DISENTANGLED ROBOT LEARNING VIA SEPARATE FORWARD AND INVERSE DYNAMICS PRETRAINING

Anonymous authors
Paper under double-blind review

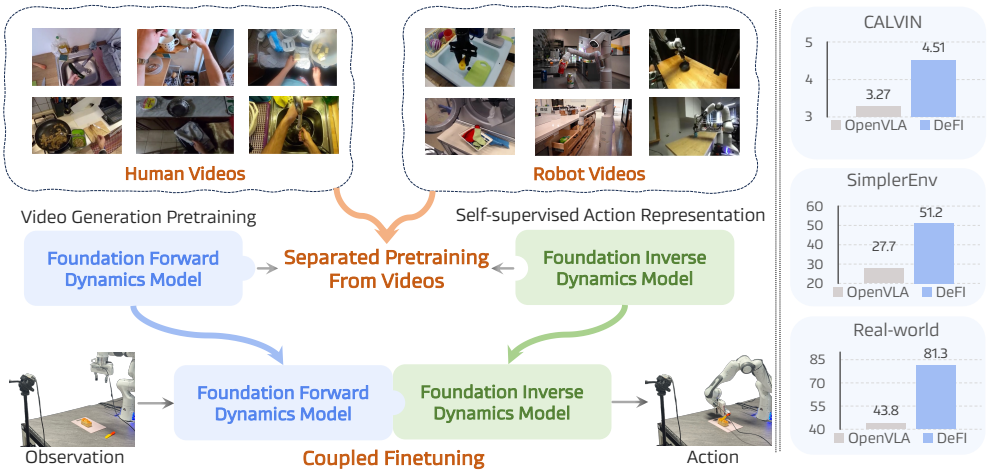


Figure 1: We disentangle robot learning via two decoupled components: a visual forward dynamics model pretrained on large-scale mixed videos via video generation, and an inverse dynamics model pretrained on mixed videos through self-supervised action representation. Then they are coupled during fine-tuning in an end-to-end manner to adapt to downstream tasks. This decoupled pretraining paradigm unleashes the potential of massive action-free videos for policy learning, while retaining robot-specific action grounding, leading to improved success rates across diverse benchmarks.

ABSTRACT

Vision-language-action (VLA) models have shown great potential in building generalist robots, but still face a dilemma–misalignment of 2D image forecasting and 3D action prediction. Besides, such a vision-action entangled training manner limits model learning from large-scale, action-free web video data. To address these issues, we propose **DeFI**, a novel framework that **De**couples visual **F**orward and **I**nverse dynamics pretraining to exploit respective data sources, wherein video generation and action prediction are disentangled. We introduce the Foundation Forward Dynamics Model (FFDM), pretrained on diverse human and robot videos for future prediction, and the Foundation Inverse Dynamics Model (FIDM), trained via self-supervised learning to infer latent actions from unlabeled video transitions. These models are then integrated into a unified architecture for end-to-end fine-tuning on downstream tasks. In this manner, FFDM and FIDM first shine separately and then cooperate for mutual benefit. Extensive experiments on CALVIN ABC-D and SimplerEnv demonstrate state-of-the-art performance, with DeFI achieving an average task length of 4.51 for CALVIN, 51.2% success rate on SimplerEnv-Fractal benchmark and 81.3% success rate in real-world deployment, significantly outperforming prior methods.

1 INTRODUCTION

Vision-language-action (VLA) models (Zitkovich et al., 2023; Kim et al., 2024; Black et al., 2024) have emerged as a promising framework for generalist robots, leveraging the strong visual and language understanding of VLMs (Karamcheti et al., 2024; Beyer et al., 2024) to generate actions

with the supervision of massive action-labeled data. A promising line of work (Tian et al., 2024; Zhao et al., 2025; Zhang et al., 2025c) seeks to integrate visual forecasting with action reasoning into an end-to-end architecture, implicitly learning a coupled representation of forward and inverse dynamics, and presents a more impressive success than conventional VLA. However, this paradigm faces two inherent challenges: (i) the competing objectives of 2D video forecasting and 3D action prediction yield unstable training (Tian et al., 2024); (ii) more critically, they hinder the model from fully exploiting these massive action-free human/web videos. We argue that human videos are indispensable for scaling VLA: they are orders of magnitude larger and more diverse than robot demonstrations, and inherently contain rich motion priors across embodiments and tasks. Unlocking their potential is therefore crucial for building truly generalist and scalable robotic agents.

Alternatively, another strategy attempts to bypass this problem by employing a video prediction model pretrained on human and robot videos for forward dynamics learning (Black et al., 2023; Du et al., 2024; Bu et al., 2024a; Liang et al., 2024; Hu et al., 2024; Feng et al., 2025), followed by a simple model for inverse action inference. This strategy reduces dependence on costly action-labeled data and inherits priors from large video generators trained on large-scale corpora. Yet it often overlooks a critical point: *accurate action inference is as important as accurate future prediction, which still needs sufficient data for pretraining to unleash its full ability*. For instance, VPP (Hu et al., 2024) omits the inverse dynamics component entirely, while Vidar (Feng et al., 2025) includes one but treats it contemptuously, without a scalable pretraining recipe—the performance gain stems largely from a powerful video generator (Bao et al., 2024) rather than principled action reasoning. As a result, the inverse dynamics module becomes the bottleneck, unable to fully exploit the predictive power of the forward model and ultimately limiting overall policy performance.

Considering all the above factors, we explore designing an approach to achieve a win-win effect w.r.t. 2D video forecasting and 3D action prediction. To this end, we propose **DeFI**, a novel paradigm that **disentangles robot learning** by decoupling forward and inverse dynamics knowledge pretraining to leverage distinct data sources, then integrating them into a unified, end-to-end architecture to adapt to downstream tasks. Conceptually, both the forward and inverse dynamics modules are pretrained on mixed human and robot data, yet they extract complementary knowledge: the forward dynamics model focuses on capturing motion-level regularities from 2D video forecasting, while the inverse dynamics model emphasizes 3D action reasoning grounded in state transitions. This first separation enables each module to specialize while still benefiting from heterogeneous data, and the following integration yields a scalable and generalizable policy framework. As shown in Figure 1, we first pretrain a visual *foundation forward dynamics model* (FFDM) built on a video generation model using a mixture of human videos and robot demonstrations. By predicting future video clips conditioned on the current observation and instruction, this FFDM could learn implicit forward dynamics. Crucially, we emphasize *inverse dynamics pretraining is as important as forward dynamics learning*. We therefore introduce a *foundation inverse dynamics model* (FIDM) with a carefully designed self-supervised recipe that scales to action-free human videos. We cast implicit action inference as a self-supervised representation learning problem: a future video reconstruction objective serves as a proxy that compels the model to distill meaningful latent action codes from visual transitions. This formulation unlocks the use of heterogeneous data to learn inverse dynamics at scale, complementing separated forward-dynamics.

During fine-tuning, we couple the pretrained forward and inverse dynamics models into a unified system that supports end-to-end optimization. This design leverages modality-specific strengths while preserving the benefits of end-to-end learning, enabling strong generalization without relying on massive robot-demonstration datasets. Our comprehensive evaluation, spanning CALVIN ABC-D (Mees et al., 2022b) and SimplerEnv-Fractal (Li et al., 2024b) underscores the framework's efficiency, scalability, and generalization, positioning it as a promising pathway toward next-generation generalist robotic policies. Also, multiple ablation studies are conducted to validate that disentangled robot learning via separate forward and inverse dynamics pretraining could fully exploit the prior knowledge of human videos.

In summary, our main contributions are three-fold: **(i)** A decoupled pretraining paradigm that breaks the reliance of conventional end-to-end VLAs on scarce action annotations, enabling us to exploit abundant, easily available unlabeled video data to learn general physical-world dynamics and action representations; **(ii)** We devise a concise architecture that integrates the separately pretrained forward and inverse dynamics models into a single framework. This design fully leverages action-free human video data while enabling end-to-end fine-tuning on downstream tasks with robot action

data **(iii)** DeFI sets a new state of the art on the CALVIN ABC-D benchmark (4.51 average task length), outperforming prior methods by up to 4.2%, and boosts SimplerEnv-Fractal benchmark to 51.2% success rate and real-world experiments to 81.3% success rate. Ablation studies confirm each component's contribution. Furthermore, benefited by pretraining, we only need a few task data to achieve efficient downstream generalization.

2 Related Works

2.1 Vision-Language-Action Models

With the vigorous development of Large Language Models (Liu et al., 2023; Karamcheti et al., 2024; Beyer et al., 2024) and the emergence of large-scale robot datasets (O'Neill et al., 2023; Ebert et al., 2021; Khazatsky et al., 2024; Deng et al., 2025), VLA has become a trend in robot learning. RT series (Brohan et al., 2023; Zitkovich et al., 2023; Belkhale et al., 2024) is the pioneering attempt to fine-tune the MLLM on robot demonstration datasets, resulting in strong accuracy and generalization. Based on this, many studies concentrate on improving the accuracy (Kim et al., 2024; Black et al., 2024; Qu et al., 2025; Liang et al., 2025; Xue et al., 2025) and extend to navigation tasks (Zhang et al., 2024b;a). Additionally, many researchers propose to employ multiple knowledge predictions as a multimodal Chain of Thought (COT) to advance the action reasoning ability of VLA. Concretely, prior efforts take several forms. One line of work first plans high-level subtasks and then outputs low-level actions (Belkhale et al., 2024; Lin et al., 2025). Another uses subgoal images or short visual rollouts that anticipate how the scene should evolve (Tian et al., 2024; Zhao et al., 2025; Cen et al., 2025; Wang et al., 2025). A third condition policies on object-centric signals (e.g., bounding boxes) that capture manipulation-relevant dynamics (Deng et al., 2025; Intelligence et al., 2025a). Others learn latent future embeddings or actions that compactly encode forthcoming motor intentions (Bu et al., 2025a; Lyu et al., 2025; Zhang et al., 2025c). Despite these advances, a central dilemma remains: the misalignment between future-knowledge forecasting and 3D action prediction. Moreover, entangling vision and action during training hampers scaling to action-free web videos. In contrast, DeFI unlocks the potential of large-scale, action-free videos by decoupling the forward and inverse dynamics pretraining, then cooperating in an end-to-end manner for mutual benefit.

2.2 Robot learning from videos

Research that leverages videos for robot learning typically falls into four branches. First, methods that learn from *explicit human hand/motion labels* (e.g., hand pose, keypoints, contact/trajectory annotations) and transfer these priors to manipulation (Bi et al., 2025; Luo et al., 2025; Kareer et al., 2024); such labels provide clean supervision but are expensive to scale and brittle under embodiment or camera shifts. Second, methods that *pretrain the policy on mixed videos* and then fine-tune on downstream tasks (Li et al., 2025; Luo & Lu, 2025; Wu et al., 2024; Cui et al., 2024; Majumdar et al., 2023). These methods solely explore using the implicit forward dynamics knowledge in videos to initialize the weights of VLA. Third, methods that *extract latent actions from human videos* to pretrain large VLA models (Ye et al., 2024; Yang et al., 2025; Bjorck et al., 2025; Chen et al., 2025; 2024b), converting video dynamics into compact tokens to amortize over human-scale data. However, this route is indirect—the latents must be consumed by sizable policy models that are costly to pretrain and fine-tune, and the learned codes are not guaranteed to align with the action manifold needed for execution across embodiments. Fourth, *video-as-policy* approaches pretrain a video or latent feature generator on mixed data to imagine future observations and then train a lightweight controller to track those futures (Wen et al., 2024; Hu et al., 2024; Bharadhwaj et al., 2024; Feng et al., 2025; Collins et al., 2025; Black et al., 2023; Du et al., 2024; Zhang et al., 2025a; Tan et al., 2025; Xie et al., 2025; Wen et al., 2023); while these methods exploit abundant action-free footage to learn forward dynamics, a prediction-to-control gap remains. In contrast, our proposed paradigm treats the inverse dynamics model as equally important, and similarly leverages large-scale action-free video data to train it, thereby completing the transfer from prediction to control.

