ACoT-VLA: Action Chain-of-Thought for Vision-Language-Action Models

Linqing Zhong[1,2] Yi Liu[2] Yifei Wei[1,2] Ziyu Xiong[2]
Si Liu[1*] Guanghui Ren[2*]

[1]Beihang University [2]AgiBot

Abstract

*Vision-Language-Action models have emerged as essential generalist robot policies for diverse manipulation tasks, conventionally relying on directly translating multimodal inputs into actions via Vision-Language Model embeddings. Recent advancements have introduced explicit intermediary reasoning—such as sub-task prediction (language) or goal image synthesis (vision)—to guide action generation. However, these intermediate reasoning are often indirect and inherently limited in their capacity to convey the full, granular information required for precise action execution. Instead, we posit that the most effective form of reasoning is one that deliberates directly in the action space. We introduce **Action Chain-of-Thought (ACoT)**, a paradigm where the reasoning process itself is formulated as a structured sequence of coarse action intents that guide the final policy. In this paper, we propose ACoT-VLA, a novel architecture that materializes the ACoT paradigm. Specifically, we introduce two complementary components: an Explicit Action Reasoner (EAR) and Implicit Action Reasoner (IAR). The former proposes coarse reference trajectories as explicit action-level reasoning steps, while the latter extracts latent action priors from internal representations of multimodal input, co-forming an ACoT that conditions the downstream action head to enable grounded policy learning. Extensive experiments in real-world and simulation environments demonstrate the superiority of our proposed method. Code is available at: `https://github.com/AgibotTech/ACoT-VLA`.*

1. Introduction

To overcome the generalization limits of task-specific robot policies [11, 48, 62], recent work has converged on Vision-Language-Action (VLA) models [5, 24, 33, 41], which always leverage a pre-trained Vision-Language Model (VLM) [1, 2, 46] to encode visual and linguistic inputs into a latent representation that conditions an action decoder. Recent advancements seek to improve the mapping from the input space to the action space by introducing the intermediate reasoning step by language generation, leading to more generalized and precise action outputs [22, 54] (Fig. 1(a)). A parallel thrust uses world models [16, 51, 66] to simulate environmental dynamics, directly enhancing the efficacy and goal-oriented nature of the generated action sequences [59, 63], as visualized in Fig. 1(b).

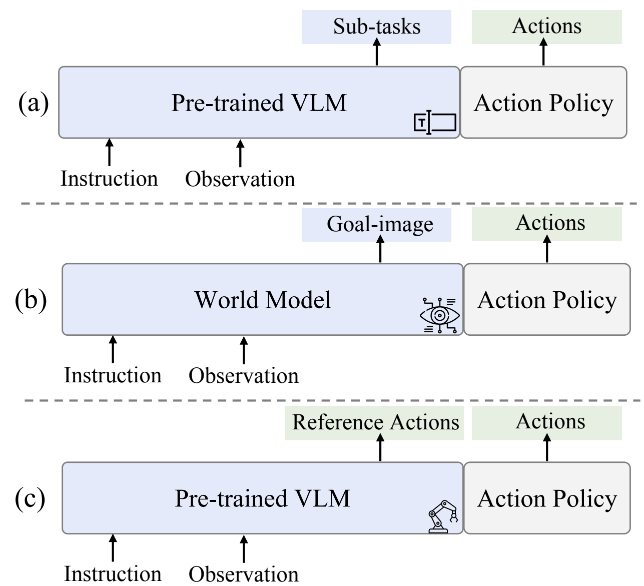

Figure 1. Chain-of-Thought in different spaces. (a) Language CoT paradigm predicts sub-tasks as intermediate reasoning. (b) Visual CoT paradigm synthesizes a goal image to provide guidance for action policy. (c) Our proposed Action CoT directly operates in the action space and provides homogeneous action guidance.

Despite the promising trajectory set by these paradigms, a critical challenge persists: existing generalist policies think predominantly in the vision-language (input) space, often failing to adequately address the inherent disparity between these rich, semantic representations and the requirements of precise, low-level action execution (output). Specifically, the knowledge encoded within the VLM backbone of VLA models is derived from pre-training on web-scale datasets focused on semantic alignment and question-

*Corresponding authors.

answering, yielding representations optimized for linguistic understanding rather than physical dynamics. Similarly, while world models forecast future visual states conditioned on inputs, their guidance remains tethered to naturally visual representations. Crucially, both semantic and visual forms of reasoning only offer suboptimal, indirect guidance for generating the necessary action sequence. Consequently, these prevailing approaches rely on an inherently constrained information conduit, struggling to convey the full, granular knowledge of the action space essential for truly grounded and accurate robotic policy learning.

The inherent semantic-kinematic gap in existing policies, *i.e*., a fundamental disconnect between high-level, abstract inputs and low-level, executable motor commands, necessitates a paradigm shift in how guidance is provided. We contend that to bridge this chasm, policies require guidance that is kinematically coherent, rather than purely semantic or visual. This core principle underpins our novel framework: **Action Chain-of-Thought (ACoT)** (Fig. 1(c)). We redefine the "thought" process not as a sequence of linguistic tokens, but as a structured chain of explicit, kinematically-grounded action intents. This approach furnishes the policy with direct motion cues, supplanting indirect representations. In a manner analogous to learning from physical demonstration, this direct conditioning on action-space information enables a substantially more efficient and veridically grounded policy learning process.

This foundational shift, however, introduces a critical and distinct research challenge: *How can we robustly and efficiently synthesize the complex, high-dimensional motion cues required for ACoT reasoning from the raw, heterogeneous multimodal inputs?*

Action-related information manifests in two complementary forms, *i.e*., explicit or implicit. The explicit form corresponds to observable motion trajectories, such as those in human demonstrations, which directly encode executable patterns of behavior. In contrast, the implicit form resides in latent cues, *e.g*., linguistic expressions like "reach out" or "grasp", as well as interaction intents embedded in visual contexts. Although these cues are not presented as explicit robotic trajectories, they implicitly define distributions over feasible actions within the action space. Building upon this insight, we introduce two synergistic mechanisms to generate both explicit and implicit guidance in the action space. We first propose the Explicit Action Reasoner (EAR), which is realized as a light-weight transformer. Particularly, EAR synthesizes coarse-grained motion trajectories conditioned on multimodal observations, offering direct and executable guidance within the action space. Secondly, we devise the Implicit Action Reasoner (IAR), which infers latent action priors through applying cross-attention modeling between downsampled multimodal representations and learnable queries, thereby providing implicit behavioral priors. Note that these two mechanisms are inherently complementary to each other. Subsequently, through jointly leveraging both EAR and IAR, we develop ACoT-VLA, an integrated Action Chain-of-Thought framework that enables grounded generalist robot policy learning. Extensive experiments across both real-world settings and three simulation benchmarks consistently demonstrate the effectiveness and versatility of our ACoT-VLA.

