CoT-VLA: Visual Chain-of-Thought Reasoning for Vision-Language-Action Models

Qingqing Zhao[1,2,*] Yao Lu[2] Moo Jin Kim[1] Zipeng Fu[1]
Zhuoyang Zhang[3] Yecheng Wu[2] Zhaoshuo Li[2] Qianli Ma[2] Song Han[2] Chelsea Finn[1]
Ankur Handa[2] Ming-Yu Liu[2] Donglai Xiang[2†] Gordon Wetzstein[1†] Tsung-Yi Lin[2†]

[1]Stanford University [2]NVIDIA [3]MIT

Abstract

Vision-language-action models (VLAs) have shown potential in leveraging pretrained vision-language models and diverse robot demonstrations for learning generalizable sensorimotor control. While this paradigm effectively utilizes large-scale data from both robotic and non-robotic sources, current VLAs primarily focus on direct input–output mappings, lacking the intermediate reasoning steps crucial for complex manipulation tasks. As a result, existing VLAs lack temporal planning or reasoning capabilities. In this paper, we introduce a method that incorporates explicit visual chain-of-thought (CoT) reasoning into vision-language-action models (VLAs) by predicting future image frames autoregressively as visual goals before generating a short action sequence to achieve these goals. We introduce CoT-VLA, a state-of-the-art 7B VLA that can understand and generate visual and action tokens. Our experimental results demonstrate that CoT-VLA achieves strong performance, outperforming the state-of-the-art VLA model by 17% in real-world manipulation tasks and 6% in simulation benchmarks. Videos are available at: https://cot-vla.github.io/.

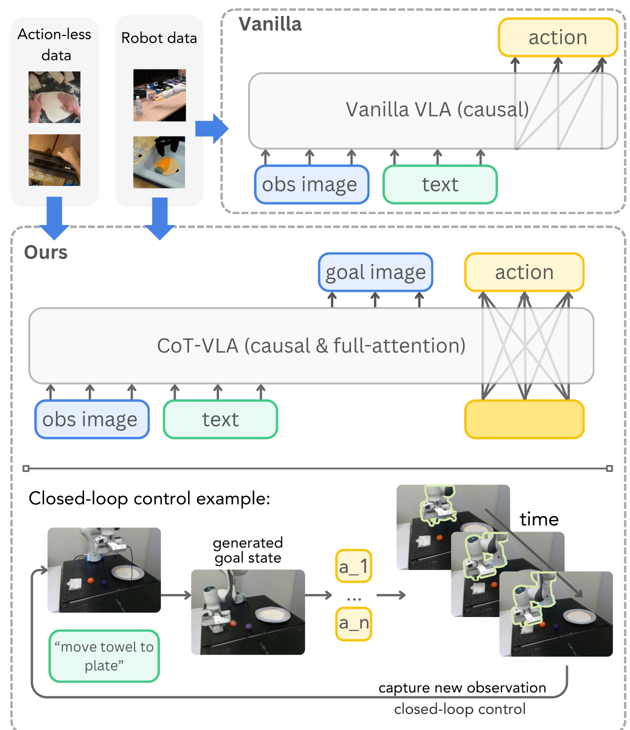

Figure 1. **Comparison between vanilla VLA and CoT-VLA frameworks**. Prior VLA models (top) directly predict robot actions from task inputs without explicit reasoning steps and only use action-annotated robot demonstration data for training. Unlike vanilla VLAs, CoT-VLA (bottom) can also leverage action-less datasets like EPIC-KITCHEN-100 [24] to enhance subgoal image generation ability, unlocking the potential of using abundant unlabeled video data to improve VLA's visual reasoning capability. CoT-VLA first generates a subgoal image as an intermediate reasoning step, and then generate a short action sequence to achieve the subgoal. We outline the robot arm for better visualization.

1. Introduction

Recent advances in robot learning have demonstrated impressive progress in training policies that can act across diverse tasks and environments [1, 4, 10, 12, 15, 26, 33, 39, 40, 43, 48, 53, 57, 60, 64, 70, 72]. One promising direction is vision-language-action (VLA) models, which leverage the rich understanding capabilities of pretrained vision-language models (VLMs) to map natural language instructions and visual observations to robot actions [10, 26, 43]. By training VLMs on robot demonstrations, VLAs inherit their ability to understand diverse scenes, objects, and natural language instructions, leading to better generalization capabilities when fine-tuned for downstream testing scenarios. While these approaches have shown impressive results, they typically map

*Work done during an internship at NVIDIA.
†Equal Advising

directly from observations to actions without explicit intermediate reasoning steps that could improve interpretability and, potentially, performance.

In the language domain, chain-of-thought (CoT) prompting has emerged as a powerful technique for improving the reasoning capabilities of large language models (LLMs) by encouraging step-by-step thinking [56, 69]. Applying these concepts to robotics presents exciting opportunities for grounding reasoning in text, visual observations, and physical actions. Recent works have made progress in this direction by incorporating intermediate reasoning steps like language descriptions, keypoints, or bounding boxes [13, 39, 40, 57]. These intermediate representations capture abstracted states of scenes, objects, and tasks and often require additional pre-processing pipelines. In our work, we explore subgoal images as an intermediate reasoning step before action generation. These images capture the state of the model's reasoning process and are naturally available within robot demonstration datasets. While prior work has explored subgoal generation and goal-conditioned imitation learning [2, 9, 41, 49], to the best of our knowledge, our approach is the first to integrate these concepts with VLAs as intermediate chain-of-thought reasoning steps.

We propose visual chain-of-thought reasoning for VLAs, a new method that uses subgoal image generation as a form of chain-of-thought reasoning for robotic tasks. Rather than directly predicting actions, our method first generates a subgoal image that represents the robot's planned state in pixel space, and then conditions its action on both the current observation and the generated subgoal image. This approach allows the model to "think visually" about how to accomplish a task before acting. By using the subgoal image as intermediate reasoning step, we leverage information that already exists in robot manipulation data with minimal preprocessing required. Furthermore, since subgoal image generation does not require action annotations, this unlocks the potential of using abundant video data for improved visual reasoning and understanding.