3 Methodology

As shown in Figure 2, our core idea is to decouple policy learning into two independent knowledge modules: a visual foundation forward dynamics model that predicts instruction-conditioned future

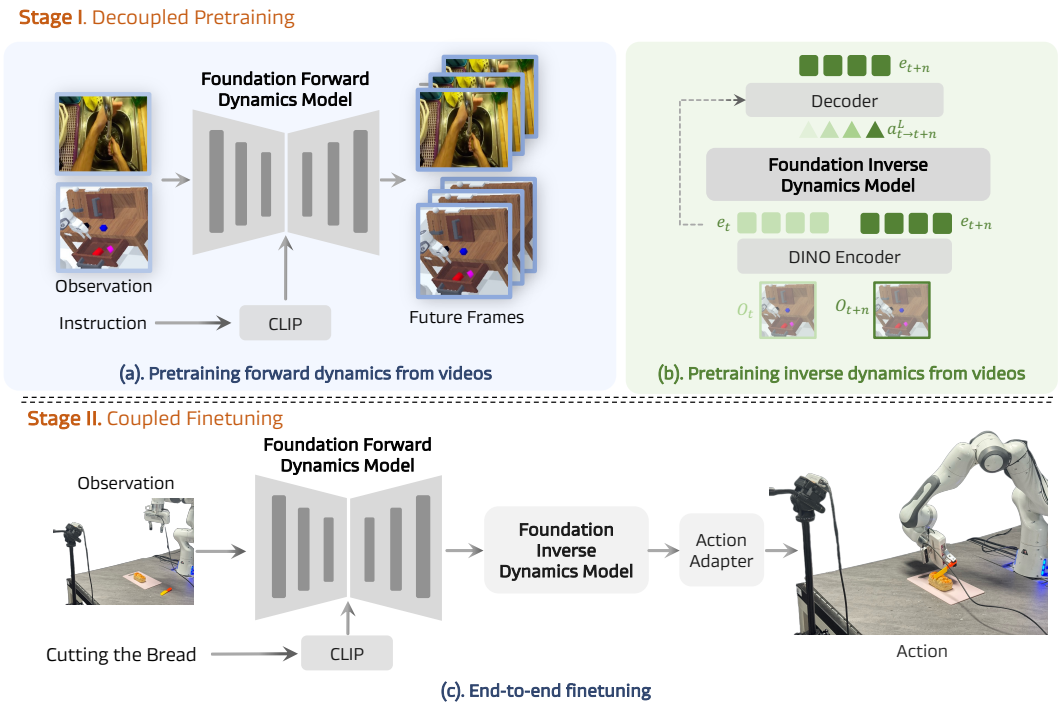


Figure 2: Overall framework of DeFI. **Stage I (Decoupled pretraining):** (a) A visual foundation forward dynamics model is pretrained with human and robot videos via a video generation objective, predicting future frames from current observations and instructions. (b) In parallel, a foundation inverse dynamics model is pretrained in a self-supervised manner to map pairs of observations (o_t, o_{t+n}) into latent actions, capturing inverse dynamics knowledge without explicit action labels. **Stage II (Coupled finetuning):** The forward and inverse models are coupled, and a diffusion-based adapter is used to generate executable robot action sequences. This two-stage framework unleashes the rich priors of human videos while grounding them in robot data for scalable policy learning.

visual states from the current state and instruction in Section 3.1, and a foundation inverse dynamics component that infers the latent actions responsible for observed visual changes in Section 3.2. Each module is pretrained on large, heterogeneous datasets to absorb complementary priors. We then post-train them together to form a complete policy that maps instructions directly to actions, while supporting end-to-end joint fine-tuning with a small amount of robot data in Section 3.3.

3.1 Pretrain FFDM to Learn Forward Dynamics

Given a current observation image o_t and a task instruction l, the objective of the visual forward dynamics model $\mathcal{F}_\theta$ is to synthesize a short-horizon video $\hat{o}_{t:t+H}$ of length $H+1$. We adopt the stable video diffusion (SVD) model with a CLIP text encoder (Radford et al., 2021) and pretrain it on mixed datasets. The model is composed of three components: (i) a video VAE $(\mathcal{E}, \mathcal{D})$ (2D or 3D) that defines the latent space, (ii) a denoiser ϵ_θ (U-Net/Transformer with temporal attention) trained under a latent-diffusion objective. We denote the diffusion time steps by $s \in \{1, \dots, S\}$, distinct from the prediction horizon H. With a variance schedule $\{\beta_s\}_{s=1}^S$, define $\alpha_s = 1 - \beta_s$ and $\bar{\alpha}_s = \prod_{i=1}^s \alpha_i$. The forward (add noise) process over the latent video sequence is as follows:

$$q\Big(z^{(s)}_{t:t+H} \mid z^{(0)}_{t:t+H}\Big) = \mathcal{N}\Big(\sqrt{\bar{\alpha}_s}\, z^{(0)}_{t:t+H},\ (1-\bar{\alpha}_s)\mathbf{I}\Big), \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{1}$$

where the ϵ denotes the Gaussian noise. The conditioning context is formed from the current observation and instruction:

$$c_t = \big(z_t,\ f_{\text{text}}(l)\big), \quad z_t = \mathcal{E}(o_t), \tag{2}$$

where z_t is obtained by encoding the current image. The denoiser is optimized via noise prediction (optionally with v-parameterization):

$$\mathcal{L}_{\text{diff}}(\theta) = \mathbb{E}_{z^{(0)}_{t:t+H},\, s,\, \epsilon} \Big\| \epsilon - \epsilon_\theta\big(z^{(s)}_{t:t+H},\ s,\ c_t\big) \Big\|_2^2. \tag{3}$$

During the inference stage, starting from Gaussian noise, a sampler (e.g., DDIM or DPM-Solver) generates the latent forecast:

$$\hat{z}_{t:t+H} = \mathcal{F}_\theta(z_t, f_{\text{text}}(l)).$$

Nevertheless, fully denoising an entire explicit video remains computationally expensive, as most of the cost is wasted on reconstructing pixel-level details irrelevant to manipulation. In contrast, the key signal for control lies in motion dynamics rather than appearance. Recent research (Zhu et al., 2025; Hu et al., 2024) further suggests that a generative model's features after a single denoising step already contain sufficient motion information to guide downstream action planning. Inspired by this, we freeze the pretrained FFDM and restrict the denoising process to a single step, yielding efficient predictions of future latent embeddings. Notably, for a robot with multiple camera views, such as a third-view and a wrist camera, we predict the future videos for each view independently.

3.2 Pretrain FIDM to Learn Inverse dynamics

To learn the knowledge of inverse dynamics from mixed videos in a fully unsupervised manner, we develop a proxy task to pretrain the foundation inverse dynamics model $\mathcal{I}_\theta$. Specifically, we start with a pair of consecutive video frames o_t, o_{t+n}, separated by a frame interval n, then extract a pair of latent states e_t and e_{t+n} using the DINOv2 (Oquab et al., 2024) visual encoder. We ensure a uniform time interval of approximately 1 second across diverse datasets. The foundation inverse dynamics model consists of an encoder built upon a spatial-temporal Transformer (Xu et al., 2020) with causal temporal masks, and a VQ-VAE codebook that enables vector-quantized action representation. We concatenate a set of learnable action queries $q_a \in \mathbb{R}^{N\times d}$ with predefined dimension d, along the sequence dimension with the DINO embeddings of the current and future frames as well as the instruction embeddings extracted by T5 (Raffel et al., 2020), and feed them into the FIDM:

$$\tilde{a}^L_{t\to t+n} = \mathcal{I}_\theta\big(e_t, e_{t+n}, l, q_a\big),$$

Following LAPA (Ye et al., 2024), we train the model using the VQ-VAE objective (Van Den Oord et al., 2017), which implicitly quantizes the latent actions. The nearest quantized representation is retrieved from a discrete embedding codebook:

$$\hat{a}^L_{t\to t+n} = \mathcal{VQ}_\theta\big(\tilde{a}^L_{t\to t+n}\big).$$

This formulation allows the latent action to be represented as discrete tokens from a vocabulary space $|C|$, making it straightforward for vision-language models to predict actions. The quantized latent action is then passed to a decoder composed of Spatial Transformers, which predicts the DINO features of the future frame. $\hat{e}_{t+n}$ The training objective minimizes the mean-squared error (MSE) loss between the predicted future states $\hat{e}_{t+n}$ and ground-truth states e_{t+n}.

3.3 Finetuning the Coupled FFDM and FIDM in an End-to-end Manner

During finetuning, we keep the FFDM $\mathcal{F}_\theta$ frozen because it was pretrained on large-scale data that already covers the downstream domain; further finetuning on the much smaller downstream split would erode these dynamics priors and hurt generalization. The frozen FFDM then serves as a stable backbone that provides temporally consistent future video representations encoding long-horizon dynamics. A lightweight MLP then projects them onto the input manifold of the FIDM, ensuring representational compatibility between forward prediction and inverse reasoning. The FIDM $\mathcal{I}_\phi$ is then optimized to interpret these aligned latents and infer latent actions that capture the underlying motion. Finally, the diffusion-based action adapter, initialized from a 30M DiT-B, is trained to translate latent actions into executable robot commands. This finetuning stage therefore couples the three modules and allows their objectives—future prediction, action inversion, and low-level control—to be jointly aligned.

3.4 Inference process

At inference time, the FFDM receives the current observation o_t and language instruction l, generates a sequence of predicted future video features $\hat{z}_{t:t+H}$ through a single-step denoising process. These features capture the anticipated future scene evolution and establish a dynamics-aware context for downstream action reasoning. The MLP then projects these future embeddings into the input space of the FIDM, which combines them with the current latent state to infer the latent action sequence

Table 1: **CALVIN ABC-D results.** We present the average success computed over 1000 rollouts for each task and the average number of completed tasks to solve 5 instructions consecutively (Avg. Len.). DeFI shows significant superiority over baselines. The best results are **bolded**. *We reproduced results of $\pi_{0.5}$, GR00T N1 and OpenVLA-OFT on CALVIN.

View	Method	1	2	3	4	5	Avg. Len. ↑
Third View	SuSIE (Black et al., 2023)	87.0	69.0	49.0	38.0	26.0	2.69
	CLOVER (Bu et al., 2024b)	96.0	83.5	70.8	57.5	45.4	3.53
	UniVLA (Bu et al., 2025b)	**95.5**	85.8	75.4	66.9	56.5	3.80
	Ours	92.9	**87.2**	**81.2**	**75.0**	**68.4**	**4.05**
Multi-View	GR-1 (Wu et al., 2024)	85.4	71.2	59.6	49.7	40.1	3.06
	OpenVLA (Kim et al., 2024)	91.3	77.8	62.0	52.1	43.5	3.27
	Vidman (Wen et al., 2024)	91.5	76.4	68.2	59.2	46.7	3.42
	π_0* (Black et al., 2024)	93.8	85.0	76.7	68.1	59.9	3.84
	$\pi_0.5$* (Intelligence et al., 2025b)	94.8	87.4	78.2	71.7	64.3	3.97
	GR00T N1* (Bjorck et al., 2025)	94.2	86.1	79.6	73.9	66.8	4.01
	OpenVLA-OFT* (Kim et al., 2025)	95.7	88.2	82.4	74.5	70.7	4.12
	UP-VLA (Zhang et al., 2025b)	92.8	86.5	81.5	76.9	69.9	4.08
	Seer (Tian et al., 2024)	96.3	91.6	86.1	80.3	74.0	4.28
	VPP (Hu et al., 2024)	96.5	90.9	86.6	82.0	76.9	4.33
	Ours	**97.9**	**94.2**	**90.7**	**87.0**	**81.2**	**4.51**

The diffusion-based action adapter then conditions on these latent actions and produces the final executable control commands.

4 Experiments

In this section, we conduct extensive experiments on both simulated and real-world environments to evaluate the effectiveness of DeFI as shown in Figure 3.

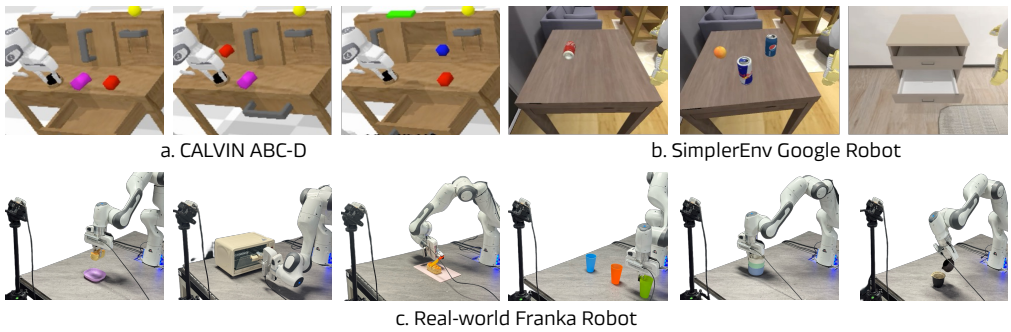


Figure 3: Experiments setup on CALVIN ABC-D, SimplerEnv Google Robot and real-world Franka Robot. We evaluate DeFI across 3 simulation environments.

4.1 Implementation Details

In the pretraining stage, we first train the foundation forward dynamics model on a diverse collection of datasets spanning both human videos (Goyal et al., 2017; Grauman et al., 2022) and robotic manipulation data (Mees et al., 2022b; O'Neill et al., 2023). In parallel, the foundation inverse dynamics model is pretrained on large-scale human egocentric datasets, including Ego4D (Grauman et al., 2022) and Open X-Embodiments (O'Neill et al., 2023). During fine-tuning, we freeze the pretrained forward dynamics model to preserve its generalization capability and generate 16-frame future predictions. All experiments are conducted on NVIDIA H100 GPUs. Detailed implementation and training protocols are provided in Appendix A.2.

