2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

CollabVLA: Self-Reflective Vision–Language–Action Model Dreaming Together with Human

Nan Sun[1,*], Yongchang Li[2,*], Chenxu Wang[1], Bo Mao[3], Huiying Li[1], Jiahe Yao[1], Kanghao Li[1], Jian Liu[4], Yifan Zhang[5], Guoying Zhang[2], Di Guo[3] and Huaping Liu[1,†]

***Abstract*— In this work, we present CollabVLA, a self-reflective vision–language–action framework that transforms a standard visuomotor policy into a collaborative assistant. CollabVLA tackles key limitations of prior VLAs, including domain overfitting, non-interpretable reasoning, and the high latency of auxiliary generative models, by integrating VLM-based reflective reasoning with diffusion-based action generation under a mixture-of-experts design. Through a two-stage training recipe of action grounding and reflection tuning, it supports explicit self-reflection and proactively solicits human guidance when confronted with uncertainty or repeated failure. It cuts normalized *Time* by $\sim 2\times$ and *Dream* counts by $\sim 4\times$ vs. generative agents, achieving higher success rates, improved interpretability, and balanced low latency compared with existing methods. This work takes a pioneering step toward shifting VLAs from opaque controllers to genuinely assistive agents capable of reasoning, acting, and collaborating with humans.**

I. Introduction

Large-scale vision–language models (VLMs) excel at open-world perception and instruction-following [1], [2], motivating vision-language-action (VLA) policies that fine-tune on robot data via autoregressive next-token prediction [3]–[5]. Yet this often degrades multimodal grounding due to domain overfitting, and scarce robot data prevents scaling generalization as in LLMs. To address this, some methods co-train VLAs on paired image–text corpora and learn latent action from internet videos [6]–[8]. However, co-training risks task interference [9], while latent actions, though compact, remain non-interpretable and add training complexity [7].

Inspired by chain-of-thought reasoning in LLMs [10], we naturally ask whether VLAs can also "think step-by-step" by making intermediate reasoning explicit. Indeed, recent work has explored this direction through textual planning, visual subgoal, and auxiliary prediction [11]–[14], showing improved transparency and generalization by aligning high-level reasoning with low-level policy.

This work was jointly supported by the National Natural Science Fund under grant nos 62025304 and 62120106005, Beijing Natural Science Foundation under grant no. L253006 and Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China under grant no. JYB2025XDXM109. * denotes the equal contribution.

[1]The author is with the Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China.
[2]The author is with the School of Artificial Intelligence, China University of Mining and Technology, Beijing, 100083, China.
[3]The author is with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, 100876, China.
[4]The author is with the Digital Navigation Center, Beihang University, Beijing, 100191, China.
[5]The author is with the Goertek Inc. AI Lab.
[†]Corresponding Author. hpliu@tsinghua.edu.cn

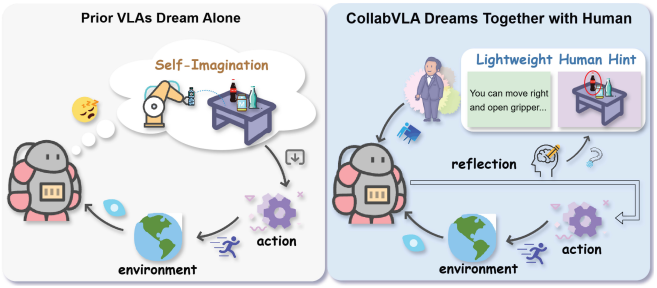

Fig. 1: CollabVLA extends beyond self-imagination by integrating human guidance with action generation, transforming a closed-loop visuomotor policy into a collaborative agent.

Yet these methods often fall short. For compact open-source VLMs or world models, explicit subgoal generation, particularly photorealistic egocentric images, usually generalizes only to seen layouts, while adding latency. Besides, their explicit reasoning remains largely a surface-level narration of the current situation, functioning as an internal monologue with limited guidance value. It lacks the deeper reflective understanding necessary for real-time failure recognition and effective interaction, as illustrated by the evolution from ReAct [15] to Reflexion [16] in LLM agents.