To summarize, our main contributions are as follows:

- Conceptually, we introduce Action Chain of Thought (ACoT), a new paradigm for generalist robot policies. To the best of our knowledge, this is the first work to formulate the deliberative process as a structured chain of explicit action-space intents, rather than abstract linguistic or visual sub-goals.
- We delve into essential action space guidance and propose the Explicit and Implicit Action Reasoners, which provide both explicit trajectory guidance and implicit behavioral inspiration for action prediction.
- Building upon these two modules, we further propose ACoT-VLA, a unified framework for grounded generalist robot policy learning.
- Empirically, we validate our approach through extensive simulation and real-world experiments, achieving state-of-the-art performance on multiple benchmarks.

2. Related Works

Vision-Language-Action Models. VLA models [14, 18, 19, 27] incorporate pre-trained VLM models to predict language-driven robotic action sequences. Early works [24, 68] formulate robot control as an autoregressive sequence generation problem, discretizing continuous actions into bins. Inspired by generative modeling [31, 38, 67], increasing works [5, 20, 33] adopt diffusion-based action policies to synthesize smooth and high-quality action trajectories. Given that robotic manipulation inherently occurs in three-dimensional space, a line of studies [30, 53, 60] have sought to enhance the spatial reasoning capability of VLA models by integrating 3D priors. For instance, SpatialVLA [41] integrates spatial embeddings to endow model with 3D awareness ,while 4D-VLA [56] incorporates both spatial and temporal information to enrich representations. Besides, due to the scarcity of large-scale real-world robot demonstrations, a series of efforts [6, 10, 12, 23, 37, 58] focus on data-centric solutions, constructing large-scale robotic datasets through simulation or real-world collection to scale up policy learning. Moreover, recent large-scale co-training approaches such as $\pi_{0.5}$ [22], GenieReasoner [35] and Gemini Robotics [47] demonstrate the potential of unifying web-scale language understanding with action learning, enhancing the policy's generalization ability while retaining the reasoning capability of pre-trained foundation models.

World-Model-based Policies. Advances in world mod-

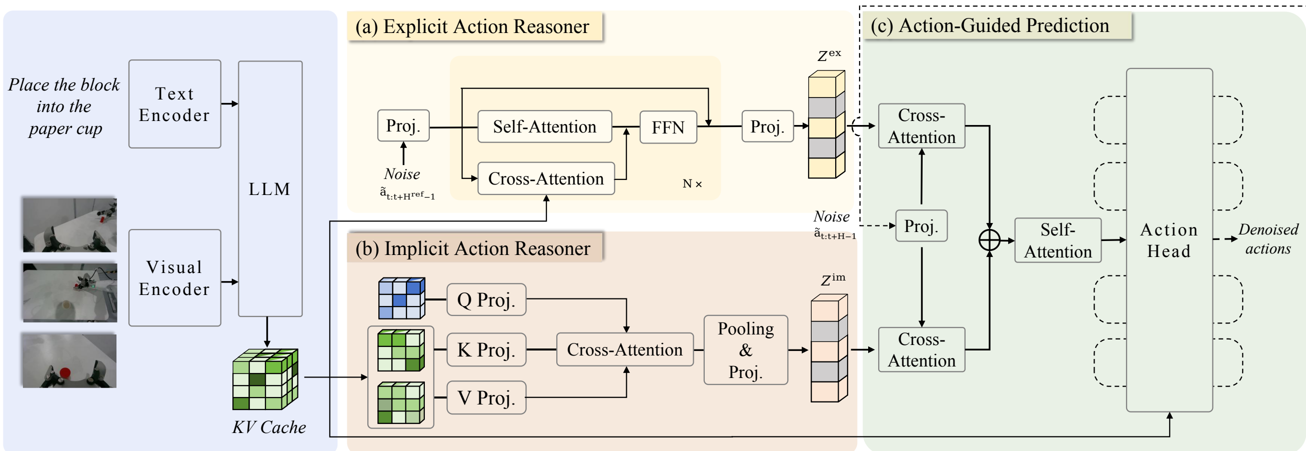

Figure 2. Architectural Overview of ACoT-VLA. The framework consists of three main components operating on features from a shared VLM backbone. (a) The Explicit Action Reasoner (EAR) is a Transformer-based module that synthesizes a coarse reference trajectory, providing explicit action-space guidance. (b) The Implicit Action Reasoner (IAR) employs a cross-attention mechanism with learnable queries to extract latent action priors from the VLM's internal representations. (c) The Action-Guided Prediction (AGP) head synergistically integrates both explicit and implicit guidances via cross-attention to condition the final denoising process, producing the executable action sequence.

els have demonstrated remarkable capability in synthesizing high-fidelity images and temporally coherent videos. Building upon such progress, emerging researches [26, 29, 36, 57] exploit their predictive dynamics to implicitly guide action generation. Specifically, CoT-VLA [61] introduces visual chain-of-thought reasoning by forecasting sub-goal images, explicitly integrating visual reasoning into action prediction. WorldVLA [8] employs an autoregressive architecture that unifies perception and action generation within a single framework. DreamVLA [59] extends beyond visual prediction and enriches world modeling with dynamic, depth, and semantic cues, improving the model's physical consistency. Collectively, existing world-model-based methods adopt knowledge-forecasting perspective, incorporating primarily visual guidance into trajectory generation.

In contrast to previous works focusing on visual or linguistic intermediaries for robotic policy learning, our key insight lies in investigating guidance directly within the action space, which intrinsically mitigates the heterogeneity between perception and action, enabling the model to effectively learn action-relevant priors.

3. Methodology

In this section, we present a detailed investigation into how to generate effective action space guidance and integrate it into robotic policy learning. We first define the robotic manipulation problem and formulate our proposed approach in Sec. 3.1. The core of our method lies in two distinct action reasoners introduced in Sec. 3.2 and Sec. 3.3, which provide explicit and implicit guidance within the action space. We conclude by illustrating the policy prediction strategy that effectively integrates this action guidance during policy learning (Sec. 3.4).