4.2 Manipulation Benchmarks on CALVIN

Experiment setup and baseline. CALVIN (Mees et al., 2022b) is a simulated benchmark designed for learning long-horizon, language-conditioned robot manipulation policies. It comprises four distinct

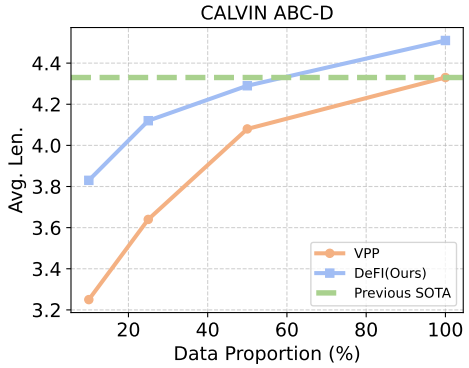


Figure 4: **Data efficiency** of DeFI's performance on CALVIN using different proportions of the action-labeled downstream data.

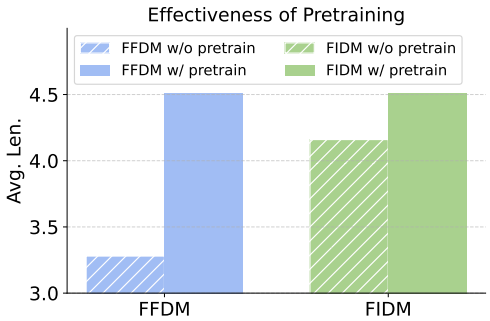


Figure 5: Ablation study for the effectiveness of decoupled forward and inverse pretraining.

manipulation environments and provides over six hours of teleoperated play data per environment, captured from multiple sensors including static and gripper-mounted RGB-D cameras, tactile images, and proprioceptive readings. We focus on the challenging ABC-D setting, where the model is trained in the ABC environment and evaluated in the unseen D environment, then report the success rate of every track and the average length of 5 tasks. We compare our model with the latest state-of-the-art generalist manipulation policies, including OpenVLA (Kim et al., 2024), Robovlm (Li et al., 2024a), π_0 (Black et al., 2024), GR1 (Wu et al., 2024), UP-VLA (Zhang et al., 2025b), Seer (Tian et al., 2024), SuSIE (Black et al., 2023), CLOVER (Bu et al., 2024b) and VPP (Hu et al., 2024). For fairness, we evaluate our approach in two setups: a static (third) view and a multi-view setting that combines static and wrist cameras.

Quantitative results and analysis. As shown in Table 1, DeFI can be effectively adapted to tasks in the CALVIN ABC-D environments under different view settings. Our method surpasses OpenVLA, π_0, and GR1, which directly project the RGB images into action signals, revealing that leveraging a powerful forward dynamics model to predict future actions would benefit current action reasoning. We compare with UniVLA, which extracts latent action labels from human videos, and then pretrains the VLA model on the large-scale datasets. It demonstrates that decoupling forward and inverse dynamics model pretraining is more effective than solely extracting latent action as pseudo labels to pretrain VLA. Additionally, DeFI surpasses UP-VLA and Seer, which integrate the visual/latent feature forecasting and action reasoning into a single VLA framework. The results demonstrate that disentangling the forward and inverse dynamics models and pretraining them on mixed videos separately would fully exploit the power of action-free videos and benefit the robot action reasoning. Compared to methods that use video generation models' predicted videos as input, like SuSIE, CLOVER and VPP, our model significantly achieves more accurate control, demonstrating that *accurate action inference is as important as accurate future prediction and a powerful inverse dynamics model leads to better performance.* Furthermore, we can find that our method is more effective in long-horizon tasks than previous methods, because our visual forward dynamics model can predict future videos and leverage a powerful FFDM to resolve the actions.

Data efficiency. Collecting robot data is both time-consuming and labor-intensive, making data efficiency crucial for robot learning. We evaluate our method on the CALVIN ABC-D benchmark, using 10%, 20%, 50%, and 100% of the available data to fine-tune pretrained policies. The results, shown in Figure 4, demonstrate that our method consistently enhances policy performance across varying data scales. Notably, under data-scarce conditions with only 10% of the training data, the pretrained policy achieves an 18% relative improvement in average task length on CALVIN ABC-D compared to VPP (Hu et al., 2024). Moreover, our method requires only about 60% of the data on CALVIN ABC-D to surpass the previous state-of-the-art baseline. These results highlight the potential of DeFI in scenarios with limited finetuning data and further push the upper bound of robot learning by introducing massive low-cost human videos.

4.3 Manipulation Benchmarks on SimplerEnv-Fractal

Experiment setup and baseline. SimplerEnv (Li et al., 2024b) features WidowX and Google Robot setups, providing diverse manipulation scenarios with varied lighting, colors, textures, and robot

Table 2: **Evaluation results across different policies on SimplerEnv**. We evaluate DeFI on 3 tasks on the Google Robot in SimplerEnv.

SimplerEnv on Google Robot Tasks								
Model	Visual Matching				Variant Aggregation			
	Pick Coke Can	Move Near	Open/Close Drawer	Avg.	Pick Coke Can	Move Near	Open/Close Drawer	Avg.
Octo-Base (Team et al., 2024)	17.0%	4.2%	22.7%	16.8%	0.6%	3.1%	1.1%	1.1%
TraceVLA (Zheng et al., 2024)	28.0%	53.7%	**57.0%**	42.0%	**60.0%**	56.4%	**31.0%**	45.0%
OpenVLA (Kim et al., 2024)	16.3%	46.2%	35.6%	27.7%	54.5%	47.7%	17.7%	39.8%
Ours	**54.2%**	**60.7%**	38.6%	**51.2%**	53.9%	**58.2%**	24.0%	**45.4%**

Table 3: **Real-world evaluation** with the Franka Robot across eight tasks.

Method	Success Rate (%)								
	Place	Open	Close	Cut	Stack Bowl	Stack Cube	Stack Bottle	Pour Water	Average
Diffusion Policy (Chi et al., 2023)	70.0	40.0	70.0	50.0	45.0	35.0	40.0	35.0	48.2
Octo-Base (Team et al., 2024)	55.0	35.0	60.0	20.0	30.0	25.0	30.0	20.0	34.4
OpenVLA (Kim et al., 2024)	50.0	40.0	65.0	40.0	30.0	35.0	45.0	45.0	43.8
Ours	**90.0**	**75.0**	**100.0**	**80.0**	**80.0**	**70.0**	**80.0**	**75.0**	**81.3**

camera pose conditions, thereby bridging the visual appearance gap between real and simulated environments. We compare our model on the Fractal branch (Google Robot) with the latest state-of-the-art generalist manipulation policies, including Octo (Team et al., 2024), TraceVLA (Zheng et al., 2024), and OpenVLA (Kim et al., 2024).

Quantitative results and analysis. Table 2 presents the SimplerEnv experimental results on the Fractal branch. DeFI also achieves state-of-the-art performance on Google robot multitasks, with an average success rate of 51.2% and 45.4% on visual matching and variant aggregation settings, respectively. However, DeFI underperforms on certain tasks. We attribute this to domain shift: the visual FFDM is pretrained on real-world datasets (Fractal (Brohan et al., 2023)) and kept frozen during finetuning, which restricts it to predicting real-world images. This mismatch propagates to the inverse dynamics model, causing it to generate erroneous actions.

4.4 Real-world Experiments

Experiment setup and baselines. As shown in Figure 6, we use the Franka Panda arm to conduct experiments evaluating the effectiveness of our method in the real world. In our setup, two RealSense D415 cameras capture RGB images: one provides a third-person view, and the other is mounted on the gripper. We collected 1,600 trajectories for 8 tasks, as shown in Table 3. In the experimental setup, each trial allows a maximum of 20 consecutive attempts. All objects are randomly positioned on the table surface. A trial is considered successful if the robotic arm grasps the target object within the specified attempts; in placement tasks, success further requires transferring the object onto a designated plate. For fair comparison, we finetune Diffusion Policy (Chi et al., 2023), Octo-Base (Team et al., 2024), OpenVLA (Kim et al., 2024) and DeFI on collected demonstration datasets.

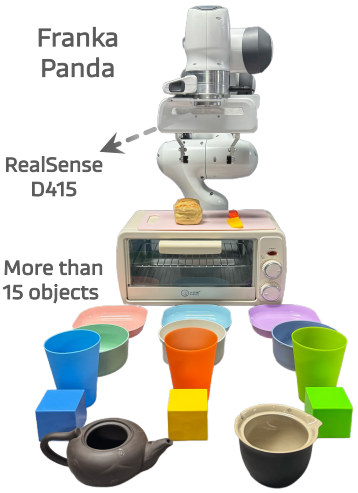


Figure 6: Real-world robot setup.

Quantitative results and analysis. As presented in Table 3, DeFI outperforms previous methods. Specifically, in simple single-task scenarios (place, open, and close), all the policies exhibit good performance($> 50\%$). However, in moderately complex tasks (cut & stack), where the models need to make the robot take or stack different colors and sizes of objects, most policies, such as DP, Octo, and OpenVLA, struggle with manipulation, frequently encountering issues like object and order misidentification. Our method surpasses these approaches thanks to the powerful prediction and generalization capabilities of the pretrained visual foundation forward dynamics model. Furthermore, in the complex long-horizon and accurate control task (pour water), our method demonstrates strong performance, accurately executing tasks like grasping up a teapot and pouring water into a cup, which relies on the powerful pretrained inverse dynamics model. Overall, DeFI achieves a higher average success rate, showcasing robust real-world operation capabilities.

Table 4: **Performance comparison** with or without decoupled pretraining.

Addition Type	Task completed in a row					
	1	2	3	4	5	Len.
FFDM w/o pre	88.0	77.6	62.4	56.8	43.2	3.28
FIDM w/o pre	96.0	88.8	83.2	76.8	71.2	4.16
All w/ pre	**97.9**	**94.2**	**90.7**	**87.0**	**81.2**	**4.51**

Table 5: **Performance comparison** with or without human videos.

Addition Type	Task completed in a row					
	1	2	3	4	5	Len.
All w/o h.v.	93.6	91.2	88.0	82.4	79.2	3.92
FFDM w/o h.v.	96.0	91.2	85.6	77.6	68.8	4.19
FIDM w/o h.v.	93.6	91.2	88.0	82.4	79.2	4.34
All w/ h.v.	**97.9**	**94.2**	**90.7**	**87.0**	**81.2**	**4.51**

Table 6: **Performance comparison** of different FFDM settings. "5 Steps" indicates five denoising steps in FFDM, while "DINO" denotes a DINO-based generative model used as FFDM.

Method	Task completed in a row					
	1	2	3	4	5	Len.
5 Steps	**98.4**	**95.2**	89.6	82.4	79.2	4.45
DINO	91.2	81.6	74.4	65.6	57.6	3.70
Ours	97.9	94.2	**90.7**	**87.0**	**81.2**	**4.51**

Table 7: **Performance comparison** of different inverse dynamics model architectures.

Method	Task completed in a row					
	1	2	3	4	5	Len.
MLP	89.6	80.8	66.4	56.0	48.8	3.42
Transformer	97.6	90.4	83.2	77.6	73.6	4.22
Ours	**97.9**	**94.2**	**90.7**	**87.0**	**81.2**	**4.51**

4.5 Ablation Study

In this section, we investigate the following questions under the multi-view setting to thoroughly evaluate the ability of our model:

Q1: What is the impact of the decoupled pretraining stage? As shown in Table 4 and Figure 5, the FFDM without pretraining achieves an average task length of 3.28, benefiting from the prior knowledge of the video generation model. However, it still suffers from limited prediction quality on robot videos. Without pretraining, the FIDM achieves an average length of 4.16, while pretraining the inverse dynamics branch provides stronger action guidance. Incorporating the full decoupled pretraining further improves performance to 4.51. These results highlight the importance of large-scale decoupled pretraining on robot and human data for stable optimization and better generalization.

Q2: What impact does human video have? As shown in Table 5, the FFDM without human videos achieves an average task length of 4.19. When the FIDM is used without human videos, the performance improves slightly to 4.34. Incorporating human videos during pretraining further increases the average task length to 4.51, yielding relative gains of +0.17 over FIDM and +0.32 over FFDM. These results indicate that the massive scale and diverse human video data provide valuable motion priors that complement robot demonstrations and enhance generalization.