In this work, we present **CollabVLA**, a collaborative VLA capable of self-reflection and can proactively seek human guidance, rather than relying solely on inefficient and imperfect self-imagination (see Fig. 1). It integrates autoregressive VLM-based language generation with diffusion-based action generation under a mixture-of-experts (MoE) adaptation. CollabVLA follows a two-stage training recipe: (1) *Action Grounding*, where a VLM-driven action policy is trained on latent action representations conditioned on multimodal goals to master acting. To facilitate this stage, we combine the data pipelines of Interleave-VLA [17] and MDT [18] to construct a diverse and hybrid dataset of multimodal goals; (2) *Reflection Tuning*, which unifies scene understanding and action generation, thereby maintaining policy performance while strengthening robust internal reflective reasoning. In the second stage, the model is jointly trained on multimodal data, manipulation tasks, and a corpus constructed following InstructVLA [19], with additional embodied scenes designed for uncertainty and failure reflection.

Through this design, CollabVLA functions not only as a visuomotor policy in a closed-loop manner, but also as a reflective collaborator that reports reasoning outcomes and proactively incorporates concise textual hints or lightweight visual cues from human input to condition the next action

chunk. This avoids reliance on auxiliary world models by *co-dreaming* with humans when facing uncertainty or failures. CollabVLA takes a pinooring step by extending multimodal reasoning in modern VLMs to deeper reflective understanding, while attaining robustness through just-in-time human guidance that steers the success in the long tail.

In summary, our contributions are as follows:

- We identify and systematize the trade-offs among (a) direct autoregressive VLAs, (b) latent-action formulations, and (c) explicit world-model approaches, highlighting the missed opportunity for *lightweight human-in-the-loop* guidance at execution time.
- We introduce CollabVLA, a collaborative VLA framework with MoE adaptation that transforms a standard visuomotor policy into a proactive assistant capable of reasoning, acting, and interacting with humans.
- We show that CollabVLA improves success rates and maintains low latency, and effectively extends its self-reflection to solicit just-in-time human guidance.

II. Related Work

A. Vision–Language–Action Models

The success of VLMs [1], [2] has motivated extensions to VLA policies by fine-tuning on robot datasets and casting control as multimodal sequence modeling. Early systems such as RT-1 [3] and OpenVLA [4] adopt autoregressive next-token prediction on robot data, but this erodes multimodal grounding and limits generalization to novel settings or long-horizon tasks. To mitigate this, some work reintroduces large-scale image-text corpora and co-trains them with robot experience to preserve grounding while aligning it with control [6]. These approaches further explore latent-action formulations that distill intermediate representations or temporally chunked primitives from unlabeled internet data [7], [8]. Yet most methods remain *black boxes*, which hinders failure analysis and interactive correction.

A subsequent wave of research makes intermediate reasoning explicit. Some methods adopt explicit intermediate or conversational planning, such as ECoT [11] and ChatVLA [9]. Others rely on visual subgoals, including CoT-VLA [12]. Alternatively, diffusion-based or world-model–driven designs, such as GR-MG [14] and RoboDreamer [13], generate imagined future states to guide control. By exposing intermediate plans or auxiliary predictions, they improve transparency and robustness. However, explicit subgoal generation often struggles to generalize to unseen layouts and introduces latency. Most reasoning also remains limited to superficial rationalization, lacking the depth required for insightful guidance. These limitations motivate lightweight human-in-the-loop interaction, which CollabVLA enables by integrating self-reflection with action generation, transforming a visuomotor policy into a collaborative assistant capable of both acting and asking (see Fig. 2).

B. Human-in-the-Loop Collaboration

Human-in-the-loop (HITL) strategies have been extensively explored for determining when and what to ask humans.