3.1. Problem Formulation

Given a natural language instruction l and current visual observation o_t, the generalist robot policy π_θ aims to predict action sequences $a_{t:t+H-1}$ that accomplishes the specified task. The process can be formally expressed as:

$$a_{t:t+H-1} = \pi_\theta(o_t, l), \tag{1}$$

where H represents the action horizon. Numerous works introduce additional guidance signals g, which encapsulates various forms of auxiliary information to enhance policy's prediction ability. Specifically, these guidance signals can be broadly categorized into two types: language-level guidance g_{lang} and vision-level guidance g_{vis}. The former is primarily adopted by VLM-based methods, *e.g.*, leveraging LLMs' reasoning capabilities to predict sub-task, while the latter is always employed by world-model-based approaches, such as simulating future observations. Such relationship can be formulated as:

$$\pi_\theta(a_{t:t+H-1}, g \mid o_t, l) = \pi_\theta(a_{t:t+H-1} \mid o_t, l, g)\pi_\theta(g \mid o_t, l), \tag{2}$$

where $g \in \{g_{\text{lang}}, g_{\text{vis}}\}$. Conversely, we shift the focus toward the action domain itself and investigate cues operating directly in the action space, symbolized as g_{action}. The above guidances are extended as $g \in \{g_{\text{lang}}, g_{\text{vis}}, g_{\text{action}}\}$.

Such action guidance can intuitively be disentangled into explicit and implicit forms. The explicit guidance $g^{\text{ex}}_{\text{action}}$ provides direct priors in the form of reference action sequences, whereas the implicit guidance $g^{\text{im}}_{\text{action}}$ arises from contextual signals, *e.g.*, action distribution inherently implied in linguistics.

3.2. Explicit Action Reasoner

To incorporate explicit action trajectories into the thinking process of π_θ to generate high-quality action predictions, we propose the Explicit Action Reasoner (EAR).

We design a mechanism that enables the model to autonomously synthesize reference action sequences as internal guidance for policy learning. Analogously, this formulation can be viewed as an action-space transfer of self-conditioning in generative models [9, 34], where incorporating prior estimates into the generation process has been shown to markedly improve sample quality. Building upon this principle, we instantiate EAR as a light-weight transformer, as shown in Fig. 2 (a), generating kinematically plausible action reference as explicit action-space guidance $g^{\text{ex}}_{\text{action}}$ for downstream action policy.

Formally, given visual observation o_t and language instruction l, a pre-trained VLM encodes them into a contextual key-value cache:

$$(K^{\text{VLM}}_{1:N}, V^{\text{VLM}}_{1:N}) = \text{VLM}(o_t, l), \tag{3}$$

where N represents the number of layers of VLM. Subsequently, the EAR, denoted as π^{ref}_θ, takes a noisy action sequence $\tilde{a}_{t:t+H^{ref}-1}$ as input, where H^{ref} indicates the horizon of reference actions. The sequence is first embedded into an initial hidden representation h^{ref}_0, which serves as the input to EAR's transformer layers. At each transformer layer i, we adopt self-attention, along with cross-attention with the contextual key-value cache from the corresponding VLM layer:

$$\tilde{h}^{\text{ref}}_i = \text{Self-Attn}(h^{\text{ref}}_{i-1}) + \text{CrossAttn}(h^{\text{ref}}_{i-1}, K^{\text{VLM}}_i, V^{\text{VLM}}_i), \tag{4}$$

where self-attention module captures temporal dependencies within the action sequence and cross-attention mechanism injects multimodal contextual priors from the VLM. Then, the intermediate representation $\tilde{h}^{\text{ref}}_i$ is processed by a feed-forward network (FFN) in a residual-parallel manner, updating the i-th EAR representation h^{ref}_i:

$$h^{\text{ref}}_i = h^{\text{ref}}_{i-1} + \text{FFN}(\tilde{h}^{\text{ref}}_i). \tag{5}$$

Through training via flow matching, π^{ref}_θ learns a distribution over action trajectories, producing a denoised action sequence:

$$a^{ref}_{t:t+H^{ref}-1} = \pi^{\text{ref}}_\theta(\tilde{a}_{t:t+H^{ref}-1}, K^{\text{VLM}}_{1:N}, V^{\text{VLM}}_{1:N}). \tag{6}$$

The generated sequence is then encoded via a MLP projector to obtain action embedding Z^{ex}, which serves as explicit action-space guidance $g^{\text{ex}}_{\text{action}}$ for action policy learning.

3.3. Implicit Action Reasoner

Beyond the explicit action trajectories, the multimodal latent space of VLM also encodes implicit motion cues [13, 40], *e.g.*, visual affordances and action-related semantics. Effectively extracting these action-relevant representations potentially offers complementary guidance. To this end, we introduce an Implicit Action Reasoner (IAR), which directly operates on the VLM's key–value cache.

Concretely, as presented in Fig. 2 (b), for each VLM layer $i \in [1, N]$, we initialize a learnable matrix $Q_i \in \mathbb{R}^{M \times d}$, where M is a hyperparameter and d represents VLM's hidden dimension. Considering the information redundancy within VLM's key–value cache and computational efficiency, we first downsample the corresponding key–value pairs into a lower-dimensional space, which is formulated as:

$$Q'_i = Q_i W^{(i)}_Q, \quad K'_i = K^{\text{VLM}}_i W^{(i)}_K, \quad V'_i = V^{\text{VLM}}_i W^{(i)}_V, \tag{7}$$

where $W^{(i)}_Q, W^{(i)}_K, W^{(i)}_V \in \mathbb{R}^{d \times d'}$ are learnable linear projectors and $d' \ll d$.

Later, cross-attention is applied to extract action-relevant information from each K'_i and V'_i. The resulting features are subsequently integrated via average pooling and transformed through a MLP projector, as visualized in Fig. 2 (b), producing compact representations that capture the implicit action semantics z^{im}_i embedded in VLM's i-th layer:

$$z^{\text{im}}_i = \text{MLP}(\text{Pool}(\text{CrossAttn}(Q'_i, K'_i, V'_i))). \tag{8}$$

Then, through aggregating these representations across layers, we obtain implicit action-related feature Z^{im}, which serves as implicit action-space guidance $g^{\text{im}}_{\text{action}}$, complementing the explicit motion priors.

3.4. Action-Guided Prediction

Building upon the explicit action embedding Z^{ex} produced by EAR and implicit action-related feature Z^{im} obtained in IAR, in this section, we introduce the Action-Guided Prediction (AGP) strategy to incorporate both action guidances into policy learning.

As illustrated in Fig. 2 (c), given a noisy action segment $\tilde{a}_{t:t+H-1}$, we first encode it into noisy action embedding via a MLP projector. Particularly, unlike previous approaches that directly feed this embedding into action head π^{head}_θ, we treat it as action query, denoted as Q_{action}, which interacts with both Z^{ex} and Z^{im} to retrieve complementary priors for conditional prediction.