Q3: How does the quality and format of the predicted image affect performance? As shown in Table 6, we study the trade-off between generation quality and latency by varying the number of denoising steps. While additional denoising slightly improves visual fidelity, it increases the latency. In our DeFI, a single denoising step requires approximately 150 ms per inference, whereas five denoising steps take around 250 ms, resulting in significantly slower performance. Notably, a single denoising step already captures sufficient semantic information about future frames, and further steps do not yield improvements in manipulation performance. We also test replacing the SVD-based video backbone with a DINO-based generative model. While DINO features converge faster and better match the FIDM feature space, they cannot integrate well with existing video-generation frameworks or leverage their pretrained knowledge, leading to inferior performance than the SVD baseline. It's a promising way to explore stronger DINO (Zhou et al., 2024) or other latent embedding prediction models (Bardes et al., 2023; Assran et al., 2025) to better understand the upper bound.

Q4: What impact do architectural variants of the inverse dynamics model have? As shown in Table 7, to rigorously evaluate the impact of inverse dynamics model architecture on overall policy performance, we compare FIDM against several common architectural variants: (i).a simple multilayer perceptron (MLP) that directly maps concatenated current and future state embeddings to actions, (ii).a Transformer that directly maps concatenated current and future state embeddings to actions, and (iii).our FIDM, which discretizes the continuous action space via a causal transformer with vector quantization using VQ-VAE. Our FIDM outperforms all alternatives. The MLP baseline fails to capture complex actions, while the Transformer accumulates errors. These results demonstrate that the design of the inverse dynamics model critically affects performance, and that our discrete action space approach with self-supervised training effectively learns a structured latent action space for precise action generation.

Figure 7: **Performance** with increasing amounts of human video data for dynamics pretraining.

Method	Task completed in a row					
	1	2	3	4	5	Len.
0.0	92.4	85.6	78.0	70.2	63.1	3.92
0.2	94.6	88.2	83.3	77.5	71.9	4.16
0.4	96.4	91.0	86.0	80.7	75.4	4.30
0.6	96.1	92.4	87.4	82.6	77.3	4.36
0.8	97.5	93.3	89.2	84.1	78.9	4.43
1.0	**97.9**	**94.2**	**90.7**	**87.0**	**81.2**	**4.51**

Table 8: **Performance comparison** of different discretization methods for stabilizing inverse dynamics learning.

Method	Task completed in a row					
	1	2	3	4	5	Len.
G.M.	94.1	90.3	84.7	78.5	72.9	4.12
S.B.	93.2	89.1	82.3	75.4	69.8	3.98
C.L.A	94.5	91.0	86.2	80.1	74.7	4.20
Ours	97.9	94.2	**90.7**	**87.0**	**81.2**	**4.51**

Figure 8: Scaling results with different amounts of human video used for dynamics pretraining.

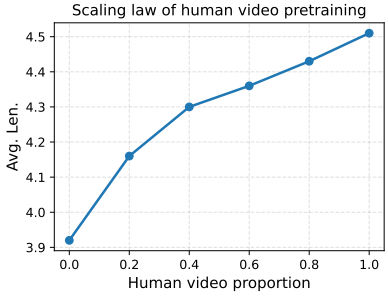


Table 9: Ablation study on finetuning FFDM, FIDM, and action adapter.

Method	Task completed in a row					
	1	2	3	4	5	Len.
Adapter Only	95.5	92.0	87.2	81.1	75.7	4.33
FFDM+Adapter	95.6	92.4	88.1	82.5	76.2	4.35
FIDM+Adapter	97.9	94.2	**90.7**	**87.0**	**81.2**	**4.51**
All Train	96.8	93.1	88.4	83.2	78.0	4.40

Q5: What is the scaling behavior of human data? To evaluate the scaling behavior of human video data, we conduct an ablation study, as shown in Table 7 and Figure 8. The results show that performance improves as the amount of human video increases, with marginal gains becoming smaller at larger scales but no saturation is observed, suggesting the potential for further improvements with even larger datasets.

Q6: How do different discretization strategies affect performance? To evaluate the effectiveness of different discretization methods, we conduct an ablation study comparing four strategies: Gaussian Mixture(G.M), Simple Binning(S.B), Continuous Latent Action(C.L.A), and our proposed Discrete VQ-VAE method. The VQ-VAE in our method serves not only as a discretization tool but also as an information-bottleneck mechanism, which stabilizes the learning of inverse dynamics. This quantization step helps prevent future-state leakage into the decoder, ensuring the model learns meaningful action representations instead of relying on low-level visual shortcuts. The results of our ablation study, shown in Table 8, demonstrate that the discrete VQ-VAE method outperforms the other discretization strategies, providing significant improvements in task performance.

Q7: What happens when different modules (FDM/IDM/Adapter) are partially fine-tuned? As shown in Table 9, Adapter Only already achieves strong performance despite having very few trainable parameters, indicating that the pretrained FFDM provides a strong and expressive latent space. FDM+Adapter yields only minor improvements, suggesting that forward prediction alone is insufficient for reliable action inference. Although "All Train" updates FFDM, IDM, and the action adapter jointly, its performance is lower than IDM+Adapter(Ours) because joint optimization introduces representation instability and gradient interference. When FFDM is finetuned, its latent outputs change throughout training, causing the input distribution of IDM to drift. As a result, the IDM must continually adapt to shifting representations, making it much harder to learn a stable and accurate action-recovery function. Additionally, the frozen forward dynamics model could provide better generalization for action reasoning.

5 Conclusion

We presented DeFI, a framework that decouples visual forward and inverse dynamics pretraining to reconcile the misalignment between 2D video forecasting and 3D action prediction while enabling learning from large-scale, action-free web videos. DeFI comprises a Foundation Forward Dynamics Model for future prediction from diverse human and robot videos and a Foundation Inverse Dynamics Model that infers latent actions from unlabeled video transitions. The two models are integrated into a unified architecture and fine-tuned end-to-end on downstream tasks, allowing them to first specialize independently and then cooperate for mutual benefit. This method consistently enhances both simulated and real tasks, showing that decoupling forward and inverse dynamics offers a scalable and effective path for VLA systems trained on Internet-scale video.

Ethic Statement

This work complies with the ICLR Code of Ethics. It does not involve human or animal subjects, nor the use of private or sensitive data. All datasets are publicly available and used under their respective licenses. The research raises no direct ethical or legal concerns, and the authors are committed to responsible and fair use of the proposed methods.

Reproducibility Statement

We have made every effort to ensure the reproducibility of our work. The proposed model, implementation details, and evaluation protocols are described in detail in the main paper and appendix. All datasets used are publicly available and properly referenced. To further support reproducibility, we will release the source code.

References

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. 9

Fan Bao, Chendong Xiang, Gang Yue, Guande He, Hongzhou Zhu, Kaiwen Zheng, Min Zhao, Shilong Liu, Yaole Wang, and Jun Zhu. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models. *arXiv preprint*, 2024. 2

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-jepa: Latent video prediction for visual representation learning. 2023. 9

Suneel Belkhale, Yuchen Cui, and Dorsa Sadigh. Hydra: Hybrid robot actions for imitation learning. In *CoRL*, 2023. 20

Suneel Belkhale, Tianli Ding, Ted Xiao, Pierre Sermanet, Quon Vuong, Jonathan Tompson, Yevgen Chebotar, Debidatta Dwibedi, and Dorsa Sadigh. Rt-h: Action hierarchies using language. *arXiv preprint*, 2024. 3

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*, 2024. 1, 3

Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2act: Predicting point tracks from internet videos enables generalizable robot manipulation. In *ECCV*, 2024. 3

Hongzhe Bi, Lingxuan Wu, Tianwei Lin, Hengkai Tan, Zhizhong Su, Hang Su, and Jun Zhu. H-rdt: Human manipulation enhanced bimanual robotic manipulation. *arXiv preprint*, 2025. 3

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 foundation model for generalist humanoid robots. *arXiv preprint*, 2025. 3, 6

Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pretrained image-editing diffusion models. *arXiv preprint*, 2023. 2, 3, 6, 7

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. pi0: A vision-language-action flow model for general robot control. *arXiv preprint*, 2024. 1, 3, 6, 7

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint*, 2023. 18

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael S. Ryoo, Grecia Salazar, Pannag R. Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong T. Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-1: robotics transformer for real-world control at scale. In *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*, 2023. 3, 8, 18, 20

Qingwen Bu, Hongyang Li, Li Chen, Jisong Cai, Jia Zeng, Heming Cui, Maoqing Yao, and Yu Qiao. Towards synergistic, generalized, and efficient dual-system for robotic manipulation. *arXiv preprint*, 2024a. 2

Qingwen Bu, Jia Zeng, Li Chen, Yanchao Yang, Guyue Zhou, Junchi Yan, Ping Luo, Heming Cui, Yi Ma, and Hongyang Li. Closed-loop visuomotor control with generative expectation for robotic manipulation. *arXiv preprint*, 2024b. 6, 7

Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xu Huang, Shu Jiang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint*, 2025a. 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*, 2025b. 6, 18

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*, 2025. 3

Lawrence Yunliang Chen, Simeon Adebola, and Ken Goldberg. Berkeley UR5 demonstration dataset. https://sites.google.com/view/berkeley-ur5/home, 2024a. 20

Xiaoyu Chen, Hangxing Wei, Pushi Zhang, Chuheng Zhang, Kaixin Wang, Yanjiang Guo, Rushuai Yang, Yucen Wang, Xinquan Xiao, Li Zhao, et al. villa-x: Enhancing latent action modeling in vision-language-action models. *arXiv preprint*, 2025. 3

Yi Chen, Yuying Ge, Yizhuo Li, Yixiao Ge, Mingyu Ding, Ying Shan, and Xihui Liu. Moto: Latent motion token as the bridging language for robot manipulation. *arXiv preprint*, 2024b. 3, 18

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*, 2023. 8

Jeremy A Collins, Loránd Cheng, Kunal Aneja, Albert Wilcox, Benjamin Joffe, and Animesh Garg. Amplify: Actionless motion priors for robot learning from videos. *arXiv preprint*, 2025. 3

Zichen Jeff Cui, Yibin Wang, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. From play to policy: Conditional behavior generation from uncurated robot data. *arXiv preprint*, 2022. 20

Zichen Jeff Cui, Hengkai Pan, Aadhithya Iyer, Siddhant Haldar, and Lerrel Pinto. Dynamo: In-domain dynamics pretraining for visuo-motor control. *NeurIPS*, 2024. 3

Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. *NeurIPS*, 2022. 20

Shivin Dass, Jullian Yapeter, Jesse Zhang, Jiahui Zhang, Karl Pertsch, Stefanos Nikolaidis, and Joseph J. Lim. Clvr jaco play dataset, 2023. URL https://github.com/clvrai/clvr_jaco_play_dataset. 20

Shengliang Deng, Mi Yan, Songlin Wei, Haixin Ma, Yuxin Yang, Jiayi Chen, Zhiqi Zhang, Taoyu Yang, Xuheng Zhang, Heming Cui, et al. Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data. *arXiv preprint*, 2025. 3

Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *NeurIPS*, 2024. 2, 3

Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. *arXiv preprint*, 2021. 3, 20

Yao Feng, Hengkai Tan, Xinyi Mao, Guodong Liu, Shuhe Huang, Chendong Xiang, Hang Su, and Jun Zhu. Generalist bimanual manipulation via foundation video diffusion models. *arXiv preprint*, 2025. 2, 3

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017. 6, 18, 20

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022. 6, 18, 20

Minho Heo, Youngwoon Lee, Doohyun Lee, and Joseph J Lim. Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation. *The International Journal of Robotics Research*, 2023. 20

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. *arXiv preprint*, 2024. 2, 3, 5, 6, 7

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. pi0.5: a vision-language-action model with open-world generalization. *arXiv preprint*, 2025a. 3

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. π0. 5: a vision-language-action model with open-world generalization. *arXiv preprint*, 2025b. 6

Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *CoRL*, 2022. 20

Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *CoRL*, 2018. 20

Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic vlms: Investigating the design space of visually-conditioned language models. *arXiv preprint*, 2024. 1, 3

Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. *arXiv preprint*, 2024. 3

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint*, 2024. 3

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*, 2024. 1, 3, 6, 7, 8

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint*, 2025. 6

Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. *arXiv preprint*, 2025. 3

Xinghang Li, Peiyan Li, Minghuan Liu, Dong Wang, Jirong Liu, Bingyi Kang, Xiao Ma, Tao Kong, Hanbo Zhang, and Huaping Liu. Towards generalist robot policies: What matters in building vision-language-action models. *arXiv preprint*, 2024a. 7

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Oier Mees, Karl Pertsch, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard (eds.), *CoRL*, 2024b. 2, 7, 18

Junbang Liang, Ruoshi Liu, Ege Ozguroglu, Sruthi Sudhakar, Achal Dave, Pavel Tokmakov, Shuran Song, and Carl Vondrick. Dreamitate: Real-world visuomotor policy learning via video generation. *arXiv preprint*, 2024. 2

Zhixuan Liang, Yizhuo Li, Tianshuo Yang, Chengyue Wu, Sitong Mao, Liuao Pei, Xiaokang Yang, Jiangmiao Pang, Yao Mu, and Ping Luo. Discrete diffusion vla: Bringing discrete diffusion to action decoding in vision-language-action policies. *arXiv preprint*, 2025. 3