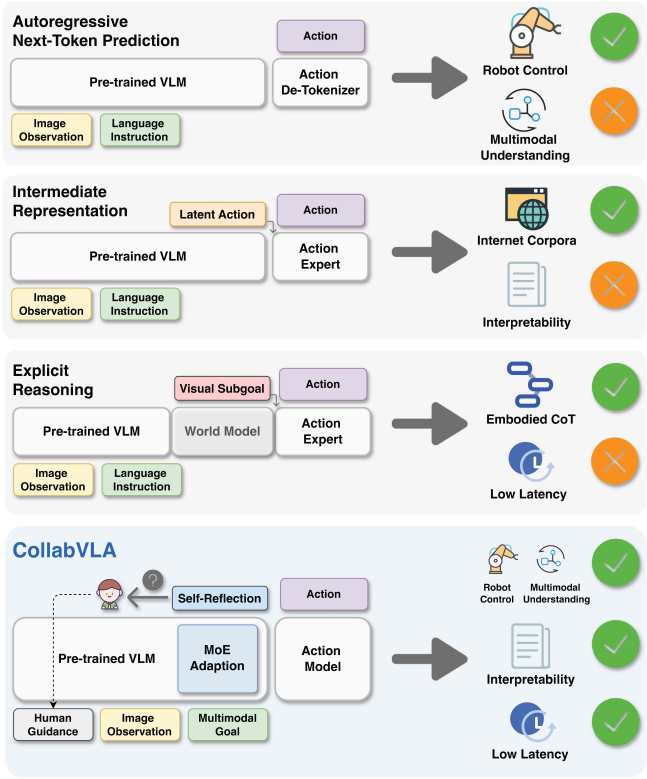

Fig. 2: **Comparison between prior methods and CollabVLA.** Prior methods often lose multimodal grounding, lack interpretability, or fail on unseen predictions and come up with high latency. In contrast, CollabVLA integrates self-reflection with lightweight human guidance to achieve robustness, transparency, and efficiency.

Notable approaches addressing uncertainty include KnowNo with conformal prediction [20] and Introspective Planning for safety [21]. In parallel, LLM-based critique methods, such as Reflexion [16], LLM-as-a-Judge [22], and critic-embedded AssistantX framework [23], show that self-reflection can be extended to trigger targeted human queries. However, most existing approaches rely on modular pipelines that combine high-level planners with low-level skill libraries, rather than unified models capable of integrated reasoning and control.

Nonetheless, they confirm that pretrained LLMs can act as reflective validators, reasoning over current observations and past progress. Prior embodied studies have partially recognized this by introducing explicit reasoning for self-guidance, but, to our knowledge, CollabVLA is the first VLA to retain a single-backbone visuomotor policy while enabling native reflective reasoning to guide action generation and incorporating real-time human guidance during execution.

III. Methodology

In this section, we first formalize the problem (Sec. III-A), then detail the construction of CollabVLA from three components: data (Sec. III-B), model architecture (Sec. III-C), and training pipeline (Sec. III-D).

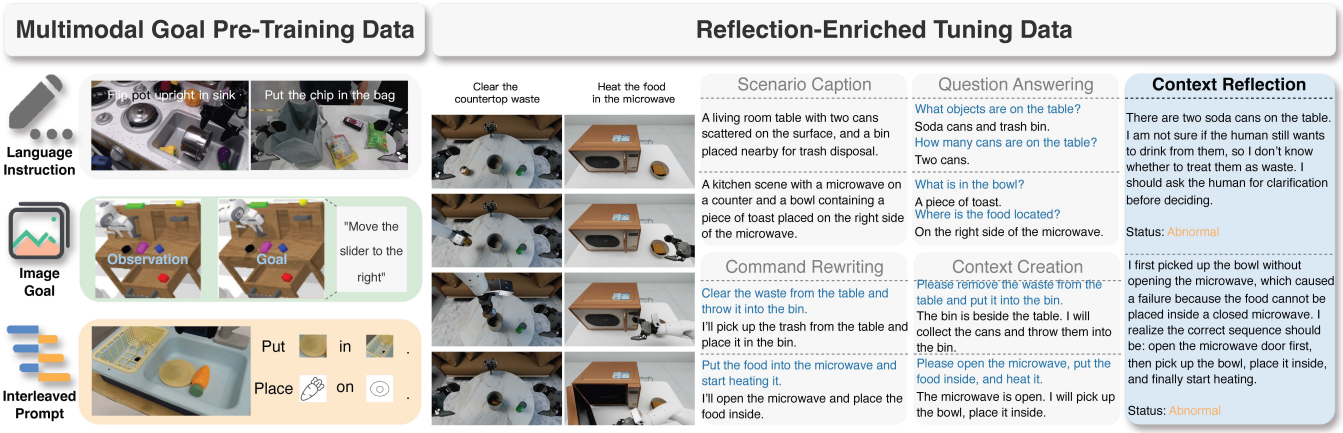

Fig. 3: **Overview of curated datasets for CollabVLA.** The left panel illustrates multimodal goal pre-training data with interleaved prompts and goal image augmentation, whereas the right panel presents reflection-enriched tuning data reformulated as *Context Reflection*, including failures due to inconsistencies between actions and states (e.g., grasping the food before opening the microwave) and ambiguities arising from multiple plausible targets (e.g., two similar cans with unclear intent).