Specifically, we perform dual cross-attention operations:

$$S^{\text{ex}} = \text{CrossAttn}(Q_{action}, Z^{\text{ex}}, Z^{\text{ex}}), \tag{9}$$

$$S^{\text{im}} = \text{CrossAttn}(Q_{action}, Z^{\text{im}}, Z^{\text{im}}), \tag{10}$$

where S^{ex} and S^{im} denote the attended representations guided by explicit and implicit priors, respectively. Note that although both encode action-relevant information, they may highlight different facets of the underlying motion. For

Methods	Guidance	Spatial		Object		Goal		Long		Avg.	
		SR ↑	Rank ↓	SR ↑	Rank ↓	SR ↑	Rank ↓	SR ↑	Rank ↓	SR ↑	Rank ↓
Diffusion Policy [11]	–	78.3	26	92.5	18	68.3	27	50.5	27	72.4	27
Octo [48]	–	78.9	25	85.7	26	84.6	20	51.1	26	75.1	25
CoT-VLA [61]	Visual	87.5	20	91.6	20	87.6	17	69.0	19	81.1	20
WorldVLA [8](256*256)	Visual	85.6	22	89.0	23	82.6	22	59.0	22	79.1	21
WorldVLA [8](512*512)	Visual	87.6	19	96.2	15	83.4	21	60.0	21	81.8	19
DreamVLA [59]	Visual	97.5	9	94.0	16	89.5	15	89.5	12	92.6	14
UniVLA [49]	Visual	95.4	14	98.8	4	93.6	12	94.0	6	95.5	10
F1 [36]	Visual	98.2	5	97.8	10	95.4	11	91.3	10	95.7	9
GE-Act [29]	Visual	98.2	5	97.6	12	95.8	9	94.4	5	96.5	7
TraceVLA [65]	Linguistics	84.6	24	85.2	27	75.1	26	54.1	24	74.8	26
OpenVLA [24]	Linguistics	84.7	23	88.4	24	79.2	23	53.7	25	76.5	24
UniAct [64]	Linguistics	77.0	27	87.0	25	77.0	25	70.0	18	77.8	23
SpatialVLA [41]	Linguistics	88.2	18	89.9	22	78.6	24	55.5	23	78.1	22
ThinkAct [17]	Linguistics	88.3	17	91.4	21	87.1	18	70.9	17	84.4	18
π_0-FAST [39]	Linguistics	96.4	12	96.8	14	88.6	16	60.2	20	85.5	17
FPC-VLA [52]	Linguistics	87.0	21	92.0	19	86.2	19	82.2	15	86.9	16
SmolVLA [43]	Linguistics	93.0	16	94.0	16	91.0	14	77.0	16	88.8	15
GR00T-N1 [4]	Linguistics	94.4	15	97.6	12	93.0	13	90.6	11	93.9	13
π_0 [5]	Linguistics	96.8	11	98.8	4	95.8	9	85.2	14	94.1	12
GO-1 [6]	Linguistics	96.2	13	97.8	10	96.0	8	89.2	13	94.8	11
DD-VLA [28]	Linguistics	97.2	10	98.6	6	97.4	5	92.0	9	96.3	8
MemoryVLA [42]	Linguistics	98.4	4	98.4	7	96.4	7	93.4	7	96.7	6
$\pi_{0.5}$ [22]	Linguistics	98.8	2	98.2	9	98.0	3	92.4	8	96.9	5
OpenVLA-OFT [25]	Linguistics	97.6	8	98.4	7	97.9	4	94.5	4	97.1	4
VLA-Adapter [50]	Linguistics	97.8	7	99.2	2	97.2	6	95.0	3	97.3	3
Ours†	**Action**	**99.4**	**1**	**99.6**	**1**	98.8	2	96.0	2	**98.5**	**1**
Ours	**Action**	98.6	3	99.0	3	**99.4**	**1**	**97.0**	**1**	**98.5**	**1**

Table 1. Comparison on the LIBERO benchmark. Our proposed approach is trained on the LIBERO dataset. † represents that the LLM backbone is frozen during training. All metrics are average success rates (%). The best results are highlighted in **bold**.

instance, explicit priors provide kinematic cues, whereas implicit priors capture latent action tendencies. Hence, to effectively combine these complementary guidance, we concatenate the two attended features and process them through self-attention fusion block, which integrates the priors into a unified representation:

$$\bar{h} = \text{Self-Attn}([S^{\text{ex}};\ S^{\text{im}}]). \tag{11}$$

Eventually, the aggregated representation $\bar{h}$ is fed into π_θ^{head}, which predicts the denoised action sequence $a_{t:t+H-1}$.

Training Objectives. The entire framework is optimized under a standard flow-matching mean-squared error (MSE) objective. The training losses consist of two parts, *i.e.*, flow-matching MSE for both Explicit Action Reasoner π_θ^{ref} and action head π_θ^{head}. Hence, the overall objective is:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\pi_\theta^{\text{ref}}} + \lambda_2 \mathcal{L}_{\pi_\theta^{\text{head}}}, \tag{12}$$

where λ_1 and λ_2 are balance factors.

Teacher Forcing Stabilization. During training, the outputs of π_θ^{ref} can be unstable. To stabilize optimization, we compute Z^{ex} directly from ground-truth reference trajectories instead of from π_θ^{ref} predictions, preventing optimization interference to π_θ^{head}. During inference, the model switches to a fully self-conditioned mode, where π_θ^{ref} autonomously generates the reference actions to guide π_θ^{head} in action prediction.

4. Experiments

In this section, we first outline the experimental setup in Sec. 4.1. In Sec. 4.2, we evaluate our approach on three simulation benchmarks, followed by comprehensive ablation studies in Sec. 4.3. Moreover, we present real-world deployment results in Sec. 4.4 to evaluate its applicability.

4.1. Experimental Setup

Data Sources. For simulation experiments, we strictly follow the official training splits provided by the corresponding benchmark (LIBERO [32], LIBERO-Plus [15], and VLABench [58]), and train our models exclusively on their standard demonstration datasets without introducing any additional data. For the real-world setting, all demonstrations used for model training are collected on our own robotic platform.

Implementation Details. We implement our approach upon $\pi_{0.5}$ [22]. Specifically, we adopt SigLIP [55] as the visual encoder, while the LLM backbone is instantiated as Gemma 2B architecture [3] with $N = 18$ layers and hidden size $d = 2048$. For frame processing, each input frame is resized to 224×224 prior to the visual encoder. Regarding the EAR, we employ a compact Transformer-based design composed of $N = 18$ layers. Concerning the IAR, each learnable matrix Q_i is configured with a row dimension of $M = 1$. The reduced dimension in the downsampling strategy is set to $d' = 128$.