Fanqi Lin, Ruiqian Nai, Yingdong Hu, Jiacheng You, Junming Zhao, and Yang Gao. Onetwovla: A unified vision-language-action model with adaptive reasoning. *arXiv preprint*, 2025. 3

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023. 3

Huihan Liu, Soroush Nasiriany, Lance Zhang, Zhiyao Bao, and Yuke Zhu. Robot learning on the job: Human-in-the-loop autonomy and learning during deployment. *The International Journal of Robotics Research*, 2022. 20

Hao Luo and Zongqing Lu. Learning video-conditioned policy on unlabelled data with joint embedding predictive transformer. In *ICLR*, 2025. 3

Hao Luo, Yicheng Feng, Wanpeng Zhang, Sipeng Zheng, Ye Wang, Haoqi Yuan, Jiazheng Liu, Chaoyi Xu, Qin Jin, and Zongqing Lu. Being-h0: Vision-language-action pretraining from large-scale human videos. *arXiv preprint*, 2025. 3

Jianlan Luo, Charles Xu, Xinyang Geng, Gilbert Feng, Kuan Fang, Liam Tan, Stefan Schaal, and Sergey Levine. Multistage cable routing through hierarchical imitation learning. *IEEE Transactions on Robotics*, 2024. 20

Corey Lynch, Ayzaan Wahid, Jonathan Tompson, Tianli Ding, James Betker, Robert Baruch, Travis Armstrong, and Pete Florence. Interactive language: Talking to robots in real time. *IEEE Robotics and Automation Letters*, 2023. 20

Jiangran Lyu, Ziming Li, Xuesong Shi, Chaoyi Xu, Yizhou Wang, and He Wang. Dywa: Dynamics-adaptive world action model for generalizable non-prehensile manipulation. *arXiv preprint*, 2025. 3

Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Tingfan Wu, Jay Vakil, et al. Where are we in the search for an artificial visual cortex for embodied intelligence? *NeurIPS*, 2023. 3

Ajay Mandlekar, Yuke Zhu, Animesh Garg, Jonathan Booher, Max Spero, Albert Tung, Julian Gao, John Emmons, Anchit Gupta, Emre Orbay, et al. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. In *CoRL*, 2018. 20

Oier Mees, Jessica Borja-Diaz, and Wolfram Burgard. Grounding language with visual affordances over unstructured data. *arXiv preprint*, 2022a. 20

Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *IEEE Robotics and Automation Letters*, 2022b. 2, 6, 18, 20, 21

Russell Mendonca, Shikhar Bahl, and Deepak Pathak. Structured world models from human videos. *arXiv preprint*, 2023. 20

Soroush Nasiriany, Tian Gao, Ajay Mandlekar, and Yuke Zhu. Learning and retrieval from prior data for skill-based imitation learning. *arXiv preprint*, 2022. 20

Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint*, 2023. 3, 6, 18, 20

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024, 2024. 5, 18

Zekun Qi, Runpei Dong, Shaochen Zhang, Haoran Geng, Chunrui Han, Zheng Ge, Li Yi, and Kaisheng Ma. Shapellm: Universal 3d object understanding for embodied interaction. In *ECCV*, 2024. 21

Zekun Qi, Wenyao Zhang, Yufei Ding, Runpei Dong, Xinqiang Yu, Jingwen Li, Lingyun Xu, Baoyu Li, Xialin He, Guofan Fan, et al. Sofar: Language-grounded orientation bridges spatial reasoning and object manipulation. *arXiv preprint*, 2025. 21

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*, 2025. 3

Gabriel Quere, Annette Hagengruber, Maged Iskandar, Samuel Bustamante, Daniel Leidner, Freek Stulp, and Jörn Vogel. Shared control templates for assistive robotics. In *ICRA*, 2020. 20

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 4, 18

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 2020. 5

Saumya Saxena, Mohit Sharma, and Oliver Kroemer. Multi-resolution sensing for real-time control with vision-language models. In *2nd Workshop on Language and Robot Learning: Language as Grounding*, 2023. 20

Nur Muhammad Mahi Shafiullah, Anant Rai, Haritheja Etukuru, Yiqian Liu, Ishan Misra, Soumith Chintala, and Lerrel Pinto. On bringing robots home. *arXiv preprint*, 2023. 20

Rutav Shah, Roberto Martín-Martín, and Yuke Zhu. Mutex: Learning unified policies from multi-modal task specifications. In *CoRL*, 2023. 20

Hengkai Tan, Yao Feng, Xinyi Mao, Shuhe Huang, Guodong Liu, Zhongkai Hao, Hang Su, and Jun Zhu. Anypos: Automated task-agnostic actions for bimanual manipulation. *arXiv preprint*, 2025. 3

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*, 2024. 8

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. *ICLR*, 2024. 2, 3, 6, 7

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 2017. 5

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 *CoRL*, 2023. 20

Yuqi Wang, Xinghang Li, Wenxuan Wang, Junbo Zhang, Yingyan Li, Yuntao Chen, Xinlong Wang, and Zhaoxiang Zhang. Unified vision-language-action model. *arXiv preprint*, 2025. 3

Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning. *arXiv preprint*, 2023. 3

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. *NeurIPS*, 2024. 3, 6

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. In *ICLR*, 2024. 3, 6, 7

Amber Xie, Oleh Rybkin, Dorsa Sadigh, and Chelsea Finn. Latent diffusion planning for imitation learning. *arXiv preprint*, 2025. 3

Mingxing Xu, Wenrui Dai, Chunmiao Liu, Xing Gao, Weiyao Lin, Guo-Jun Qi, and Hongkai Xiong. Spatial-temporal transformer networks for traffic flow forecasting. *arXiv preprint*, 2020. 5

Haoru Xue, Xiaoyu Huang, Dantong Niu, Qiayuan Liao, Thomas Kragerud, Jan Tommy Gravdahl, Xue Bin Peng, Guanya Shi, Trevor Darrell, Koushil Sreenath, et al. Leverb: Humanoid whole-body control with latent vision-language instruction. *arXiv preprint arXiv:2506.13751*, 2025. 3

Jiange Yang, Yansong Shi, Haoyi Zhu, Mingyu Liu, Kaijing Ma, Yating Wang, Gangshan Wu, Tong He, and Limin Wang. Como: Learning continuous latent motion from internet videos for scalable robot learning. *arXiv preprint*, 2025. 3

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, et al. Latent action pretraining from videos. *arXiv preprint*, 2024. 3, 5, 18

Hongyin Zhang, Pengxiang Ding, Shangke Lyu, Ying Peng, and Donglin Wang. Gevrm: Goal-expressive video generation model for robust visual manipulation. *arXiv preprint*, 2025a. 3

Jianke Zhang, Yanjiang Guo, Yucheng Hu, Xiaoyu Chen, Xiang Zhu, and Jianyu Chen. Up-vla: A unified understanding and prediction model for embodied agent. *arXiv preprint*, 2025b. 6, 7

Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks. *arXiv preprint*, 2024a. 3

Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. Navid: Video-based vlm plans the next step for vision-and-language navigation. *Robotics: Science and Systems*, 2024b. 3

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*, 2025c. 2, 3

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. *arXiv preprint*, 2025. 2, 3

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*, 2024. 8

Gaoyue Zhou, Victoria Dean, Mohan Kumar Srirama, Aravind Rajeswaran, Jyothish Pari, Kyle Hatch, Aryan Jain, Tianhe Yu, Pieter Abbeel, Lerrel Pinto, et al. Train offline, test online: A real robot learning benchmark. *arXiv preprint*, 2023. 20

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024. 9

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. *arXiv preprint*, 2025. 5

Xinghao Zhu, Ran Tian, Chenfeng Xu, Mingxiao Huo, Wei Zhan, Masayoshi Tomizuka, and Mingyu Ding. Fanuc manipulation: A dataset for learning-based manipulation with fanuc mate 200id robot. https://sites.google.com/berkeley.edu/fanuc-manipulation, 2023a. 20

Yifeng Zhu, Peter Stone, and Yuke Zhu. Bottom-up skill discovery from unsegmented demonstrations for long-horizon robot manipulation. *IEEE Robotics and Automation Letters*, 2022. 20

Yifeng Zhu, Abhishek Joshi, Peter Stone, and Yuke Zhu. Viola: Imitation learning for vision-based manipulation with object proposal priors. In *CoRL*, 2023b. 20

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong T. Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: vision-language-action models transfer web knowledge to robotic control. In *CoRL*, 2023. 1, 3

A Appendix

A.1 The Use of Large Language Models

Large language models are used solely as writing assistants for grammar refinement and expression polishing. They do not contribute to research ideation, methodology design, experiments, or analysis.

A.2 Implementation Details

FFDM Pretraining Details. For pretraining the foundation forward dynamics model, we use a mixture of robot video datasets (Open X-Embodiment (O'Neill et al., 2023), CALVIN (Mees et al., 2022b)) and human video datasets (Something-Something-v2 (Goyal et al., 2017), Ego4D (Grauman et al., 2022)). To appropriately balance the contributions of different datasets, we adopt varying sampling ratios. Detailed dataset information is provided in Table 11.

FIDM Pretraining Details. For pretraining the foundation inverse dynamics model, we use a subset of the Open X-Embodiment dataset (O'Neill et al., 2023) containing single-arm end-effector control. Although actions and proprioceptive states are available in these robot datasets, we exclude them during pretraining and rely only on episode frames and text instructions. We further incorporate open-world human videos, specifically egocentric recordings of daily activities from the Ego4D dataset (Grauman et al., 2022). Except for the SimplerEnv benchmark (Li et al., 2024b), which replicates the environment of the Fractal dataset (Brohan et al., 2023), none of the downstream evaluation environments (e.g., CALVIN (Mees et al., 2022b)) are seen during pretraining, thereby requiring strong generalization capabilities from the model. The dataset composition and sampling ratios are detailed in Table 12.

Coupled Finetuning Details For the coupled finetuning stage, we freeze the foundation forward dynamics model while finetuning the foundation inverse dynamics model and the latent action adapter. Training is conducted on the CALVIN-ABC dataset for evaluation on the CALVIN benchmark (Mees et al., 2022b), and on the Fractal dataset (Brohan et al., 2023) for evaluation on the SimplerEnv benchmark (Li et al., 2024b).

We summarize the training and model parameters of each component of DeFI in Table 10.

A.3 Model Architecture

Foundation forward dynamics model. We adopt the open-sourced Stable Video Diffusion (SVD) (Blattmann et al., 2023) as the foundation for the forward dynamics model. We further enhance it by incorporating language instructions through CLIP (Radford et al., 2021) and adjusting the output video resolution to 256×256, aligning with the resolution of robot datasets (O'Neill et al., 2023; Mees et al., 2022b).

Foundation inverse dynamics model. We adopt a Transformer architecture as the foundation inverse dynamics model and train it in the DINO feature (Oquab et al., 2024) space to obtain semantically rich representations, following prior work (Bu et al., 2025b). The pseudo-code for the pretraining process is shown in Algorithm 1. Following previous works (Ye et al., 2024; Chen et al., 2024b; Bu et al., 2025b), we use a Transformer decoder to reconstruct the features of future frames when training the foundation inverse dynamics model. However, the Transformer decoder is discarded during the fine-tuning stage of the coupled foundation forward dynamics model (FFDM) and foundation inverse dynamics model (FIDM).

Diffusion-based action adapter. We adopt a Diffusion Transformer architecture as the action adapter to decode latent action features into robot actions. The language instruction, encoded by the CLIP encoder, is combined with the latent action features obtained from the foundation inverse dynamics model and serves as conditioning for the action denoising process, which generates the final robot actions.

Table 10: Training and model parameters used in our DeFI.

Train parameter	**Value**
GPU	NVIDIA H100
Number of GPUs	8
Pretraining time of FFDM	3 days
Pretraining time of FIDM	1.5 days
Finetuning time on CALVIN	0.5 days
Training memory on CALVIN	64G
Inference memory on CALVIN	7G
Batch size	32
Learning rate	1×10^{-4}
Weight decay	1×10^{-2}
Optimizer	AdamW
Pretraining epochs	20
Finetuning epochs	12
Model parameter	**Value**
Foundation Forward Dynamics Model	
Model type	Stable Video Diffusion
Image size	256×256
Predicted future frames	16
Foundation Inverse Dynamics Model	
Model type	Transformer
Feature dimension	768
Vocabulary size of VQ codebook	128
Number of layers	16
Action Adapter	
Model type	Diffusion Transformer
Feature dimension	384
Number of layers	12
Sampling steps	10
Action dimension	7

Algorithm 1: Foundation Inverse Dynamics Model Training

Input: Current frame o_t, future frame o_{t+n}, language instruction L
Output: Predicted DINO feature of o_{t+n}

$F_t \leftarrow \text{DINOEncoder}(o_t)$; `// DINO feature of current frame`
$F_{t+n} \leftarrow \text{DINOEncoder}(o_{t+n})$; `// DINO feature of future frame`
$F_L \leftarrow \text{TextEncoder}(L)$; `// Instruction embedding`
$H \leftarrow \text{Spatial-temporal Transformer}(F_t, F_{t+n}, F_L)$; `// Foundation Inverse Dynamics Model`
$A \leftarrow \text{VQ-VAE}(H)$; `// Latent action feature`
$\hat{F}_{t+n} \leftarrow \text{TransformerDecoder}(F_t, A)$; `// Decoded future DINO feature`
$\mathcal{L}_{\text{pred}} \leftarrow \text{Loss}(\hat{F}_{t+n}, F_{t+n})$; `// Prediction loss`
$\mathcal{L}_{\text{VQ}} \leftarrow \text{Loss}(H, A)$; `// VQ-VAE loss`
$\mathcal{L} \leftarrow \mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{VQ}}$; `// Total loss`

A.4 Experiments

A.4.1 Real-world Experiments.