We build our CoT-VLA system that leverages visual chain-of-thought reasoning upon recent advances in unified multimodal foundation models that can understand and generate text and images [36, 52, 55, 61, 63]. We train our base model [61] on both the Open X-Embodiment dataset [43] and action-less video datasets [17, 24], and then fine-tune the model on task demonstrations collected on downstream robot setups used for deployment and evaluation. We design a hybrid attention mechanism for CoT-VLA: we use causal attention with next-token prediction for text and image generation, and leverage full attention to predict all action dimensions at once. Additionally, inspired by recent advances in robot learning [8, 14, 71], we predict sequences of actions (action chunking) rather than a single action at each timestep. We demonstrate that both action chunking and the hybrid attention mechanism improve the model's performance.

Through extensive experiments in both simulation benchmarks [34] and real-world experiments[43, 54], we demonstrate that our visual chain-of-thought reasoning helps improve policy performance compared to prior VLA approaches. Our key contributions include:

- We introduce a method of visual chain-of-thought reasoning through subgoal image generation as an intermediate reasoning step for robotic control.
- We introduce a system CoT-VLA that incorporates visual chain-of-thought reasoning, and a hybrid attention mechanism that combines causal attention for pixel and text generation and full attention for action prediction.
- We conduct comprehensive evaluations in both simulation and the real world, demonstrating that visual chain-of-thought reasoning improves VLA performance, and our system achieves state-of-the-art performance across multiple robot platforms and tasks.

2. Related Work

Chain-of-Thought (CoT) Reasoning CoT reasoning has gained prominence in natural language processing, particularly for enabling models to perform complex, multi-step reasoning tasks by breaking down problem-solving into sequential, explainable steps. Early work on CoT reasoning [56] has demonstrated the effectiveness of prompting large language models to generate intermediate reasoning steps before arriving at a final answer. Extending this paradigm to the visual domain, researchers have explored multimodal chain-of-thought methods, where visual information is processed iteratively in a stepwise fashion to reason about future outcomes or states, including generating bounding boxes [47], intermediate image infillments using Stable Diffusion [45] or standard Python packages [21], or generating CLIP embeddings [19]. Recently, CoT reasoning has been explored in embodied applications. It can generate textual plans for multi-stage execution [39, 40], point trajectories [57], label bounding boxes of objects and gripper positions as additional observations [39], generate future image trajectories for open-loop following [32, 42], and generate fine-grained reward guidance for reinforcement learning [70]. In this work, we introduce Visual-CoT reasoning for robotic manipulation, where predicted subgoal images serve as intermediate reasoning steps for closed-loop action generation. This approach leverages demonstration videos as natural intermediate reasoning states without requiring additional annotations.

Vision-Language-Action Models Large pretrained vision-language models (VLMs) [7, 25, 35] have emerged as a powerful tool for robot learning, and recent works have explored various approaches to integrate them into robot systems. Several works utilize VLMs as intermediate

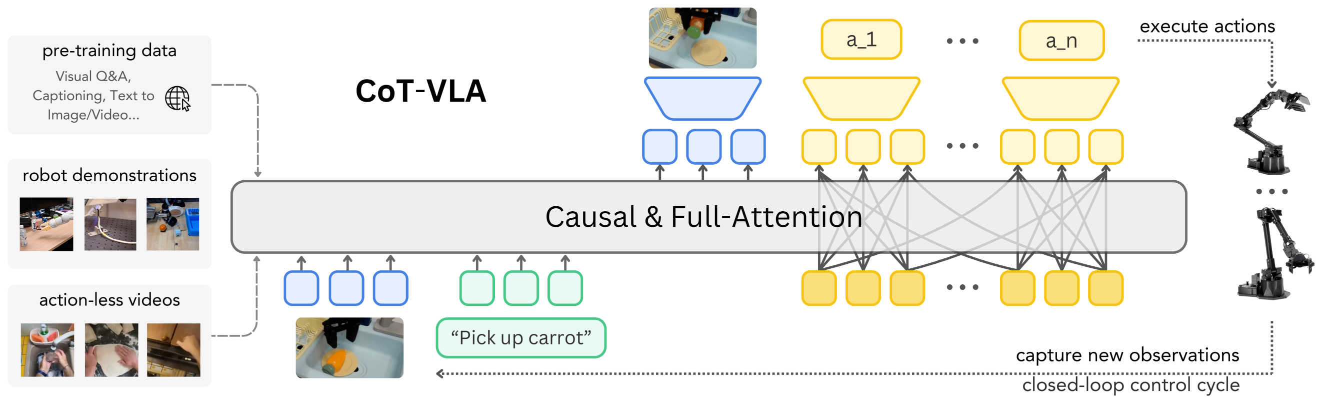

Figure 2. **Overview of CoT-VLA framework**. We build our model on VILA-U [61], a generative multimodal model pretrained on interleaved text–image data. The base model then trains on robot demonstrations [43] and action-less videos [17, 24]. During deployment, given a visual observation and a text instruction, the model performs visual chain-of-thought reasoning by generating a subgoal image (upper blue) with causal attention. It then generates a short action sequence with full attention ($\mathbf{a}_1 \cdots \mathbf{a}_n$) for robot execution. The system operates in a closed-loop control manner by capturing new observations after executing predicted action sequences.

components for perception and control, leveraging their strong semantic understanding and reasoning capabilities to decompose complex tasks [18, 23, 31, 50], detect objects [16, 22], or generate dense rewards [11, 37, 68] or goals [2, 9, 13, 41, 42, 49, 65, 74]. Some approaches incorporate VLMs [3, 44, 58] into end-to-end trainable policies by using them as pretrained backbone for better visuo-language representation [10, 26, 38, 43, 53]. Most relevant to our work are recent approaches that fine-tune pretrained VLMs on robot demonstration data for direct action prediction [10, 26, 43]. These VLAs demonstrate improved generalization to novel objects, environments, and natural language instructions through pretraining on internet-scale vision-language datasets, providing a promising direction for transferring visual and language knowledge to robotic control tasks. However, most existing VLAs do not leverage the step-by-step reasoning capabilities demonstrated in large language models, which have been shown to significantly improve performance across various tasks [56]. In the past, researchers have used chain-of-thought reasoning on language instructions or intermediate keypoints/bounding boxes for robotics [5, 11, 40, 57]. We introduce visual chain-of-thought reasoning to the VLA frameworks, using subgoal images as intermediate reasoning steps before action generation.