A. Problem Formulation

We formalize CollabVLA as a goal-conditioned vision–language–action policy that outputs not only an action sequence but also explicit reflective reasoning and optional human queries. Given a current state o^t (image observations), past states o^{t-1}, proprioception p^t, and a multimodal goal g, the policy π_θ predicts:

$$\pi_\theta(o^t, o^{t-1}, p^t, g) \rightarrow \{\bar{a}^t, r^t, q^t\} \tag{1}$$

where $\bar{a}^t = (a_i^t, \ldots, a_{i+k-1}^t)$ denotes a short action chunk, r^t is a reflective reasoning trace, and q^t is a binary query indicator. The query token $q^t \in \{0, 1\}$ specifies whether to solicit human input: $q^t = 0$ allows the agent to proceed autonomously, while $q^t = 1$ triggers a follow-up question.

B. Data

VLA training data are largely language-only or only weakly aligned across vision, language, and action, with sparse, non-instructional labels; these limitations impede free-form instruction following and multimodal goal grounding. Prior work also lacks explicit supervision for *when* to self-diagnose, *when* to ask, and *how* to revise. We therefore curate two complementary corpora as illustrated in Fig. 3.

Multimodal Goal Pre-training. Built on simulation and real-world manipulation datasets [24]–[29], we add two augmentations: (i) *interleaved multimodal prompts* (as in Interleave-VLA [17]) that reformulate demonstrations into mixed text–image instructions; and (ii) *goal-image augmentation* (inspired by MDT [18] and GR-MG [14]) that samples future frames as explicit visual goals. We further adopt Diffusion-VLA-style reasoning augmentation [30] to attach concise language rationales to trajectories, injecting planning-oriented signals that act as lightweight self-reflection for successful cases.

Reflection-Enriched Tuning. To couple action with reflective behavior, we extend the InstructVLA pipeline [19] with a *Context Reflection* task: the agent explains past/current observations and diagnoses uncertainty or failure. We synthesize hard cases by (i) inserting irrelevant or shuffled frames to break temporal consistency, (ii) adding key objects to induce perceptual ambiguity and multiple choices, and (iii) perturbing action labels or goal descriptions to simulate failure trajectories. Each synthetic sample is formatted as a `{observation sequence, instruction, action trace, reflection}` tuple, where the reflection field captures how the agent should articulate uncertainty. We follow a reflection-oriented data generation pipeline: environment rollouts are paired with large language models to generate natural reflection answers, while visual states are synthesized with generative models and subsequently verified by human annotators for correctness. In addition, we interleave synthetic and real failure cases to balance distribution, and vary reflection styles to add diversity.

C. Model Architecture

CollabVLA couples a VLM backbone with MoE adaptation and a diffusion-based action expert for conditional, low-latency trajectory generation (see Fig. 4).

VLM Backbone. We build on InternVL2.5 [31], which natively supports image–text interleaving. The backbone consumes a single sequence that concatenates: (i) robot observations (current RGB o^t and optionally a past frame o^{t-1}) tagged as "`[NOW]`" and "`[PAST]`"; cached embeddings for o^{t-1} are reused to reduce latency; (ii) a multimodal goal g plus optional human guidance wrapped by "`<HumanTip> ... </HumanTip>`"; (iii) proprioception p^t projected by a small MLP; and (iv) K learnable `[ACT]` queries. The model outputs: (a) a reflection string; (b) a binary ask indicator from a classifier over the pooled reflection representation; and (c) K latent action embeddings from the final hidden states of the `[ACT]` queries, which seed the action expert.