As shown in Figure 17, success is recorded only if both the grasping and placement operations are completed within the allowed attempts. For the articulated object manipulation tasks(open & close), the microwave is randomly placed in front of the robotic arm. The experiment is considered

Table 11: The foundation forward dynamics model of our DeFI is trained on a mixture of data from the Open X-Embodiment (O'Neill et al., 2023) and CALVIN (Mees et al., 2022b) robot video datasets, as well as the Ego4D (Grauman et al., 2022) and Something-Something-v2 (Goyal et al., 2017) human video datasets. The proportions are normalized to sum to 100%.

Category	Training dataset mixture	Proportion
Robot Videos	Fractal (Brohan et al., 2023)	30%
	Bridge (Ebert et al., 2021; Walke et al., 2023)	10%
	CALVIN-ABC (Mees et al., 2022b)	30%
Human Videos	Something-Something-v2 Goyal et al. (2017)	15%
	Ego4D Grauman et al. (2022)	15%

Table 12: The foundation inverse dynamics model of our DeFI is trained on a mixture of data from the Open X-Embodiment (O'Neill et al., 2023) robot video dataset and the Ego4D (Grauman et al., 2022) human video dataset. The proportions are normalized to sum to 100%.

Category	Training dataset mixture	Proportion
Robot Video	Fractal (Brohan et al., 2023)	16.3%
	Kuka (Kalashnikov et al., 2018)	7.4%
	Bridge (Ebert et al., 2021; Walke et al., 2023)	8.0%
	Taco Play (Mees et al., 2022a)	4.1%
	Jaco Play (Dass et al., 2023)	0.7%
	Berkeley Cable Routing (Luo et al., 2024)	0.4%
	Roboturk (Mandlekar et al., 2018)	3.3%
	Viola (Zhu et al., 2023b)	1.3%
	Berkeley Autolab UR5 (Chen et al., 2024a)	1.6%
	Toto (Zhou et al., 2023)	2.8%
	Language Table (Lynch et al., 2023)	6.1%
	Stanford Hydra Dataset (Belkhale et al., 2023)	6.2%
	Austin Buds Dataset (Zhu et al., 2022)	0.4%
	NYU Franka Play Dataset (Cui et al., 2022)	1.2%
	Furniture Bench Dataset (Heo et al., 2023)	3.4%
	UCSD Kitchen Dataset (Darkhalil et al., 2022)	0.1%
	Austin Sailor Dataset (Nasiriany et al., 2022)	3.0%
	Austin Sirius Dataset (Liu et al., 2022)	2.3%
	DLR EDAN Shared Control (Quere et al., 2020)	0.1%
	IAMLab CMU Pickup Insert (Saxena et al., 2023)	1.3%
	UTAustin Mutex (Shah et al., 2023)	3.0%
	Berkeley Fanuc Manipulation (Zhu et al., 2023a)	1.1%
	CMU Stretch (Mendonca et al., 2023)	0.2%
	BC-Z (Jang et al., 2022)	10.3%
	FMB Dataset (Lynch et al., 2023)	9.8%
	Dobbe (Shafiullah et al., 2023)	2.0%
Human Video	Ego4D (Grauman et al., 2022)	3.5%

successful if the door displacement exceeds 5 cm, indicating effective interaction. For the cutting task, the robot has to take the knife to cut the bread.

A.4.2 Additional Ablation Study

Q8: The visualization of different denoising step. As shown in Figure 9, we present the visualization of the forward dynamics model across different denoising steps(1, 10, 20). Although a single denoising step may blur the background regions, the motion-critical information—such as the object trajectory and the robot end-effector path—remains well preserved, which explains why task performance remains stable. In addition, finetuning the inverse dynamics model and the adapter

further mitigates potential information loss, enabling the model to operate reliably even under this aggressive optimization.

Q9: The statistics of failure cases. We examined 200 failure cases on CALVIN (failures defined as not completing the task within 280 steps). The errors can be grouped into two major categories:

(i) Forward-dynamics failures (62%): More challenging scenarios occur in contact-rich or cluttered interactions, where the FDM may generate hallucinated or physically implausible predictions and mismatch videos for multi-view. In these cases, the prediction error becomes too large for the IDM to compensate. Figure 10 shows such a scenario, where accumulated inconsistencies in the imagined future scene ultimately mislead the controller. These failures indicate that long-horizon consistency and contact modeling remain bottlenecks for world-model-based approaches.

(ii) Inverse-dynamics failures (38%): Even when the predicted future is accurate, the IDM may still produce incorrect actions—for instance, misplacing objects, failing to grasp, or causing collisions—as shown in Figure 11. These failures highlight that action inference is an equally critical component and can bottleneck overall performance independent of the world model's accuracy.

This highlights our motivation: accurate inverse dynamics is as essential as accurate forward prediction for reliable control.

A.5 Inference Latency

As shown in Table 13, we evaluate the inference time of our model on the CALVIN (Mees et al., 2022b) benchmark. The inference time of each component of the model is reported, averaged over five runs.

Table 13: The inference time of our DeFI is measured on an NVIDIA GeForce RTX 4090 GPU, averaged over five runs.

Model part	Inference time
Foundation Forward Dynamics Model	86.1ms
Foundation Inverse Dynamics Model	42.9ms
Action Adapter	24.3ms

A.6 Discussions and Future Work

Our model introduces a new framework that disentangles robot learning into a foundation forward dynamics model and a foundation inverse dynamics model, enabling full utilization of large-scale action-free videos from both humans and robots. This represents a fundamental improvement over existing vision-language-action (VLA) architectures, particularly in scenarios where embodied intelligence data is costly. For the limitation, our model does not incorporate a large language model, and therefore lacks the ability to support language-based interaction. Interaction and embodied reasoning (Qi et al., 2025; 2024) are crucial for complex robotic tasks. For future work, we aim to integrate our FFDM and FIDM with a large language model as a foundation understanding module, enabling the model to unify prediction, interaction, and action execution capabilities.

A.7 Qualitative Results

Heatmap of FIDM. As shown in Figure 12, we visualize the attention maps of the foundation inverse dynamics model (FIDM) to demonstrate its ability to capture actions from both robot and human videos. The results indicate that the model consistently attends to the robot arm or human arm across different time steps, enabling it to extract latent actions that guide the generation of executable actions. This highlights the benefit of large-scale pretraining in grounding the model's action understanding.

Qualitative results on the benchmark. As shown in Figure 13 and Figure 14, we visualize the results of the CALVIN benchmark on long-horizon tasks. Our DeFI performs well on sequences consisting of five consecutive tasks. Similarly, Figure 15 and Figure 16 present results on the

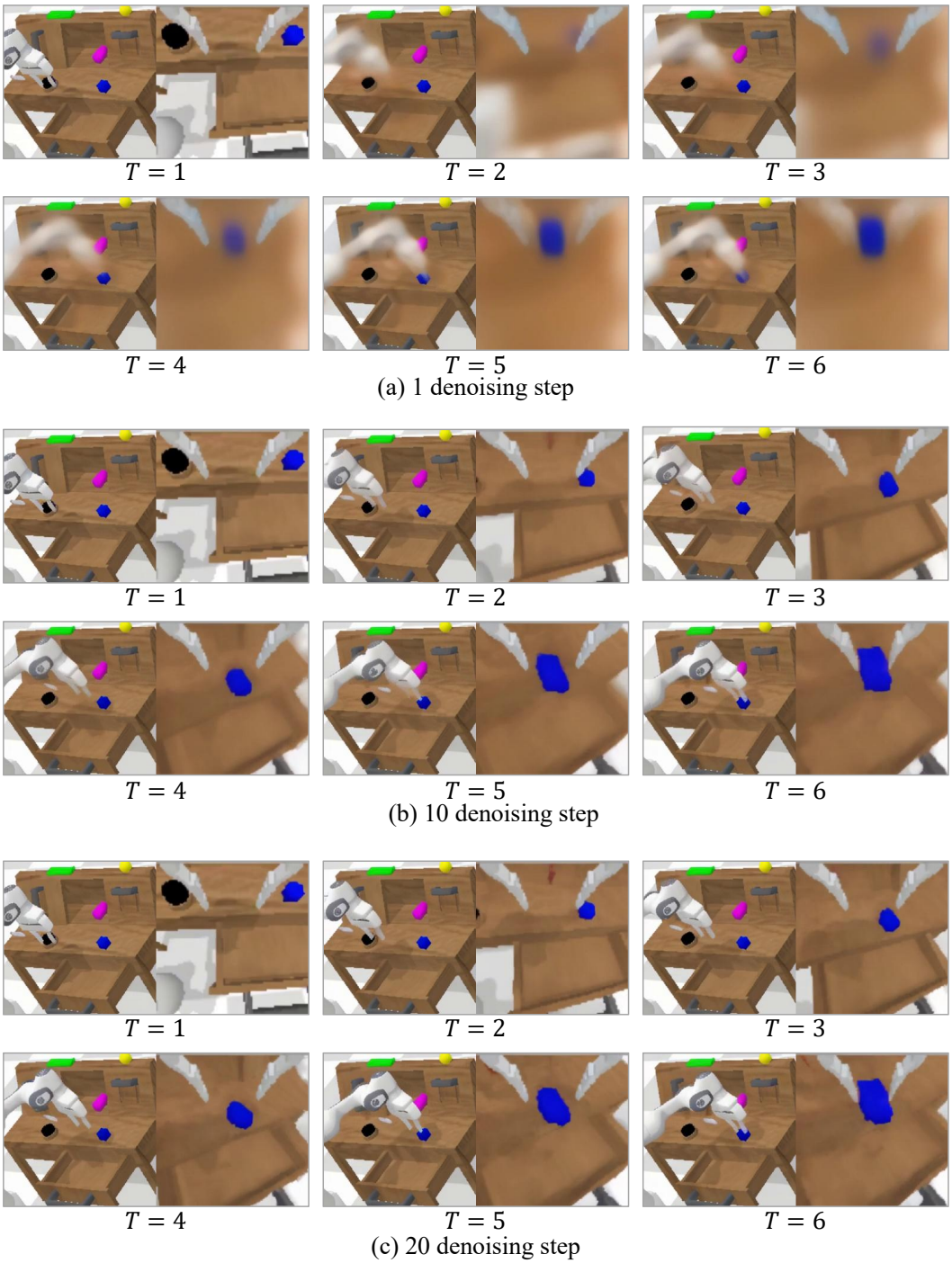
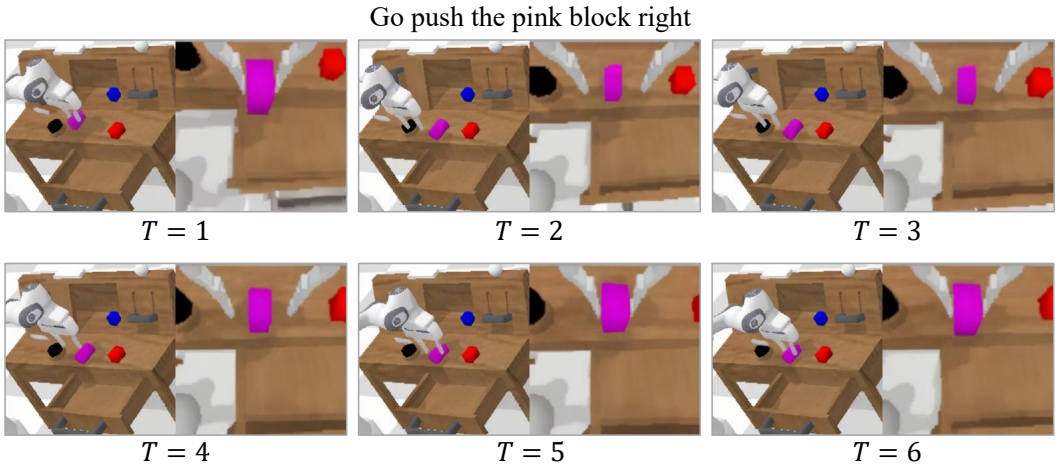


Figure 9: Qualitative results of different denosing step for the forward dynamics model.