3. CoT-VLA

In this section, we present our visual chain-of-thought reasoning framework for VLAs. We begin with the formulation of our method (3.1), followed by a detailed description of the system architecture (3.2). We then explain our training procedures (3.3) and outline the deployment strategy for downstream tasks (3.3).

3.1. Visual Chain-of-Thought Reasoning

We consider two types of training data for VLA pretraining: robot demonstrations dataset D_r and action-less videos dataset D_v. Robot demonstrations are represented as $D_r = \{(l, \mathbf{a}_{1...T}, \mathbf{s}_{1...T})\}$, where l denotes the natural language instruction, $\mathbf{a}_{1...T} = \{\mathbf{a}_1, ..., \mathbf{a}_T\}$ denotes the sequence of robot actions, and $\mathbf{s}_{1...T} = \{\mathbf{s}_1, ..., \mathbf{s}_T\}$ denotes the visual observations as a sequence of images. Action-less videos $D_v = \{(l, \mathbf{s}_{1...T})\}$ consist of language descriptions and images without action annotations.

VLA: Vanilla VLA approaches fine-tune a pretrained VLM, P_θ, on D_r, learning to predict actions $\hat{\mathbf{a}}_{t+1}$ directly from the current observation $\mathbf{s}_t$ and language instruction l (Figure 1, top):

$$\hat{\mathbf{a}}_t \sim P_\theta(\mathbf{a}_t | \mathbf{s}_t, l) \tag{1}$$

CoT-VLA: Our key insight is to incorporate explicit visual reasoning before action generation. As illustrated in Figure 2, our approach operates in two sequential phases:

$$\hat{\mathbf{s}}_{t+n} \sim P_\theta(\mathbf{s}_{t+n} | \mathbf{s}_t, l) \tag{2}$$

$$\{\hat{\mathbf{a}}_t, ..., \hat{\mathbf{a}}_{t+m}\} \sim P_\theta(\{\mathbf{a}_t, ..., \mathbf{a}_{t+m} | \mathbf{s}_t, l, \mathbf{s}_{t+n}) \tag{3}$$

where we first predict a subgoal image $\hat{\mathbf{s}}_{t+n}$, n frames ahead, as an intermediate visual reasoning step (Equation 2). Then we generates a sequence of m actions to achieve this subgoal state (Equation 3). This enables the model to "think visually" first by explicitly reasoning about desired future states before predicting the actions. The visual reasoning step, Equation (2), is trained on both robot demonstrations D_r and action-less videos D_v, and the action generation step, Equation (3), is trained on robot demonstrations D_r only.

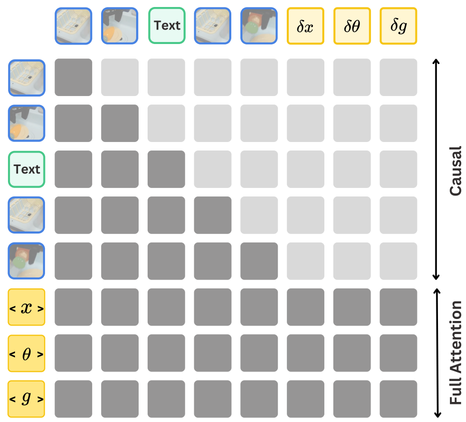

Figure 3. **Hybrid attention mechanism in CoT-VLA**. We use causal attention for image or text generation and full attention for action generation. $[x]$, $[\theta]$ and $[g]$ are special tokens for parallel decoding of actions.

3.2. The Base Vision-Language Model

To enable the visual reasoning capabilities described in Equation (2), we build upon VILA-U [61], an unified multimodal foundation model capable of both understanding and generating image and text tokens.

VILA-U unifies video, image, and language understanding through an autoregressive next-token prediction framework. At its core is a unified vision tower that encodes visual inputs as discrete tokens aligned with textual information. This enables autoregressive image and video generation while significantly enhancing the understanding capabilities of VLMs that leverage discrete visual features. VILA-U utilizes residual quantization [29] to improve the representational capacity of discrete visual features - incorporating a depth transformer, as introduced in RQ-VAE [29], to gradually predict the residual tokens. The extracted visual features are then passed through a projector before being processed by the LLM backbone. The base model is trained on multimodal pairs including [image, text], [text, image], [video, text], and [text, video]. We use the VILA-U model trained on 256×256 resolution images, where each image is encoded into $16 \times 16 \times 4$ tokens with a residual depth of 4 [29]. For detailed information about VILA-U training and architecture, we refer readers to [61].

3.3. Training Procedures

We pretrain the base 7B VILA-U model on a combination of robot demonstrations D_r and action-less videos D_v. During training, we optimize three components, the LLM backbone, projector, and depth transformer, while keeping the vision tower fixed. Our training objective has two key components: subgoal image generation with causal attention (2) and action generation with full attention (3).

Visual Tokens Prediction For subgoal image generation, each training sequence is of form $(l, \mathbf{s}_t, \mathbf{s}_{t+n})$. We follow the training objective used in [61]. At each visual position j, the depth transformer, P_δ, autoregressively predicts D residual tokens $(k_{j1}, ..., k_{jD})$ based on the LLM-generated code embedding h_j. The training objective for visual tokens is then formulated as:

$$\mathcal{L}_{\text{visual}} = -\sum_{j} \sum_{d=1}^{D} \log P_\delta(k_{jd} | k_{j,<d}) \tag{4}$$

where j indexes the positions containing visual tokens. For a more detailed explanation of this loss function, we refer readers to [61], Section 3.2.