MoE Adaption. Inside the backbone, we introduce a *Mixture-of-Experts design* that enables the model to alternate between reflection and forward planning. Unlike ChatVLA [9], which

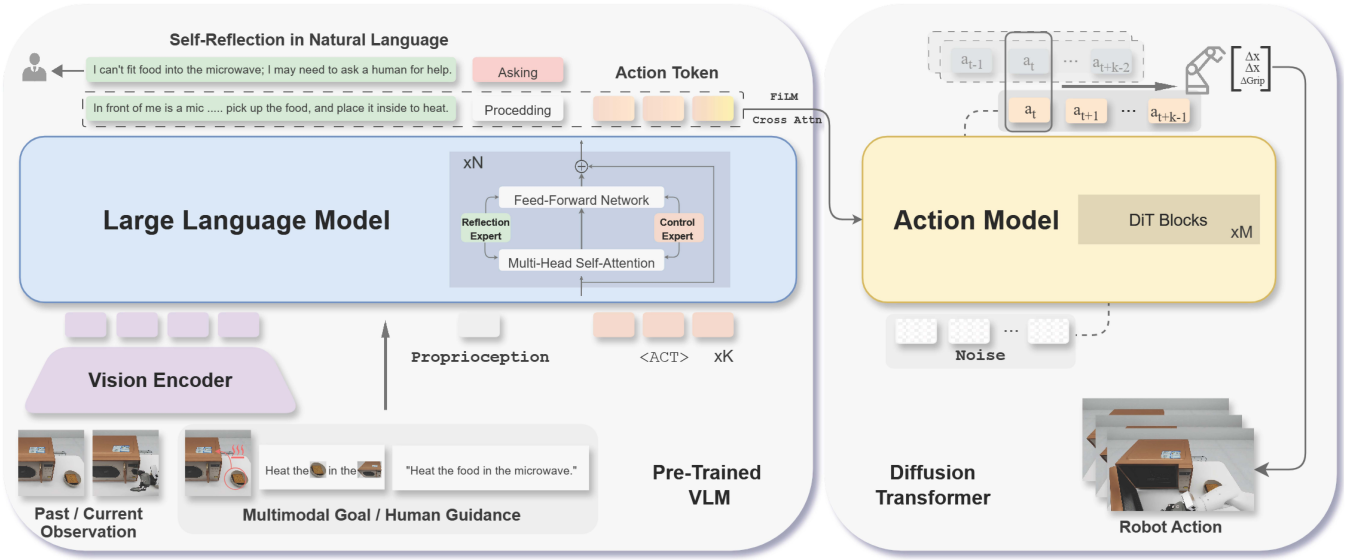

Fig. 4: **Overall architecture of CollabVLA.** The VLM backbone is augmented with LoRA-based *Control* and *Reflection Experts*, adaptively gated inside each Transformer block. Reflection provides natural-language reasoning and human queries, while latent action tokens condition a DiT via cross-attention by being injected as key–value memories that the denoiser queries at each step to retrieve fine-grained intent. The reflection embedding modulates all DiT layers through FiLM.

uses static routing to activate either a control or a conversational FFN, we adopt an adaptive gating scheme similar to InstructVLA [19]. We insert LoRA experts [32] into the linear projections of MHA and FFN: a *Control Expert* and a *Reflection Expert*. For hidden state x, the output is:

$$h = W_0 x + (B_{\text{ctrl}} A_{\text{ctrl}} x)\alpha_{\text{ctrl}}\lambda_{\text{ctrl}} + (B_{\text{ref}} A_{\text{ref}} x)\alpha_{\text{ref}}\lambda_{\text{ref}} \quad (2)$$

where W_0 is the frozen pretrained weight, $A_\cdot \in \mathbb{R}^{r\times d}$ and $B_\cdot \in \mathbb{R}^{d\times r}$ are LoRA parameters, $\alpha_\cdot$ are scaling factors, and $\lambda_\cdot$ are gating coefficients predicted by:

$$\lambda = \text{softmax}\Big(W_g \sum_j \text{softmax}(w^\top h_j)\, h_j\Big) \quad (3)$$

where W_g is a lightweight gating head, h_j are token hidden states of the current layer, and w is a learned attention vector. This adaptive gating favors the Control Expert during routine control and shifts to the Reflection Expert under uncertainty, allowing the model to balance reasoning and acting.