Methods	Guidance	Camera	Robot	Language	Light	Background	Noise	Layout	Avg.
Zero-Shot Transfer									
WorldVLA [8]	Visual	0.1	27.9	41.6	43.7	17.1	10.9	38.0	25.0
OpenVLA [24]	Linguistics	0.8	3.5	23.0	8.1	34.8	15.2	28.5	15.6
NORA [21]	Linguistics	2.2	37.0	65.1	45.7	58.6	12.8	62.1	39.0
UniVLA [7]	Linguistics	1.8	46.2	69.6	69.0	81.0	21.2	31.9	42.9
π_0-Fast [39]	Linguistics	65.1	21.6	61.0	73.2	73.2	74.4	68.8	61.6
RIPT-VLA [44]	Linguistics	55.2	31.2	77.6	88.4	91.6	73.5	74.2	68.4
OpenVLA-OFT [25]	Linguistics	56.4	31.9	79.5	88.7	93.3	75.8	74.2	69.6
π_0^* [5]	Linguistics	61.0	40.8	63.5	89.3	84.1	80.1	76.4	69.4
$\pi_{0.5}^*$ [22]	Linguistics	**75.8**	79.4	83.3	95.5	95.0	**89.6**	87.0	85.7
Ours†	**Action**	68.9	80.3	84.1	95.6	93.1	81.5	**88.3**	83.6
Ours	**Action**	72.6	**82.6**	**87.5**	**97.7**	**96.5**	87.8	88.1	**86.6**
Supervised Fine-Tuning									
$\pi_0^\dagger$ [5]	Linguistics	79.6	21.1	72.5	84.7	86.2	68.3	69.4	67.4
$\pi_{0.5}^\dagger$ [22]	Linguistics	70.3	41.7	**81.1**	**97.3**	94.6	71.8	84.9	75.7
Ours†	**Action**	91.2	62.5	80.3	95.1	91.5	88.3	84.9	84.1
Ours	**Action**	**96.6**	**70.4**	79.7	95.1	**97.1**	**95.9**	**85.0**	**88.0**

Table 2. Comparison on the LIBERO-Plus benchmark. Methods under *Zero-Shot Transfer* are trained on LIBERO dataset and directly evaluated on LIBERO-Plus. *Supervised Fine-Tuning* denotes models trained on the LIBERO-Plus training set. An asterisk (*) denotes results reproduced by utilizing officially released checkpoints, while † represents that the LLM backbone is frozen during training. The best results are highlighted in **bold**.

In terms of model training, unless explicitly specified, the horizon of predicted reference actions H^{ref} and action policy output H are fixed to 15 and 10, with action shift set to 2 and 1, respectively. To clarify, the action shift specifies the temporal interval relative to the expert demonstration. For instance, a shift of 1 yields frame-aligned predictions, whereas a shift of 2 skips one intermediate frame. We set the balance factors in training losses as $\lambda_1 = \lambda_2 = 0.5$.

Training Configuration. We adopt a unified set of training hyperparameters across all experiments unless explicitly specified. Concretely, the learning rate follows a cosine-decay schedule with a warm-up phase of 10K steps, a peak learning rate of 5e$-$5, and a decay toward 5e$-$5 over 10K steps. Optimization is performed with AdamW with gradient-norm clipping set to 1.0. An exponential moving average (EMA) of model parameters is maintained with a decay rate of 0.999. Regarding hardware settings, model training is performed on a single node equipped with 8 NVIDIA H100 GPUs using bfloat16 precision. And the inference is conducted on a single NVIDIA RTX 4090.

4.2. Simulation Experiments

In this section, we conduct the simulation evaluations across three benchmarks, *i.e*., LIBERO [32], LIBERO-Plus [15], and VLABench [58], to comprehensively evaluate our approach's performance and generalization capabilities.

LIBERO Benchmark. We evaluate our approach on LIBERO benchmark, which targets four distinct robot capabilities: spatial awareness (Spatial), object manipulation (Object), goal completion (Goal), and long-horizon reasoning (Long). Each task suite consists of 10 tasks and provides 50 human-teleoperated demonstrations per task for policy training. The evaluation is conducted following the official evaluation protocol. For each task, the policy is evaluated over 50 trials, amounting to $2,000$ total rollouts.

As reported in Table 1, the quantitative evaluation results demonstrate that our proposed approach outperforms existing methods across all tracks. Compared to previous state-of-the-art method $\pi_{0.5}$, our approach achieves a 1.6% absolute improvement in average. Notably, we observe a pronounced improvement on LIBERO-Long suite, where tasks require long-horizon manipulation with strict error control. Particularly, unlike Language- or Visual-CoT, whose intermediate reasoning remains abstract or indirect with respect to action execution, our proposed ACoT naturally operates in precise representation. Through leveraging actions as intermediate reasoning, the model feeds the action head with structured action guidance, which significantly enhances the robustness in long-horizon manipulation tasks.

LIBERO-Plus Benchmark. Built upon LIBERO, LIBERO-Plus is designed to systematically evaluate robotic policies under controlled distribution shifts. Concretely, LIBERO-Plus introduces 7 perturbation dimensions, *i.e*., camera-viewpoints (Camera), robot-initial-states (Robot), language-variations (Language), lighting-conditions (Light), background-textures (Background), sensor-noise (Noise) and object-layout (Layout), exposing hidden failure modes under standard evaluations. Notably, LIBERO-Plus consists of $10,030$ evaluation episodes, providing statistically reliable evaluation.

We consider two evaluation protocols: (i) *Zero-Shot Transfer*, where policies trained on the LIBERO dataset

Methods	Guidance	In-dist.		Category		Commonsense		Instruction		Texture		Avg.	
		IS ↑	PS ↑	IS ↑	PS ↑	IS ↑	PS ↑	IS ↑	PS ↑	IS ↑	PS ↑	IS ↑	PS ↑
$\pi_0^\dagger$ [5]	Linguistics	67.8	62.7	44.0	33.6	54.9	**43.0**	**58.0**	38.7	50.6	42.5	55.0	44.1
$\pi_{0.5}^\dagger$ [22]	Linguistics	75.0	60.8	49.6	35.3	**57.5**	41.6	57.1	30.3	62.0	47.4	60.2	43.1
Ours†	**Action**	**79.8**	**66.1**	**54.1**	**38.9**	52.3	37.8	56.8	**39.6**	**74.6**	**54.6**	**63.5**	**47.4**

Table 3. Comparison on the VLABench benchmark. IS and PS represent Intention score and Progress score, respectively. All models are trained for 60K steps. † indicates that the LLM backbone is frozen during training. The best results are highlighted in **bold**.

Name	EAR	IAR	Spatial	Object	Goal	Long	Avg.
Baseline			98.8	98.2	98.0	92.4	96.9
#1	✓		99.0	99.4	98.0	**96.6**	98.3
#2		✓	99.2	99.2	98.2	95.6	98.1
#3	✓	✓	**99.4**	**99.6**	**98.8**	96.0	**98.5**

Table 4. Module ablations. The performance is gradually improved with the continuous addition of proposed methods.

are directly evaluated on LIBERO-Plus to assess generalization. (ii) *Supervised Fine-Tuning*, where models are directly optimized on LIBERO-Plus dataset. Technically, we follow evaluation configuration in LIBERO-Plus [15], *i.e.*, each episode is executed once without repeated rollouts.