Action Tokens Prediction For action prediction, each training sequence takes the form $(l, \mathbf{s}_t, \mathbf{s}_{t+n}, \mathbf{a}_t, ..., \mathbf{a}_{t+m})$. Each action $\mathbf{a}_i$ is represented by 7 tokens, with each action dimension independently discretized. Following [26], we map each continuous action dimension into 256 discrete bins, with bin widths determined by uniformly dividing the interval between the 1st and 99th percentiles of the training data's action distribution. We repurpose the 256 least frequently used tokens in the text tokenizer's vocabulary as action bin tokens. Unlike prior works [10, 26, 43], we employ full attention for processing and predicting action tokens, enabling all action tokens to interact with each other. This attention mechanism is illustrated in Figure 3. During training, we minimize the cross-entropy loss for action predictions:

$$\mathcal{L}_{\text{action}} = -\sum_{i=1}^{m} \log P_\theta(\mathbf{a}_t ... \mathbf{a}_{t+m} | l, s_t, s_{t+n}) \tag{5}$$

Given a batch of input sequences, The overall training objective combines the action and visual losses:

$$\mathcal{L} = \mathcal{L}_{\text{action}} + \mathcal{L}_{\text{visual}} \tag{6}$$

Pretraining Phase We pretrain CoT-VLA on both robot demonstrations D_r and action-less videos D_v as described in Section 3.1. For robot demonstrations, we curate a subset of the Open X-Embodiment dataset [43] (OpenX). Following the pre-processing pipeline established in OpenVLA [26], we select and process datasets with third-person camera views and single-arm end-effector control (7-DoF). For action-less videos D_v, we incorporate the EPIC-KITCHENS [24] and Something-Something V2 [17] datasets. All images are processed at 256×256 resolution. For visual reasoning, we use subgoal images at future timestep n uniformly sampled from a dataset-specific range $[n_l, n_u]$, where n_l and n_u define the lower and upper bounds of the prediction horizon. We use an action chunk size of 10. For complete dataset specifications and training hyperparameters, please refer to the supplementary material.

	Average (↑)	Spatial (↑)	Object (↑)	Goal (↑)	Long (↑)
Diffusion Policy	$72.4 \pm 0.7\%$	$78.3 \pm 1.1\%$	$\mathbf{92.5 \pm 0.7\%}$	$68.3 \pm 1.2\%$	$50.5 \pm 1.3\%$
Octo fine-tuned	$75.1 \pm 0.6\%$	$78.9 \pm 1.0\%$	$85.7 \pm 0.9\%$	$84.6 \pm 0.9\%$	$51.1 \pm 1.3\%$
OpenVLA fine-tuned	$76.5 \pm 0.6\%$	$84.7 \pm 0.9\%$	$88.4 \pm 0.8\%$	$79.2 \pm 1.0\%$	$53.7 \pm 1.3\%$
CoT-VLA-7B (ours)	$\mathbf{81.13 \pm 0.6}$ **%**	$\mathbf{87.5 \pm 1.4\%}$	$91.6 \pm 0.5\%$	$\mathbf{87.6 \pm 0.6\%}$	$\mathbf{69.0 \pm 0.8\%}$

Table 1. **LIBERO benchmark experimental results.** For each task suite (Spatial, Object, Goal, Long), we report the average success rate and standard error across 3 seeds with 500 episodes each. CoT-VLA achieves the best or competitive performance across all LIBERO benchmarks suites compared to baseline approaches. The bolded entries correspond to highest success rates while underlined entries correspond to second-highest.

Adaptation Phase for Downstream Closed-Loop Deployment For adaptation to downstream tasks, we fine-tune our pretrained model using task-specific robot demonstration data D_r collected on the target robot setups. During this phase, we optimize the LLM backbone, projector, and depth transformer while keeping the vision tower frozen, maintaining the same training setup as the pretraining stage. The resulting model can execute new manipulation tasks based on natural language commands l. Algorithm 1 describes our robot control procedure at test time.

Algorithm 1 CoT-VLA test-time closed-loop control

Require: CoT-VLA Model P_θ, initial state s_0^{obs}, language instruction l

0: $t \leftarrow 0$
0: **while** True **do**
0: sample $\hat{\mathbf{s}}_{t+n} \sim P_\theta(\mathbf{s}_{t+n} \mid l, \mathbf{s}_t^{\text{obs}})$
0: sample $[\hat{\mathbf{a}}_t, .., \hat{\mathbf{a}}_{t+m}] \sim P_\theta(\mathbf{a}_t, .., \mathbf{a}_{t+m} \mid l, \mathbf{s}_t^{\text{obs}}, \mathbf{s}_{t+n})$
0: **for** $j = 0$ to m **do**
0: execute $\hat{\mathbf{a}}_{t+j}$
0: **end for**
0: $t \leftarrow t + m + 1$
0: $\mathbf{s}_t^{\text{obs}} \leftarrow$ robot observation

4. Experiments

We evaluate the effectiveness of our approach and our system through a set of experiments spanning both simulation benchmarks and real-world robot manipulation tasks. Our experiments aim to addresses following questions:

- How does our system perform compared to state-of-the-art baselines across multiple benchmarks and embodiments? (Section 4.2)
- What is the impact of pretraining, visual chain-of-thought reasoning and hybrid attention on task performance? (Section 4.3)
- To what extent does improved generalization in visual reasoning enhance the action prediction capabilities? (Section 4.4)

4.1. Experimental Setup

We conduct evaluations across three complementary settings: the LIBERO benchmark [34] for evaluation in simulation environments, the Bridge-V2 platform [54] with its dataset of 45k robot demonstrations, and the Franka-Tabletop setup with a stationary, table-mounted Franka Emika Panda robot with limited 10 to 150 robot demonstrations for each testing scenario.

LIBERO Simulation Benchmark We perform evaluation on LIBERO [34], a simulation benchmark comprising four distinct task suites: LIBERO-Spatial, LIBERO-Object, LIBERO-Goal, and LIBERO-Long. Each suite contains 10 diverse tasks with 50 human-teleoperated demonstrations per task, aiming to evaluate the robot's comprehension of spatial relationships, object interactions, and task-specific objectives. We follow the same preprocessed pipeline as in [26]: (1) removing pause intervals from trajectories, (2) standardizing image resolution to 256×256 pixels, and (3) applying a 180-degree rotation to all images.

Bridge-V2 Real-Robot Experiments We use a 6-DoF WidowX robotic arm, following the experimental setup from Bridge-V2 [54]. Our training data has 45k language-annotated trajectories from the Bridge-V2 dataset, encompassing diverse manipulation tasks. While the dataset was incorporated into the pretraining phase alongside OpenX, we performed additional task-specific fine-tuning exclusively on Bridge-V2 until achieving a training action prediction accuracy threshold of 95%. Following [26], we evaluate on four tasks designed in [26] to evaluate visual robustness (varying distractors), motion generalization (novel object positions), semantic generalization (unseen language instruction), and language grounding (instruction following).