SimplerEnv benchmark for the tasks "pick coke can" and "close drawer", where DeFI completes the tasks coherently according to the given instructions.

Qualitative results on the real-world environments. As shown in Figure 17, we visualize the results of the real-world environments across eight tasks.



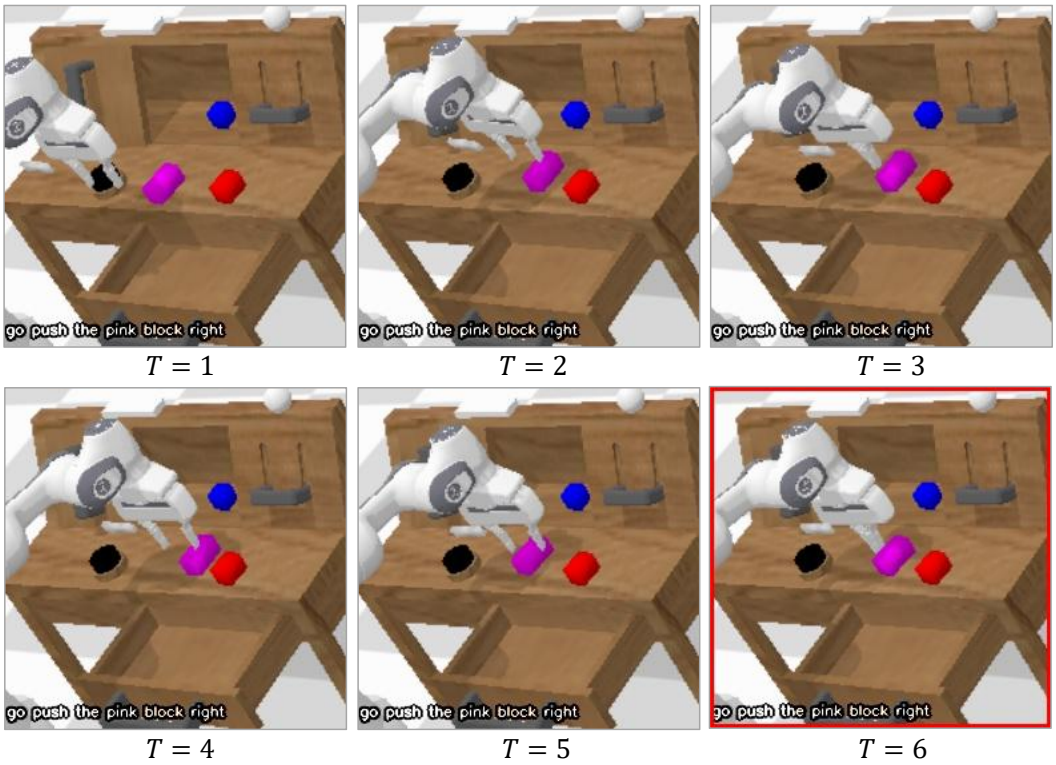
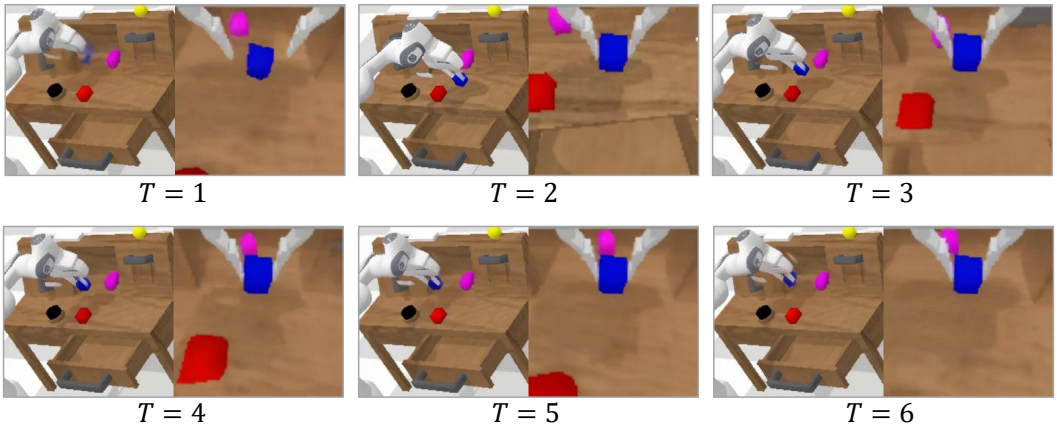
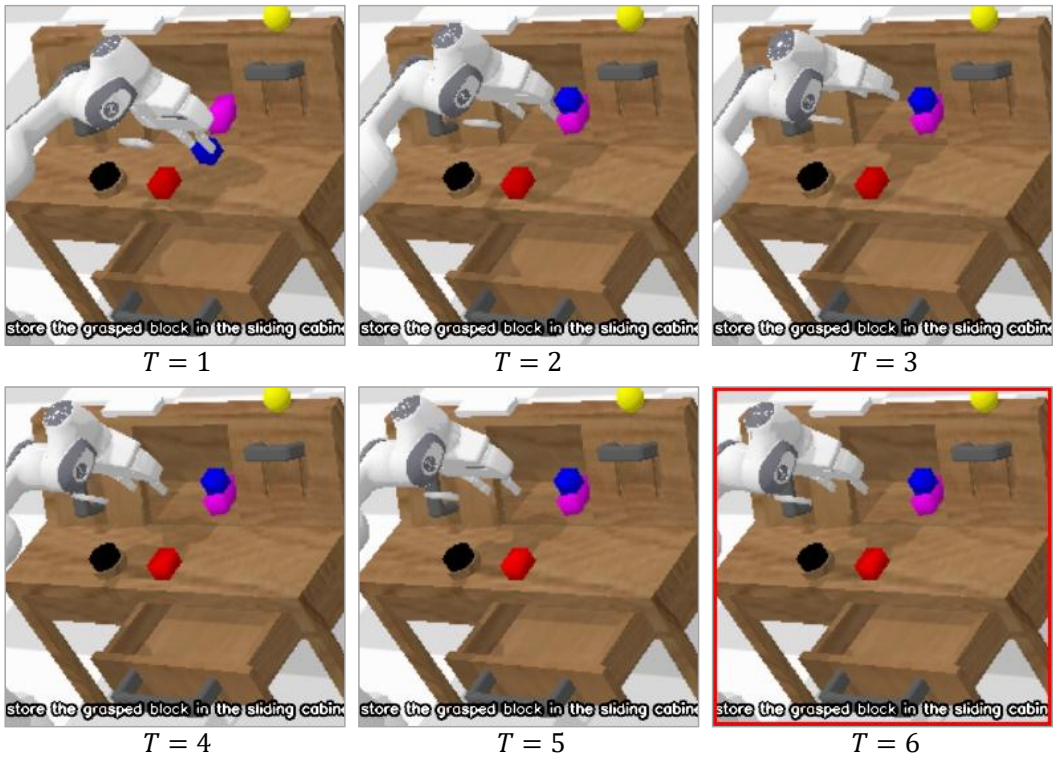


Figure 10: Qualitative results of failure cases.

Store the grasped block in the sliding cabinet



Forward Dynamics Model generation (success)



Inverse Dynamics Model action (fail)

Figure 11: Qualitative results of failure cases.

Figure 12: Qualitative attention heatmap results of the foundation inverse dynamics model on robot and human videos.

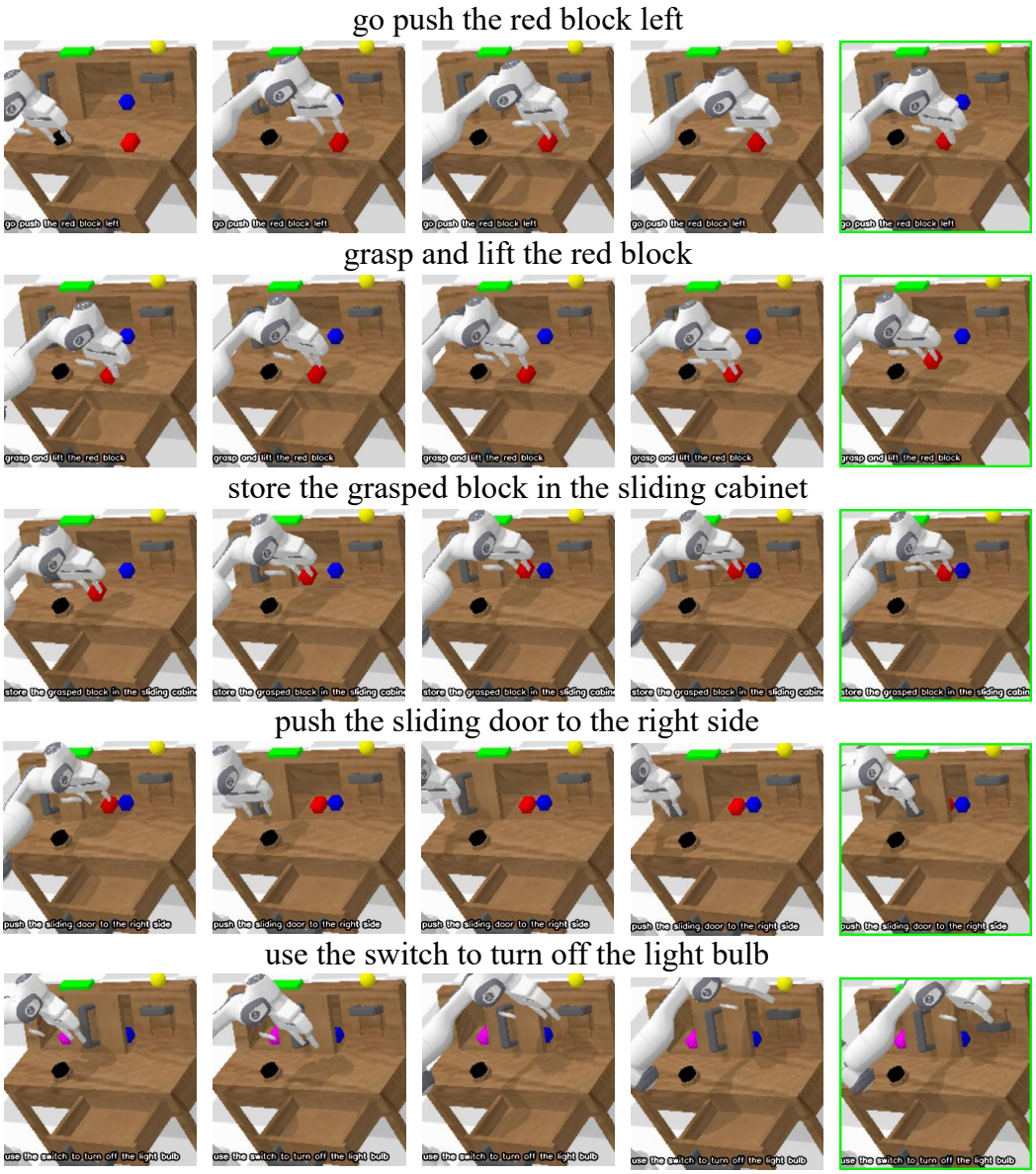


Figure 13: Qualitative results of the CALVIN long-horizon task.