As shown in Table 2, our method significantly boosts the policy's performance, surpassing all previous methods in both settings. Specifically, under the *Zero-Shot* regime, our approach demonstrates pronounced robustness against distribution shifts such as robot initial-state perturbations (+3.2%) and language variations (+4.2%), where existing language-guided or vision-guided policies exhibit significant degradation. Furthermore, our method maintains exceptional performance under the *Supervised Fine-Tuning* setting, reaching an 88.0% average success rate. These results highlight the effectiveness of our action-space reasoning in improving generalization and robust policy learning.

VLABench Benchmark. Built on ManiSkill3 [45], VLABench is designed to benchmark both VLAs and VLMs on diverse robotic tasks. The standard evaluation is organized into 5 public tracks, *i.e.*, in-distribution, cross-category (category-level generalization), commonsense reasoning, semantic-instruction (language understanding), and unseen-texture (appearance robustness). Particularly, VLABench proposes Intention Score (IS) and Progress Score (PS) to evaluate robot policies.

In our context, we train π_0, $\pi_{0.5}$, along with our method in a unified training setup on VLABench's official training data. We present quantitative results in Table 3. Overall, our method achieves the best performance across both IS (63.5%) and PS (47.4%). Notably, under the challenging unseen-texture track, it delivers substantial gains, *i.e.*, +12.6% in IS and +7.2% in PS, indicating strong robustness to distribution shifts. Together, these results further confirm the effectiveness of our proposed approach.

4.3. Ablation Study

We examine each component's contribution via systematic ablation experiments on the LIBERO benchmark, which are shown in Table 4, Table 5, and Table 6. Note that we adopt $\pi_{0.5}$ as the "Baseline" method. More ablations in different benchmarks are in the supplementary material.

Name	Action shift	Action horizon	Equi. horizon	Spatial	Object	Goal	Long	Avg.
Baseline	1	10	10	98.6	99.0	96.4	92.2	96.6
+EAR	1	10	10	99.4	99.4	**98.8**	95.0	98.2
	2	5	10	**99.6**	**99.6**	98.4	94.4	98.0
	1	30	30	99.2	99.2	97.6	95.6	97.9
	2	15	30	99.0	99.4	98.0	**96.6**	**98.3**
	2	30	60	99.4	99.0	98.2	95.0	97.9
	3	30	90	98.8	99.4	97.4	96.2	98.0

Table 5. Reference action parameter ablation. We observe that different reference-action configurations within EAR generally lead to performance improvements.

Methods	Spatial	Object	Goal	Long	Avg.
Baseline	98.8	98.2	98.0	92.4	96.9
Query	98.8	99.0	97.2	92.8	97.0
Attention Pooling	**99.4**	98.6	**98.2**	92.8	97.3
Downsample	99.2	**99.2**	**98.2**	**95.6**	**98.1**

Table 6. Comparison of KV-cache interaction strategies in IAR.

EAR. As shown in Table 4, compared with the baseline, the experiment "#1" introduces the Explicit Action Reasoner (EAR) module into policy learning, which lifts the average success rate from 96.9% to 98.3%, demonstrating that the explicit action-space guidance benefits the robotic action sequence prediction. A plausible explanation is that EAR introduces an intermediate reference action sequence, which injects strong inductive bias on the behavior and thereby reduces ambiguity in mapping from observations to actions.

IAR. Analogously, with the Implicit Action Reasoner (IAR) module added in "#2", the average success rate increases from 96.9% to 98.1%. This gain suggests that exploiting the implicit action distribution encoded in vision–language representations can also provide effective guidance for policy learning. This performance gain can be partly attributed to the fact that IAR distills action-related clues implicitly encoded within the vision–language backbone, which potentially reflects the distribution of feasible actions. Such priors encourage the policy to remain closer to coherent, task-consistent behavioral patterns.

EAR + IAR. In Table 4, experiment "#3" incorporates both EAR and IAR, achieving the highest average success rate of 98.5%. The consistent improvements demonstrate that explicit action guidance and implicit action cues extracted from VLM's key-value cache are complementary, jointly providing stronger guidance for accurate action prediction.

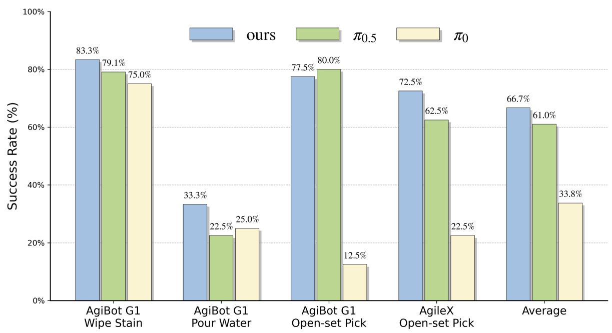

Figure 3. Evaluation results of real-world experiments.

Reference Action Parameters. To further examine the effect of explicit action references in EAR, we investigate different settings of action shift and action horizon, as summarized in Table 5. We observe that various parameter combinations consistently bring improvements over the baseline, indicating that providing action cues is broadly beneficial for policy learning. Besides, we find that shorter horizons combined with moderate shifts tend to produce relatively stronger gains. These observations offer further insight into how explicit action guidance influence policy learning.

KV-cache Interaction Strategies. We compare three strategies for extracting action-relevant cues from VLM's key-value cache within IAR module, as presented in Table 6. Concretely, "Query" method uses learnable queries to attend to VLM's original key-value cache. "Attention Pooling" method forms a pooled query by averaging key-value cache and then applies cross-attention. "Downsample" method first downsample VLM's key-value cache and then aggregates them using learnable matrix.

As shown in Table 6, all three variants outperform the baseline, indicating that extracting implicit action cues from VLM benefits policy learning. Notably, the "Downsample" strategy achieves the best performance, suggesting that VLM's features may contain redundancy for action prediction. This also highlights the importance of designing appropriate interaction mechanisms.

4.4. Real-World Deployment

To further validate the effectiveness of our framework, we conduct extensive real-world experiments on the AgiBot G1 robot. We consider three manipulation tasks, *i.e.*, "Wipe Stain", "Pour Water", and "Open-set Pick", which respectively assess contact-rich manipulation, fine-grained object handling, and instruction-following abilities.

Specifically, (i) the "Wipe Stain" task requires the robot to pick up a sponge from the table and wipe away the stain until the surface is clean. (ii) The "Pour Water" task requires the robot to grasp the kettle by its handle, locate the target cup, pour water into it without causing overflow, and finally return the kettle to the table in a stable manner. (iii) The "Open-set Pick" task instructs the robot to pick up the correct tabletop object according to given natural-language command. Additionally, to examine the cross-embodiment adaptability, we also perform the "Open-set Pick" task on the AgileX robotic platform. More details about training and evaluation are provided in the supplementary material.