Franka-Tabletop Real-Robot Experiments We use a stationary, table-mounted Franka Emika Panda 7-DoF robot arm denoted as Franka-Tabletop. The setup is not seen during the pretraining stage and is designed to assess our model's adaptation capability to novel real-world environments with small amounts of robot demonstrations. We

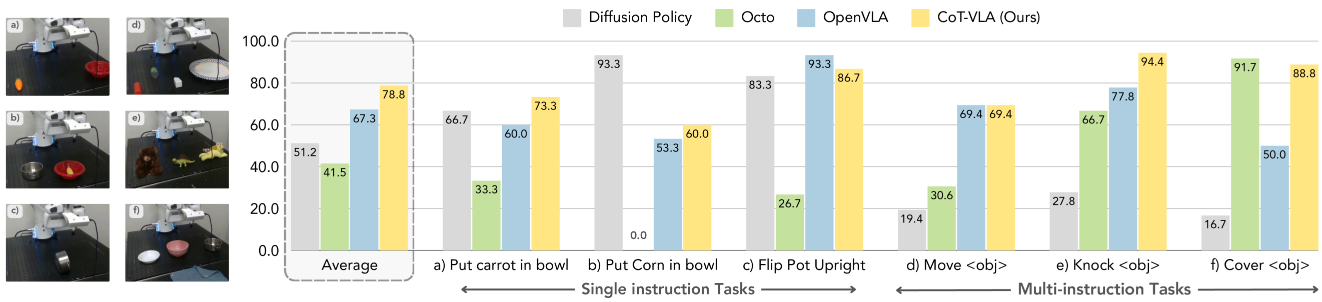

Figure 4. **Franka-Tabletop comparisons.** Evaluation across six distinct manipulation tasks, with separate models trained per task. Left: Representative initial states for each task setup. Right: Task-specific success rates and cross-task averages for our method and baselines. CoT-VLA achieves best average performance and demonstrates strong capabilities in both single-instruction and multi-instruction scenarios.

perform evaluations across 6 tasks: 3 narrow domain single-instruction tasks for and 3 diverse multi-instruction tasks outlined in Figure 4 and introduced in [26]. For each task, the dataset contains between 10 and 150 demonstrations.

Baselines We evaluate our approach against four state-of-the-art baselines. **Diffusion Policy** [8], a state-of-the-art imitation learning algorithm, is trained from scratch for each test scenario in LIBERO and Franka-Tabletop. The implementation incorporates action chunking and proprioception while conditioning on DistilBERT [46] language embeddings. **OpenVLA** [26] is an open-source VLA model that fine-tunes pretrained vision-language models on the OpenX dataset; and **Octo** [53] is a generalist model pretrained on OpenX without VLM initialization. For both OpenVLA and Octo, we use their published checkpoints for Bridge-V2 evaluations and fine-tune them for our LIBERO and Franka-Tabletop experiments. **SUSIE** [2], a two-stage approach, combines instruction-guided image editing for goal generation with a goal-conditioned policy for action generation. We evaluate SUSIE using their published checkpoint on Bridge-V2.

4.2. Evaluations Results

LIBERO We present quantitative results in Table 1, where each method is evaluated over 500 trials per task suite, with 3 random seeds. Success rates are reported with means and standard error. Qualitative examples of our method's reasoning and execution trajectories are illustrated in Figure 5. Results demonstrate that CoT-VLA effectively adapts to tasks in the LIBERO simulation environment, achieving best or competitive performance compared to baseline approaches. By analyzing rollout videos of failure cases, we found that baseline methods occasionally overfit to visual cues while disregarding language instructions. Specifically, when initial states appear visually similar across different tasks (e.g., in LIBERO-Spatial), baseline methods execute a different task compared to the commanded task in some episodes. CoT-VLA exhibits better instruction following ability by first reasoning visually about the desired actions via language-grounded subgoal generation, and then predicting the relevant actions for achieving the goal.

Bridge-V2 We evaluate CoT-VLA and baselines on the Bridge-V2 benchmark across four generalization categories identified in [26]: visual generalization ("put eggplant into pot" with cluttered environments), motion generalization ("put carrot on plate" with height variations), semantic generalization ("take purple grapes out of pot"), and language grounding ("put eggplant or red bottle into pot"). We report the quantitative results in Table 2, where each task is tested with 10 trials. SUSIE [2] generates visually higher-quality goal images through its diffusion prior (see Section 5 for a detailed discussion on our limitations) but achieves lower success rates on tasks involving novel objects or requiring complex language grounding. Compared to OpenVLA [26], CoT-VLA shows slightly lower success rates in visual and language generalization tasks due to grasping failures from action chunking (see Section 5) rather than errors in visual reasoning. However, CoT-VLA demonstrates competitive performance across all four generalization categories overall, achieving comparable or better results to baseline approaches.

Franka-Tabletop We present quantitative results in Table 4 and example execution trajectories in Figure 5. In this experiment, models are fine-tuned on a relatively small set of demonstrations. While Diffusion Policy achieves top performance on single-instruction tasks (e.g., "put corn in bowl"), its performance degrades on multi-instruction tasks involving diverse objects and complex language instructions. Models pretrained on the OpenX dataset - Octo, OpenVLA, and CoT-VLA - demonstrate better adaptation and performance on multi-instruction tasks where language grounding is critical. Overall, CoT-VLA achieves the highest average performance compared to baseline approaches, showing improvements in both single and multi-instruction scenarios.

