extensive robotics datasets O'Neill et al. (2024). Architectures like RT-2 Zitkovich et al. (2023) and OpenVLA Kim et al. (2024) treat action generation as a sequence modeling problem, directly mapping visual and textual inputs to discretized action tokens. Other recent works have focused on improving the action representation itself, using techniques like diffusion policies Chi et al. (2023) or flow matching Black et al. (2024). While this end-to-end approach has demonstrated remarkable generalization, it often treats the environment's physical dynamics as a "black box". The policy is learned reactively, without an explicit, underlying model of how the world functions or evolves. FlowVLA diverges from this standard VLA formulation by prioritizing world understanding over immediate action generation. Its pre-training objective is not to learn a policy ($v_t \rightarrow a_t$), but to build a robust world model by learning the physical transition function of the environment ($v_t \rightarrow v_{t+1}$). This "dynamics-first" approach establishes a solid foundation of physical knowledge before it is adapted for downstream control.

4.2. World Models for Robotics

The concept of a world model, which learns a model of its environment to plan or imagine future outcomes Ha and Schmidhuber (2018), is increasingly vital in robotics. Recent works have leveraged this idea for policy learning. For example, some models use video prediction as a form of self-supervised pre-training to improve downstream task performance Wang et al. (2025), Wu et al. (2023). Others, like WorldVLA Cen et al. (2025), propose architectures that jointly learn to predict both the next frame and the next action, creating a tight loop between prediction and control. A common thread in these approaches is the direct prediction of the next frame, modeling the transition as $v_t \rightarrow v_{t+1}$. However, this direct objective forces a single network to simultaneously handle two distinct problems: understanding static scene properties (appearance, texture, lighting) and modeling complex physical dynamics (motion, interaction, causality). This entanglement can result in inefficient learning and physically implausible predictions, such as blurry or distorted futures. In contrast, FlowVLA avoids this entanglement with its Visual Chain of Thought framework. We decompose the prediction into a "frame → flow → frame" reasoning process. By forcing the model to first predict an intermediate optical flow field (f_t), we explicitly decouple the learning of dynamics (*how* things move) from appearance (*what* they look like), resulting in a more causally-grounded world model.

4.3. Embodied Reasoning for Robotics

To move beyond simple reactive policies, a significant line of research has focused on endowing agents with explicit reasoning capabilities. These approaches can be broadly categorized. One category focuses on high-level semantic reasoning, where models generate linguistic or abstract plans. For instance, ECoT Zawalski et al. (2024) and ThinkAct Huang et al. (2025b) leverage Chain-of-Thought prompting to generate textual sub-goals that guide the agent's behavior. A second category focuses on mid-level geometric reasoning, where models produce intermediate spatial representations to guide actions. MolmoAct Lee et al. (2025), for example, generates depth maps and 2D end-effector trajectory traces as part of its "Action Reasoning" pipeline to make planning more concrete. FlowVLA introduces a more fundamental form of reasoning: low-level physical reasoning. Unlike high-level semantic or geometric planning, our Visual CoT operates at the pixel level. By predicting the dense optical flow field, it learns a general, causal model of the world's dynamics, independent of any specific task or action. This provides a foundational understanding of physics that is complementary to, and arguably a prerequisite for, effective high-level control.

5. Conclusion

We proposed the Visual Chain of Thought (Visual CoT) as a new principle for world model learning, instantiated in FlowVLA. By decomposing prediction into a motion–then–appearance process, FlowVLA learns more coherent dynamics within a unified autoregressive model. Experiments on robotics benchmarks demonstrate state-of-the-art performance and improved sample efficiency, highlighting the value of explicit motion reasoning for bridging perception and control.

References

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control, 2024. URL https://arxiv.org/abs/2410.24164.

Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, et al. Worldvla: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025.

Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. VideoJAM: Joint Appearance-Motion Representations for Enhanced Motion Generation in Video Models. *arXiv preprint arXiv:2502.02492*, 2025. URL https://arxiv.org/abs/2502.02492.