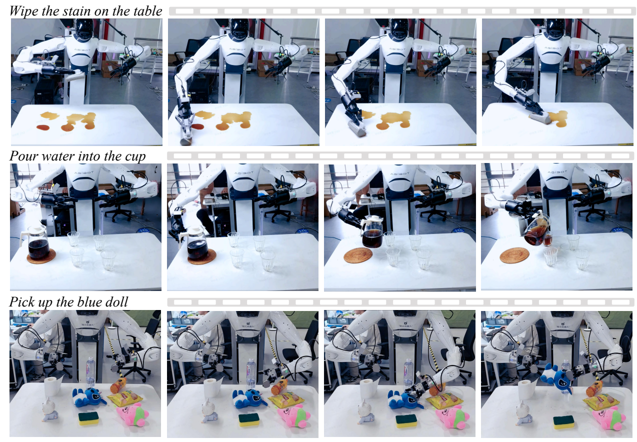

Figure 4. Visualization of three manipulation tasks in real world.

As shown in Fig. 3, our approach achieves consistently higher average success rates than both $\pi_{0.5}$ and π_0, *i.e.*, 66.7% against 61.0% and 33.8%. These results demonstrate the proposed framework maintains effectiveness under real-world sensing conditions. Moreover, the aligned improvements observed on both AgiBot G1 and AgileX also indicate that our method exhibits adaptability across different robotic embodiments. Visualization is provided in Fig. 4.

5. Conclusion

In this work, we address the fundamental semantic-kinematic gap in modern robotic policies by proposing a new paradigm: Action Chain-of-Thought (ACoT). We argue that for physically grounded intelligence, deliberation should occur not in the abstract space of language or vision, but directly in the kinematically grounded space of actions. We materialize this concept in our ACoT-VLA framework, which leverages two synergistic modules—an Explicit Action Reasoner (EAR) and an Implicit Action Reasoner (IAR)—to generate and fuse both explicit trajectory plans and implicit behavioral priors. This action-centric guidance mechanism creates a direct, information-rich conduit between high-level intent and low-level motor control. Extensive experiments across multiple simulation and real-world benchmarks demonstrate that our proposed approach yields state-of-the-art performance. Through shifting the locus of reasoning from perception to action, our work not only provides a more effective and grounded method for robotic policy learning, but also opens a new avenue for research into more structured, interpretable, and capable embodied agents. We believe that learning to "think" in the language of actions is a critical step towards developing the next generation of generalist robots.

Acknowledgements

This research is supported in part by National Key R&D Program of China (2022ZD0115502), National Natural Science Foundation of China (No. 62461160308, No. 62576024, U23B2010), "the Fundamental Research Funds for the Central Universities" (No. 501RCQD2025), "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2024C01161), Beijing Natural Science Foundation (QY25227), Ningbo Science and Technology Innovation 2025 Major Project (2025Z034), NSFCRGC Project (N CUHK498/24).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1

[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 arXiv:2407.07726*, 2024. 5

[4] 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 arXiv:2503.14734*, 2025. 5

[5] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. corr, abs/2410.24164, 2024. doi: 10.48550. *arXiv preprint ARXIV.2410.24164*. 1, 2, 5, 6, 7

[6] Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025. 2, 5

[7] 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, 2025. 6

[8] Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, et al. Worldvla: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025. 3, 5, 6

[9] Ting Chen, Ruixiang Zhang, and Geoffrey Hinton. Analog bits: Generating discrete data using diffusion models with self-conditioning. *arXiv preprint arXiv:2208.04202*, 2022. 4

[10] Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025. 2

[11] 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*, 44 (10-11):1684–1704, 2025. 1, 5

[12] Shengliang Deng, Mi Yan, Songlin Wei, Haixin Ma, Yuxin Yang, Jiayi Chen, Zhiqi Zhang, Taoyu Yang, Xuheng Zhang, Wenhao Zhang, et al. Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data. *arXiv preprint arXiv:2505.03233*, 2025. 2

[13] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. Palm-e: An embodied multimodal language model. 2023. 4

[14] Jiafei Duan, Wentao Yuan, Wilbert Pumacay, Yi Ru Wang, Kiana Ehsani, Dieter Fox, and Ranjay Krishna. Manipulate-anything: Automating real-world robots using vision-language models. *arXiv preprint arXiv:2406.18915*, 2024. 2

[15] Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, et al. Libero-plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025. 5, 6, 7

[16] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018. 1

[17] Chi-Pin Huang, Yueh-Hua Wu, Min-Hung Chen, Yu-Chiang Frank Wang, and Fu-En Yang. Thinkact: Vision-language-action reasoning via reinforced visual latent planning. *arXiv preprint arXiv:2507.16815*, 2025. 5

[18] Siyuan Huang, Haonan Chang, Yuhan Liu, Yimeng Zhu, Hao Dong, Peng Gao, Abdeslam Boularias, and Hongsheng Li. A3vlm: Actionable articulation-aware vision language model. *arXiv preprint arXiv:2406.07549*, 2024. 2

[19] 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. 2

[20] Wenhui Huang, Changhe Chen, Han Qi, Chen Lv, Yilun Du, and Heng Yang. Motvla: A vision-language-action model with unified fast-slow reasoning. *arXiv preprint arXiv:2510.18337*, 2025. 2

[21] Chia-Yu Hung, Qi Sun, Pengfei Hong, Amir Zadeh, Chuan Li, U Tan, Navonil Majumder, Soujanya Poria, et al. Nora: A small open-sourced generalist vision language action model for embodied tasks. *arXiv preprint arXiv:2504.19854*, 2025. 6

[22] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025. 1, 2, 5, 6, 7

[23] Tao Jiang, Tianyuan Yuan, Yicheng Liu, Chenhao Lu, Jianning Cui, Xiao Liu, Shuiqi Cheng, Jiyang Gao, Huazhe Xu, and Hang Zhao. Galaxea open-world dataset and g0 dual-system vla model. *arXiv preprint arXiv:2509.00576*, 2025. 2

[24] 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, 2, 5, 6

[25] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025. 5, 6

[26] Hengtao Li, Pengxiang Ding, Runze Suo, Yihao Wang, Zirui Ge, Dongyuan Zang, Kexian Yu, Mingyang Sun, Hongyin Zhang, Donglin Wang, et al. Vla-rft: Vision-language-action reinforcement fine-tuning with verified rewards in world simulators. *arXiv preprint arXiv:2510.00406*, 2025. 3

[27] Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18061–18070, 2024. 2

[28] 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 arXiv:2508.20072*, 2025. 5