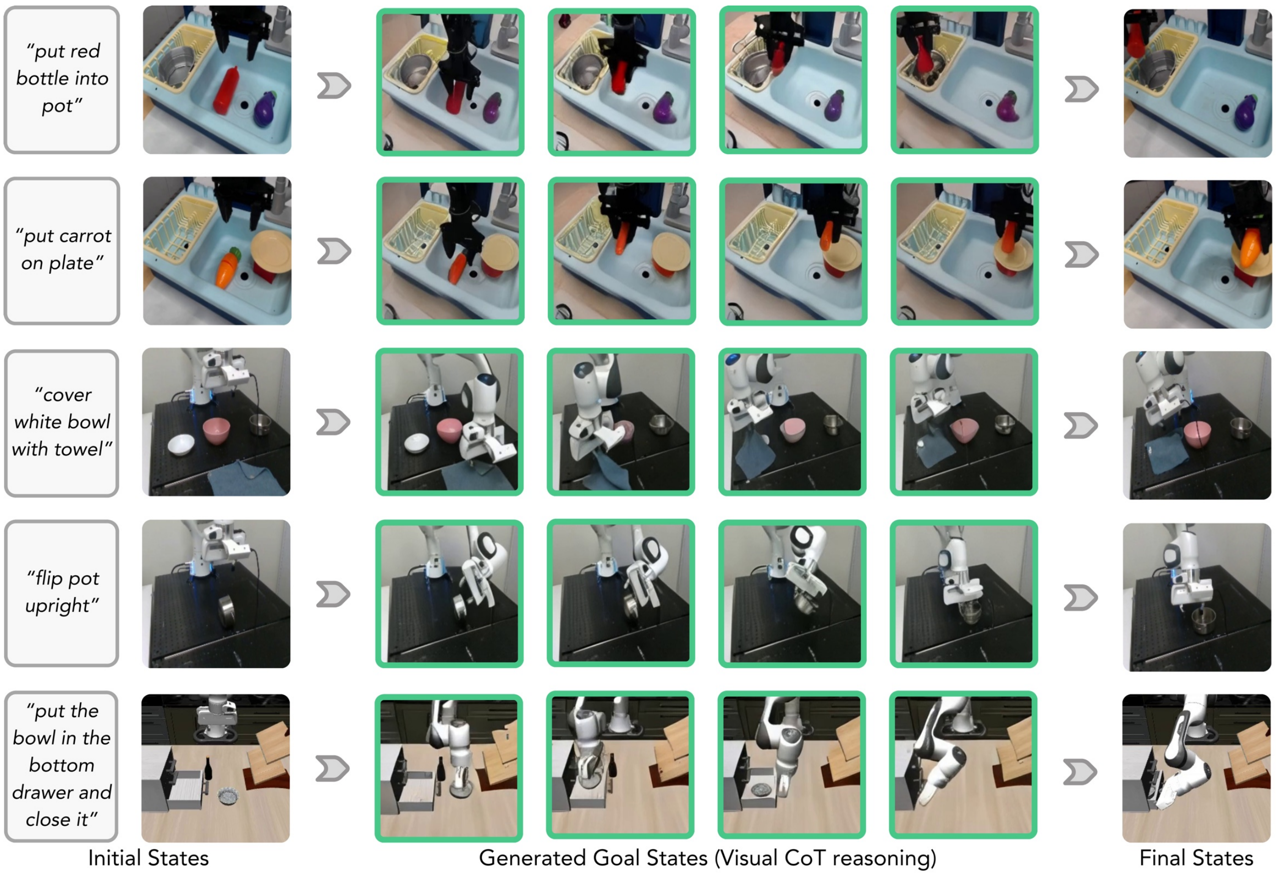

Figure 5. **Task execution examples for LIBERO, Bridge-V2, and Franka-Tabletop using CoT-VLA.** For each task: Left: text instruction (l) and initial state (s_0^{obs}). Middle: generated intermediate goal states ($\hat{s}_t$) demonstrating visual chain-of-thought reasoning, where each goal image is conditioned on both the instruction and the most recent observation. Right: final state (s_T^{obs}) upon task completion. Complete execution trajectories are available in the supplementary video.

4.3. Ablation Study

Visual CoT, Hybrid Attention, and Action Chunking We conduct comprehensive ablation studies on two LIBERO benchmark suites: LIBERO-Spatial and LIBERO-Goal. We evaluate four model variants: **VLA** - a baseline implementation following the standard VLA framework [26], with the same VILA-U backbone but without chain-of-thought reasoning and action chunking; **+ action chunking** - extending the vanilla VLA to predict action sequences of length m; **+ hybrid attention** - further adding full attention mechanisms for action sequence prediction, as illustrated in Figure 3; and **+ CoT (ours)**: our complete approach with hybrid attention mechanism and visual chain-of-thought reasoning.

As shown in Figure 6, both benchmark suites demonstrate that action sequence prediction consistently outperforms single-action prediction. The addition of hybrid attention mechanisms further improves performance. Our CoT-VLA achieves the best results validating the effectiveness of visual chain-of-thought reasoning for VLA tasks.

Category	SUSIE	Octo	OpenVLA	CoT-VLA
Visual	30%	35%	**75%**	65%
Motion	10%	10%	45%	**60%**
Semantic	20%	0%	40%	**50%**
Language	40%	40%	**75%**	70%

Table 2. **Bridge-V2 Comparison.** Success rates across four generalization categories, with 10 trials per category and partial credit scoring following [26]. **Visual**: "put eggplant into pot" with cluttered environments; **Motion**: "put carrot on plate" with height variations; **Semantic**: "take purple grapes out of pot"; **Language**: "put eggplant or red bottle into pot".

Pretraining Our training pipeline has two stages, pretraining VILA-U on the OpenX dataset augmented with actionless video data (Section 3.3), and task-specific post-training on robot demonstration data. To assess the importance of our pretraining stage, we conduct ablation studies on the Franka-Tabletop setup. We report the quantitative results

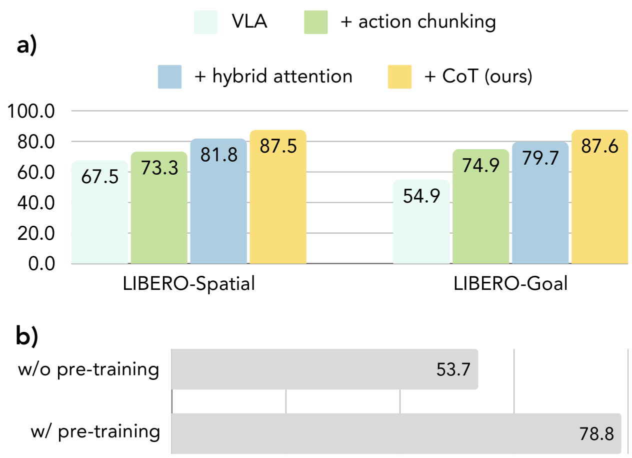

Figure 6. **Ablation studies of CoT-VLA components.** a) Results on LIBERO-Spatial and LIBERO-Goal benchmarks demonstrate the effectiveness of three components: action chunking, hybrid attention, and visual chain-of-thought reasoning. b) Pretraining ablation experiments on Franka-Tabletop show performance improvements from the OpenX and action-less video pretraining process.

in Figure 6. Our results show that CoT-VLA with our pretraining stage achieves a 46.7% relative improvement, from 53.7% to 78.8%, compared to directly fine-tuning the base VILA-U model on Franka-Tabletop demonstrations, demonstrating better downstream task adaptation.