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.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018.

Zhi Hou, Tianyi Zhang, Yuwen Xiong, Haonan Duan, Hengjun Pu, Ronglei Tong, Chengyang Zhao, Xizhou Zhu, Yu Qiao, Jifeng Dai, et al. Dita: Scaling diffusion transformer for generalist vision-language-action policy. *arXiv preprint arXiv:2503.19757*, 2025.

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. *arXiv preprint arXiv:2507.16815*, 2025a.

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, 2025b. URL https://arxiv.org/abs/2507.16815.

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. *arXiv preprint arXiv:2406.09246*, 2024.

Jason Lee, Jiafei Duan, Haoquan Fang, Yuquan Deng, Shuo Liu, Boyang Li, Bohan Fang, Jieyu Zhang, Yi Ru Wang, Sangho Lee, Winson Han, Wilbert Pumacay, Angelica Wu, Rose Hendrix, Karen Farley, Eli VanderBilt, Ali Farhadi, Dieter Fox, and Ranjay Krishna. Molmoact: Action reasoning models that can reason in space, 2025. URL https://arxiv.org/abs/2508.07917.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, et al. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36: 44776–44791, 2023.

Huaping Liu, Xinghang Li, Peiyan Li, Minghuan Liu, Dong Wang, Jirong Liu, Bingyi Kang, Xiao Ma, Tao Kong, and Hanbo Zhang. Towards generalist robot policies: What matters in building vision-language-action models. 2025.

Ruibo Ming, Zhewei Huang, Zhuoxuan Ju, Jianming Hu, Lihui Peng, and Shuchang Zhou. A survey on video prediction: From deterministic to generative approaches. *CoRR*, 2024.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024.

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.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025.

Wenxuan Song, Jiayi Chen, Pengxiang Ding, Han Zhao, Wei Zhao, Zhide Zhong, Zongyuan Ge, Jun Ma, and Haoang Li. Accelerating vision-language-action model integrated with action chunking via parallel decoding, 2025a. URL https://arxiv.org/abs/2503.02310.

Wenxuan Song, Ziyang Zhou, Han Zhao, Jiayi Chen, Pengxiang Ding, Haodong Yan, Yuxin Huang, Feilong Tang, Donglin Wang, and Haoang Li. Reconvla: Reconstructive vision-language-action model as effective robot perceiver. *arXiv preprint arXiv:2508.10333*, 2025b.

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.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020.

Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning*, pages 1723–1736. PMLR, 2023.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need, 2024. URL https://arxiv.org/abs/2409.18869.

Yuqi Wang, Xinghang Li, Wenxuan Wang, Junbo Zhang, Yingyan Li, Yuntao Chen, Xinlong Wang, and Zhaoxiang Zhang. Unified vision-language-action model, 2025. URL https://arxiv.org/abs/2506.19850.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

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.

Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction for robotics. *arXiv preprint arXiv:2406.10721*, 2024.

Yifu Yuan, Haiqin Cui, Yibin Chen, Zibin Dong, Fei Ni, Longxin Kou, Jinyi Liu, Pengyi Li, Yan Zheng, and Jianye Hao. From seeing to doing: Bridging reasoning and decision for robotic manipulation. *arXiv preprint arXiv:2505.08548*, 2025a.

Yifu Yuan, Haiqin Cui, Yaoting Huang, Yibin Chen, Fei Ni, Zibin Dong, Pengyi Li, Yan Zheng, and Jianye Hao. Embodied-r1: Reinforced embodied reasoning for general robotic manipulation, 2025b. URL https://arxiv.org/abs/2508.13998.

Michał Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*, 2024.

Jia Zeng, Qingwen Bu, Bangjun Wang, Wenke Xia, Li Chen, Hao Dong, Haoming Song, Dong Wang, Di Hu, Ping Luo, et al. Learning manipulation by predicting interaction. *arXiv preprint arXiv:2406.00439*, 2024.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1702–1713, 2025.

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. *arXiv preprint arXiv:2412.10345*, 2024.

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.