[29] Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Jingbin Cai, Si Liu, Jianlan Luo, et al. Genie envisioner: A unified world foundation platform for robotic manipulation. *arXiv preprint arXiv:2508.05635*, 2025. 3, 5

[30] Tao Lin, Gen Li, Yilei Zhong, Yanwen Zou, Yuxin Du, Jiting Liu, Encheng Gu, and Bo Zhao. Evo-0: Vision-language-action model with implicit spatial understanding. *arXiv preprint arXiv:2507.00416*, 2025. 2

[31] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 2

[32] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023. 5, 6

[33] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024. 1, 2

[34] Yunzhe Liu, Rinon Gal, Amit H Bermano, Baoquan Chen, and Daniel Cohen-Or. Self-conditioned generative adversarial networks for image editing. *arXiv preprint arXiv:2202.04040*, 2022. 4

[35] Yi Liu, Sukai Wang, Dafeng Wei, Xiaowei Cai, Linqing Zhong, Jiange Yang, Guanghui Ren, Jinyu Zhang, Maoqing Yao, Chuankang Li, et al. Unified embodied vlm reasoning with robotic action via autoregressive discretized pre-training. *arXiv preprint arXiv:2512.24125*, 2025. 2

[36] Qi Lv, Weijie Kong, Hao Li, Jia Zeng, Zherui Qiu, Delin Qu, Haoming Song, Qizhi Chen, Xiang Deng, and Jiangmiao Pang. F1: A vision-language-action model bridging understanding and generation to actions. *arXiv preprint arXiv:2509.06951*, 2025. 3, 5

[37] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024. 2

[38] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 2

[39] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025. 5, 6

[40] Shengyi Qian, Weifeng Chen, Min Bai, Xiong Zhou, Zhuowen Tu, and Li Erran Li. Affordancellm: Grounding affordance from vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7587–7597, 2024. 4

[41] Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025. 1, 2, 5

[42] Hao Shi, Bin Xie, Yingfei Liu, Lin Sun, Fengrong Liu, Tiancai Wang, Erjin Zhou, Haoqiang Fan, Xiangyu Zhang, and Gao Huang. Memoryvla: Perceptual-cognitive memory in vision-language-action models for robotic manipulation. *arXiv preprint arXiv:2508.19236*, 2025. 5

[43] Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, et al. Smolvla: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844*, 2025. 5

[44] Shuhan Tan, Kairan Dou, Yue Zhao, and Philipp Krähenbühl. Interactive post-training for vision-language-action models. *arXiv preprint arXiv:2505.17016*, 2025. 6

[45] Stone Tao, Fanbo Xiang, Arth Shukla, Yuzhe Qin, Xander Hinrichsen, Xiaodi Yuan, Chen Bao, Xinsong Lin, Yulin Liu, Tse-kai Chan, et al. Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai. *arXiv preprint arXiv:2410.00425*, 2024. 7

[46] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1

[47] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025. 2

[48] 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, 5

[49] Yuqi Wang, Xinghang Li, Wenxuan Wang, Junbo Zhang, Yingyan Li, Yuntao Chen, Xinlong Wang, and Zhaoxiang Zhang. Unified vision-language-action model. *arXiv preprint arXiv:2506.19850*, 2025. 5

[50] Yihao Wang, Pengxiang Ding, Lingxiao Li, Can Cui, Zirui Ge, Xinyang Tong, Wenxuan Song, Han Zhao, Wei Zhao, Pengxu Hou, et al. Vla-adapter: An effective paradigm for tiny-scale vision-language-action model. In *Proceedings of the AAAI conference on artificial intelligence*, pages 18638–18646, 2026. 5

[51] Ziyang Yan, Wenzhen Dong, Yihua Shao, Yuhang Lu, Haiyang Liu, Jingwen Liu, Haozhe Wang, Zhe Wang, Yan Wang, Fabio Remondino, et al. Renderworld: World model with self-supervised 3d label. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6063–6070. IEEE, 2025. 1

[52] Yifan Yang, Zhixiang Duan, Tianshi Xie, Fuyu Cao, Pinxi Shen, Peili Song, Piaopiao Jin, Guokang Sun, Shaoqing Xu, Yangwei You, et al. Fpc-vla: A vision-language-action framework with a supervisor for failure prediction and correction. *arXiv preprint arXiv:2509.04018*, 2025. 5

[53] Tianyuan Yuan, Yicheng Liu, Chenhao Lu, Zhuoguang Chen, Tao Jiang, and Hang Zhao. Depthvla: Enhancing vision-language-action models with depth-aware spatial reasoning. *arXiv preprint arXiv:2510.13375*, 2025. 2

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

[55] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 5

[56] Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Yanpeng Zhou, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, et al. 4d-vla: Spatiotemporal vision-language-action pretraining with cross-scene calibration. *arXiv preprint arXiv:2506.22242*, 2025. 2

[57] 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 arXiv:2501.18867*, 2025. 3

[58] Shiduo Zhang, Zhe Xu, Peiju Liu, Xiaopeng Yu, Yuan Li, Qinghui Gao, Zhaoye Fei, Zhangyue Yin, Zuxuan Wu, Yu-Gang Jiang, et al. Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11142–11152, 2025. 2, 5, 6

[59] Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, Fan Lu, He Wang, et al. Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge. *arXiv preprint arXiv:2507.04447*, 2025. 1, 3, 5

[60] Zhengshen Zhang, Hao Li, Yalun Dai, Zhengbang Zhu, Lei Zhou, Chenchen Liu, Dong Wang, Francis EH Tay, Sijin Chen, Ziwei Liu, et al. From spatial to actions: Grounding vision-language-action model in spatial foundation priors. *arXiv preprint arXiv:2510.17439*, 2025. 2

[61] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1702–1713, 2025. 3, 5

[62] 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. 1

[63] 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

[64] Jinliang Zheng, Jianxiong Li, Dongxiu Liu, Yinan Zheng, Zhihao Wang, Zhonghong Ou, Yu Liu, Jingjing Liu, Ya-Qin Zhang, and Xianyuan Zhan. Universal actions for enhanced embodied foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22508–22519, 2025. 5

[65] Ruijie Zheng, Yongyuan Liang, Shuaiyi Huang, Jianfeng Gao, Hal Daumé III, Andrey Kolobov, Furong Huang, and Jianwei Yang. Tracevla: Visual trace prompting enhances spatial-temporal awareness for generalist robotic policies. *arXiv preprint arXiv:2412.10345*, 2024. 5

[66] Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. Occworld: Learning a 3d occupancy world model for autonomous driving. In *European conference on computer vision*, pages 55–72. Springer, 2024. 1

[67] 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. *arXiv preprint arXiv:2408.11039*, 2024. 2

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