4.4. Better Visual Reasoning Helps

Unlike prior VLAs that only use robot demonstration data D_r during training, CoT-VLA also leverages action-less video data D_v for pretraining through its intermediate visual chain-of-thought reasoning steps. This enables learning of both dynamics and instruction following from captioned videos alone, which are substantially more abundant than robot demonstrations. To investigate how visual reasoning capabilities transfer to robot performance, we conduct an ablation study on the Franka-Tabletop setup using novel, long-horizon tasks that combine two unseen subtasks. We design two tasks – (1) "move the green scallion to the apple-covered book" and (2) "move the green cauliflower to the bear-covered book" – that are challenging for our model's out-of-distribution generalization. For each task, we collect one demonstration trajectory to obtain ground-truth goal images. We evaluate each task across 5 trials under two conditions: (1) CoT-VLA using its generated goal images and (2) CoT-VLA using ground-truth goal images from the collected demonstrations. As shown in Table 3, using ground-truth goal images improves the absolute success rate by 40% for both tasks. This performance boost suggests that advances in visual reasoning and goal image generation could directly translate to better robotic task performance. While our method still struggles with out-of-distribution subgoal generation, recent advances in large-scale video and image models show promising directions for improving visual reasoning capabilities with scaling.

	Sub-task 1	Sub-task 2
Generated Goal Images	20%	0%
Ground-truth Goal Images	**60%**	**40%**

Table 3. **Better visual reasoning helps.** Success rates comparing CoT-VLA using generated versus ground-truth goal images on out-of-distribution tasks. Results demonstrate that improved visual reasoning (simulated by ground-truth goals) leads to better task performance, suggesting that advances in goal generation can translate to improved action execution.

5. Conclusion, Limitations and Future Work

In this work, we introduce CoT-VLA, bridging vision-language-action models with chain-of-thought reasoning by introducing intermediate visual goals as explicit reasoning steps. Rather than using abstract representations like bounding boxes or keypoints, we propose using subgoal images sampled from videos as an interpretable and effective intermediate representation. We build our system upon VILA-U, demonstrating strong performance across diverse robotic manipulation tasks.

While our approach demonstrates effectiveness, there are certain limitations. First, generating intermediate image tokens during inference introduces significant computational overhead compared to direct action generation approaches. Our method requires generating 256 image tokens before action tokens, leading to a 7× slowdown on average with an action chunk size of 10. While action chunking and parallel decoding improve inference speed, image generation remains the primary bottleneck. Recent advancement in fast image generation or fast LLM inference techniques could potentially improve the throughput of the model [6, 28, 30, 51, 67] and be integrated into our system. Second, our autoregressive image generation produces lower visual quality compared to state-of-the-art diffusion-based models. Recent advances in unified multimodal models [55, 59, 63, 73] suggest promising directions for improvements. Additionally, while effective, our action chunking approach can introduce discontinuous actions between chunks and lacks high-frequency feedback during execution. These limitations could be addressed through temporal smoothing techniques and per-step prediction approaches similar to those proposed in [8]. Finally, while CoT-VLA leverages action-less video data during pretraining, current computational constraints limit its ability to achieve visual-reasoning generalization for entirely new tasks. Looking forward, we believe recent advances in video/image generation and world models [13, 20, 27, 62, 66] present promising opportunities to enhance generalization capabilities through improved visual reasoning and predictive modeling.

References

[1] Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. *arXiv preprint arXiv:2409.16283*, 2024. 1

[2] 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 arXiv:2310.10639*, 2023. 2, 3, 6

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 3

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

[5] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14455–14465, 2024. 3

[6] Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, Yao Lu, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. *arXiv preprint arXiv:2410.10733*, 2024. 8

[7] Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, et al. Pali-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*, 2023. 2

[8] 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. 2, 6, 8

[9] Yiming Ding, Carlos Florensa, Pieter Abbeel, and Mariano Phielipp. Goal-conditioned imitation learning. *Advances in neural information processing systems*, 32, 2019. 2, 3

[10] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023. 1, 3, 4

[11] Yuqing Du, Ksenia Konyushkova, Misha Denil, Akhil Raju, Jessica Landon, Felix Hill, Nando de Freitas, and Serkan Cabi. Vision-language models as success detectors. *arXiv preprint arXiv:2303.07280*, 2023. 3

[12] Yilun Du, Mengjiao Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B Tenenbaum, et al. Video language planning. *arXiv preprint arXiv:2310.10625*, 2023. 1

[13] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 8

[14] Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. Humanplus: Humanoid shadowing and imitation from humans. *arXiv preprint arXiv:2406.10454*, 2024. 2

[15] Zipeng Fu, Tony Z. Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. In *Conference on Robot Learning (CoRL)*, 2024. 1

[16] Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23171–23181, 2023. 3

[17] 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 *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 2, 3, 4

[18] Huy Ha, Pete Florence, and Shuran Song. Scaling up and distilling down: Language-guided robot skill acquisition. In *Conference on Robot Learning*, pages 3766–3777. PMLR, 2023. 3

[19] William Harvey and Frank Wood. Visual chain-of-thought diffusion models. *arXiv preprint arXiv:2303.16187*, 2023. 2

[20] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023. 8

[21] Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *arXiv preprint arXiv:2406.09403*, 2024. 2

[22] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10608–10615. IEEE, 2023. 3