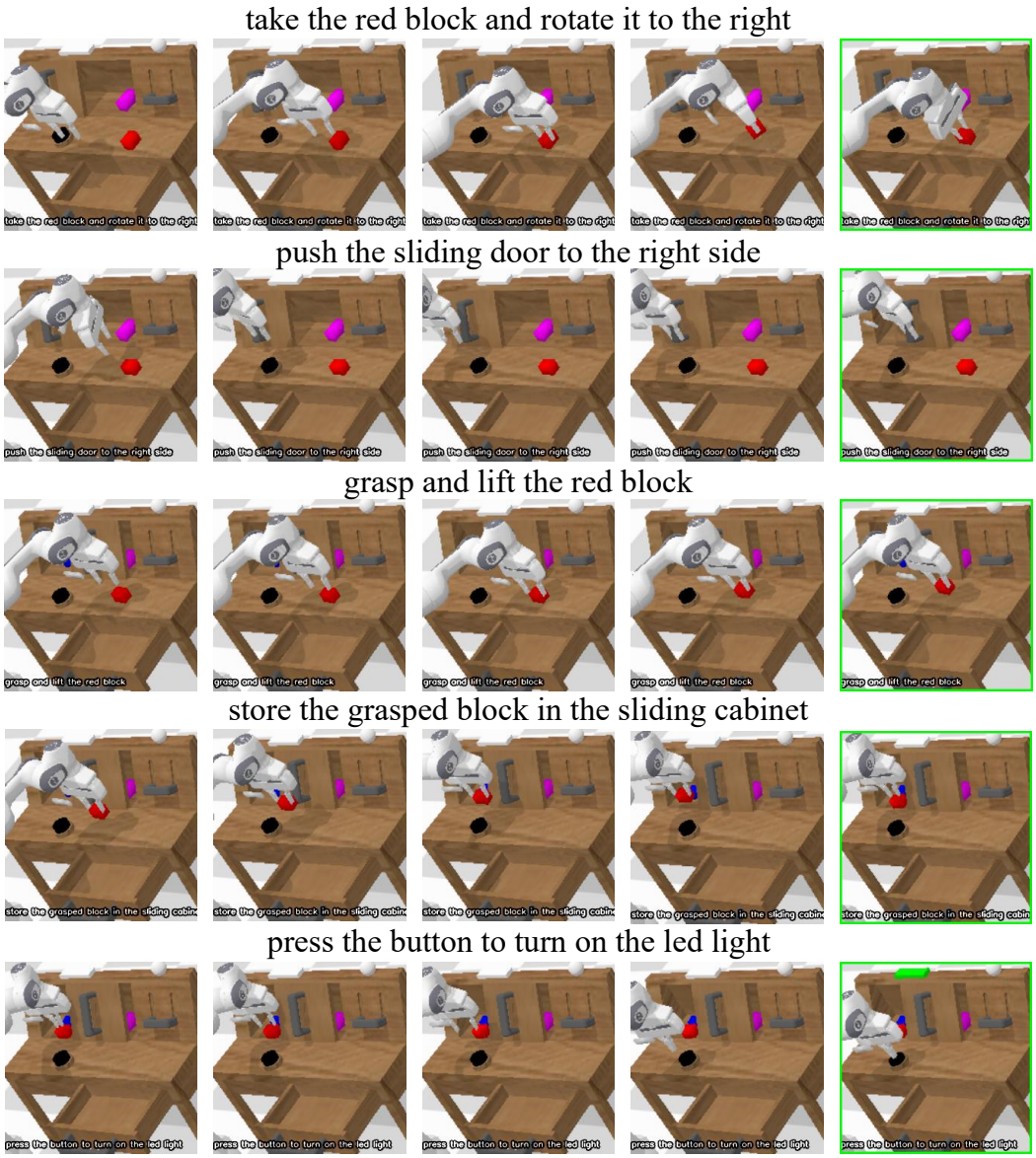


Figure 14: Qualitative results of the CALVIN long-horizon task.

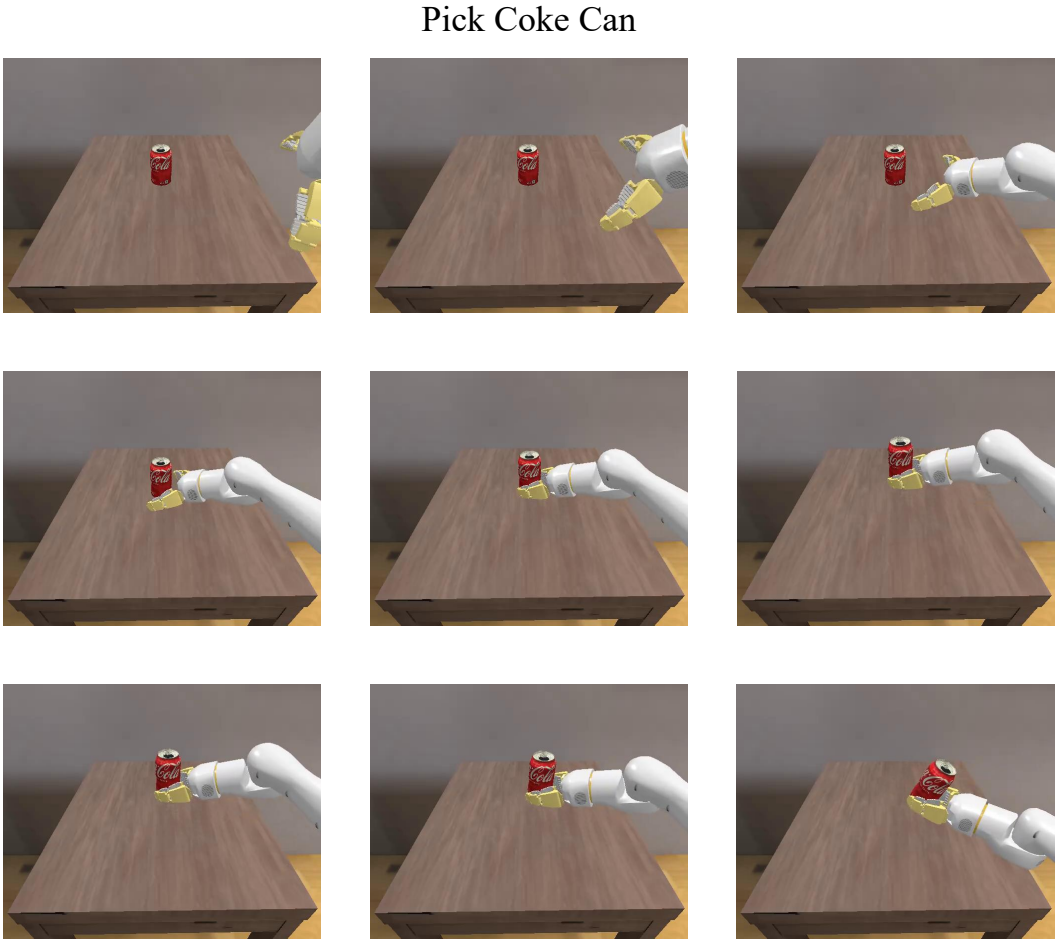


Figure 15: Qualitative results of SimplerEnv evaluation on Google Robot.

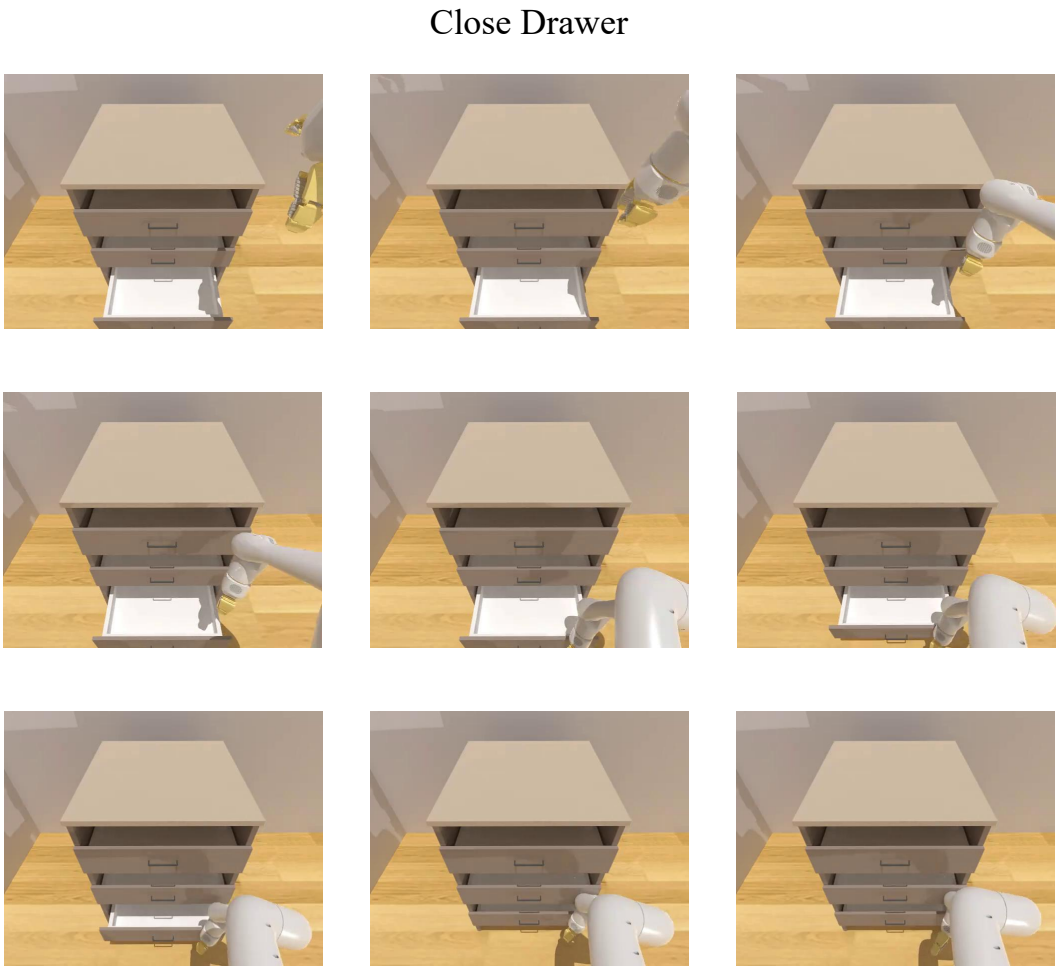


Figure 16: Qualitative results of SimplerEnv evaluation on Google Robot.

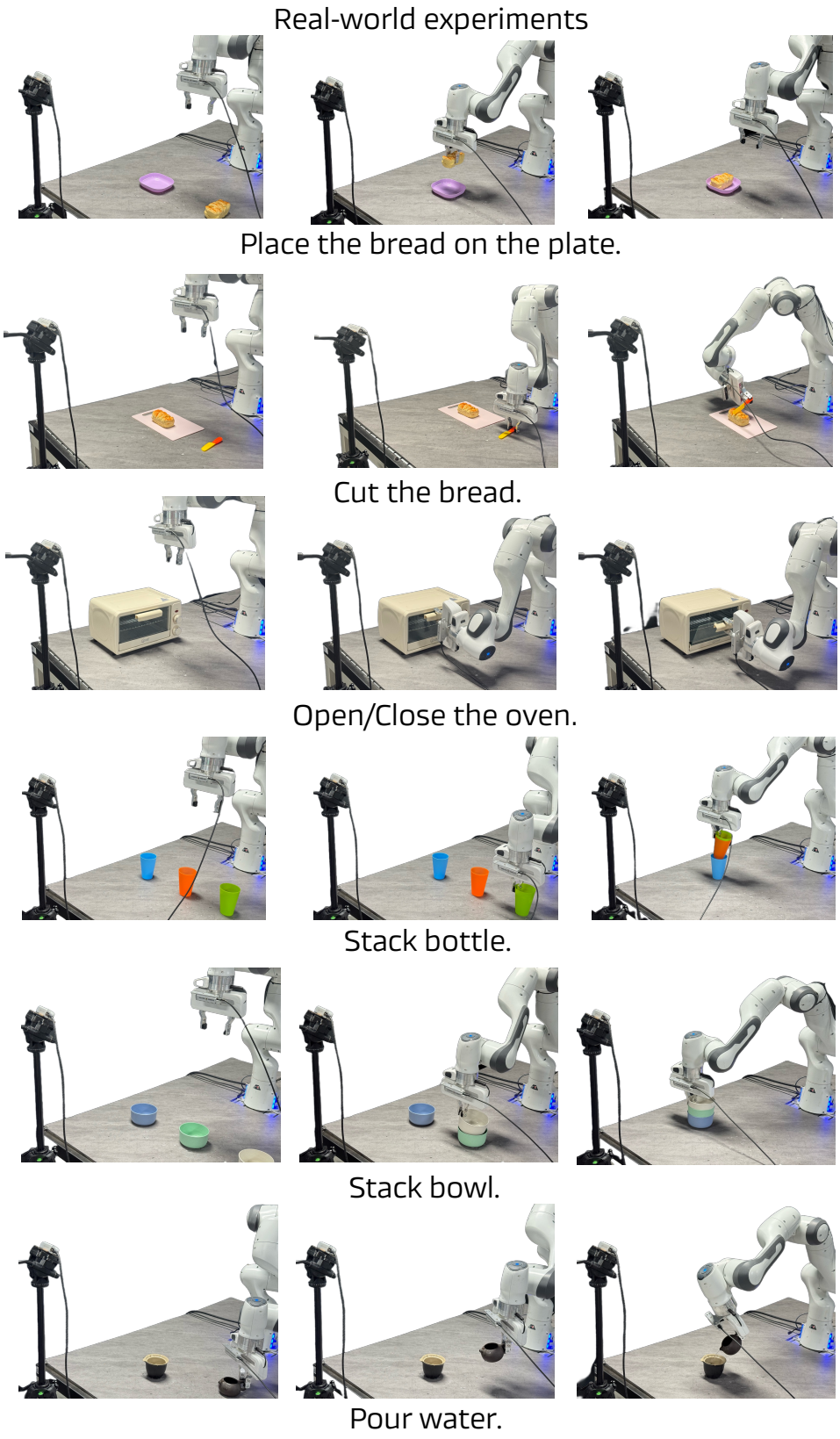


Figure 17: Qualitative results of real-world experiments.