[23] Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. *arXiv preprint arXiv:2409.01652*, 2024. 3

[24] Georgios Kapidis, Ronald Poppe, Elsbeth Van Dam, Lucas Noldus, and Remco Veltkamp. Egocentric hand track and object-based human action recognition. In *2019 IEEE SmartWorld, Ubiquitous Intelligence & Com-*

puting, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), pages 922–929. IEEE, 2019. 1, 2, 3, 4

[25] 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 arXiv:2402.07865*, 2024. 2

[26] 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. 1, 3, 4, 5, 6, 7

[27] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023. 8

[28] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023. 8

[29] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11523–11532, 2022. 4

[30] Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, pages 19274–19286. PMLR, 2023. 8

[31] Boyi Li, Yue Wang, Jiageng Mao, Boris Ivanovic, Sushant Veer, Karen Leung, and Marco Pavone. Driving everywhere with large language model policy adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14948–14957, 2024. 3

[32] 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 arXiv:2406.16862*, 2024. 2

[33] Fanqi Lin, Yingdong Hu, Pingyue Sheng, Chuan Wen, Jiacheng You, and Yang Gao. Data scaling laws in imitation learning for robotic manipulation. *arXiv preprint arXiv:2410.18647*, 2024. 1

[34] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *arXiv preprint arXiv:2306.03310*, 2023. 2, 5

[35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2

[36] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26439–26455, 2024. 2

[37] Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Eureka: Human-level reward design via coding large language models. *arXiv preprint arXiv: Arxiv-2310.12931*, 2023. 3

[38] 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? *Advances in Neural Information Processing Systems*, 36:655–677, 2023. 3

[39] Zawalski Michał, Chen William, Pertsch Karl, Mees Oier, Finn Chelsea, and Levine Sergey. Robotic control via embodied chain-of-thought reasoning. *arXiv preprint arXiv:2407.08693*, 2024. 1, 2

[40] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 3

[41] Ashvin V Nair, Vitchyr Pong, Murtaza Dalal, Shikhar Bahl, Steven Lin, and Sergey Levine. Visual reinforcement learning with imagined goals. *Advances in neural information processing systems*, 31, 2018. 2, 3

[42] Fei Ni, Jianye Hao, Shiguang Wu, Longxin Kou, Jiashun Liu, Yan Zheng, Bin Wang, and Yuzheng Zhuang. Generate subgoal images before act: Unlocking the chain-of-thought reasoning in diffusion model for robot manipulation with multimodal prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 3

[43] 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 arXiv:2310.08864*, 2023. 1, 2, 3, 4

[44] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[45] Daniel Rose, Vaishnavi Himakunthala, Andy Ouyang, Ryan He, Alex Mei, Yujie Lu, Michael Saxon, Chinmay Sonar, Diba Mirza, and William Yang Wang. Visual chain of thought: bridging logical gaps with multimodal infillings. *arXiv preprint arXiv:2305.02317*, 2023. 2

[46] V Sanh. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 6

[47] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. *arXiv preprint arXiv:2403.16999*, 2024. 2

[48] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on robot learning*, pages 894–906. PMLR, 2022. 1

[49] Mohit Shridhar, Yat Long Lo, and Stephen James. Generative image as action models. *arXiv preprint arXiv:2407.07875*, 2024. 2, 3

[50] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11523–11530. IEEE, 2023. 3

[51] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023. 8

[52] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 2

[53] 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. 1, 3, 6

[54] 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. 2, 5

[55] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 2, 8

[56] 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*, 2022. 2, 3

[57] Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning. *arXiv preprint arXiv:2401.00025*, 2023. 1, 2, 3

[58] Junjie Wen, Yichen Zhu, Jinming Li, Minjie Zhu, Kun Wu, Zhiyuan Xu, Ran Cheng, Chaomin Shen, Yaxin Peng, Feifei Feng, et al. Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation. *arXiv preprint arXiv:2409.12514*, 2024. 3

[59] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024. 8

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

[61] Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. *arXiv preprint arXiv:2409.04429*, 2024. 2, 3, 4

[62] Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting Ning, Yuheng Zha, Zeyu Feng, Tianhua Tao, Shibo Hao, Yemin Shi, et al. Pandora: Towards general world model with natural language actions and video states. *arXiv preprint arXiv:2406.09455*, 2024. 8

[63] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 2, 8

[64] Jonathan Yang, Catherine Glossop, Arjun Bhorkar, Dhruv Shah, Quan Vuong, Chelsea Finn, Dorsa Sadigh, and Sergey Levine. Pushing the limits of cross-embodiment learning for manipulation and navigation. *arXiv preprint arXiv:2402.19432*, 2024. 1

[65] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023. 3

[66] Sherry Yang, Jacob Walker, Jack Parker-Holder, Yilun Du, Jake Bruce, Andre Barreto, Pieter Abbeel, and Dale Schuurmans. Video as the new language for real-world decision making. *arXiv preprint arXiv:2402.17139*, 2024. 8

[67] Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers, and Liang-Chieh Chen. An image is worth 32 tokens for reconstruction and generation. *arxiv: 2406.07550*, 2024. 8

[68] Wenhao Yu, Nimrod Gileadi, Chuyuan Fu, Sean Kirmani, Kuang-Huei Lee, Montse Gonzalez Arenas, Hao-Tien Lewis Chiang, Tom Erez, Leonard Hasenclever, Jan Humplik, et al. Language to rewards for robotic skill synthesis. *arXiv preprint arXiv:2306.08647*, 2023. 3

[69] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022. 2

[70] Kaifeng Zhang, Zhao-Heng Yin, Weirui Ye, and Yang Gao. Learning manipulation skills through robot chain-of-thought with sparse failure guidance. *arXiv preprint arXiv:2405.13573*, 2024. 1, 2

[71] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023. 2

[72] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024. 1

[73] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. 2024. 8

[74] Enshen Zhou, Yiran Qin, Zhenfei Yin, Yuzhou Huang, Ruimao Zhang, Lu Sheng, Yu Qiao, and Jing Shao. Mine-

dreamer: Learning to follow instructions via chain-of-imagination for simulated-world control. *arXiv preprint arXiv:2403.12037*, 2024. 3