Diffusion-Based Action Model. We employ a Diffusion Transformer (DiT) [33] as the action generator. The VLM backbone provides two signals: (i) *latent action tokens* from learnable `[ACT]` queries, encoding structured intentions and injected as key–value memories in cross-attention for fine-grained trajectory control; and (ii) a *reflection embedding*, the final hidden states of reflection tokens, broadcast via FiLM [34] to modulate hidden activations with global semantic guidance. Starting from Gaussian noise in action space, the DiT iteratively refines trajectories conditioned on both signals, yielding outputs that are dynamically consistent, physically feasible, and reasoning-aligned.

Inference. During inference, CollabVLA executes a concise *reflect–ask/act* two-pass loop. The backbone first decodes a short reflection, pools it into an embedding, and a binary head predicts the ask indicator $\hat{q} \in \{0, 1\}$. If $\hat{q} = 1$, the reflection (which already includes the uncertainty information) is shown to the user; the reply is appended to the goal and a second forward pass is run. If $\hat{q} = 0$, the reflection embedding and K latent `[ACT]` queries are sent directly to the diffusion expert. To reduce latency, we support (i) stopping autoregressive decoding once the first `[ACT]` appears and decoding the remaining queries in parallel, and (ii) caching the reflection and `[ACT]` tokens across steps. To smooth control when new guidance arrives, we blend the current action with cached predictions using a similarity-weighted average [35]:

$$\hat{a}_t = \frac{\sum_{k=0}^{K} \exp\big(\alpha \langle a_t(o_t),\, a_t(o_{t-k})\rangle\big)\, a_t(o_{t-k})}{\sum_{j=0}^{K} \exp\big(\alpha \langle a_t(o_t),\, a_t(o_{t-j})\rangle\big)} \quad (4)$$

where $\langle \cdot, \cdot \rangle$ is cosine similarity and $\alpha = 0.1$. This yields smooth, mode-consistent trajectories with negligible overhead.

D. Training Pipeline

We adopt a two-stage recipe that first grounds visuomotor control and then enhances reflective reasoning, ensuring that reflection does not compromise action performance.

Action Grounding. This stage aligns multimodal perception with motor actions and lightweight planning language, enabling the agent to describe and execute task steps. Training uses the multimodal goal pre-training data in Sec. III-B, which provide paired observations, multimodal goals, trajectories, and rationales, while only the Control Expert is activated. The VLM backbone with Control-LoRA θ_{ctrl} outputs a planning string $\hat{r}_t$ and a set of latent action tokens z from learnable `[ACT]` queries. Unlike UniVLA [36], which first trains a separate latent-action model to annotate demonstrations, here z has no explicit ground-truth labels: it is learned implicitly as a conditioning signal for the diffusion action model to

reconstruct ground-truth trajectories. The overall loss is:

$$\mathcal{L}_{\text{Stage1}} = \lambda_{\text{lang}} \cdot \mathcal{L}_{\text{lang}} + \lambda_{\text{act}} \cdot \mathcal{L}_{\text{diff}} \tag{5}$$

where:

$$\mathcal{L}_{\text{lang}} = -\sum_{j} \log P_{\theta_{\text{ctrl}}}(r_{t,j} \mid o^t, o^{t-1}, p^t, g) \tag{6}$$

is the cross-entropy loss on forward planning traces, and:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{s,\epsilon}\Big[\big\|\epsilon - \varepsilon_\phi(\mathbf{x}_s, z, s)\big\|^2\Big], \quad \mathbf{x}_s = \alpha_s \tau_t + \sigma_s \epsilon \tag{7}$$

is the diffusion denoising loss on robot trajectories (with τ_t the ground-truth trajectory at time t). Here, $\mathcal{L}_{\text{lang}}$ updates the Control-Expert LoRA θ_{ctrl}, while $\mathcal{L}_{\text{diff}}$ updates the DiT parameters ϕ and θ_{ctrl}.

Reflection Tuning. In this stage, we freeze the diffusion action model and the VLM backbone, while jointly training the two VLM-side LoRA experts and the auxiliary heads. Training leverages a hybrid corpus that combines diverse multimodal tasks (e.g., VQA, captioning, referring expressions, retrieval, commonsense reasoning, and manipulation tasks) with the reflection-enriched data described in Sec. III-B. The objective is to unify action generation with reflective reasoning, particularly when the policy encounters failures or ambiguities, enabling the model to not only execute but also assess, revise, and query. Let θ_{ctrl} and θ_{refl} denote the Control- and Reflection-Expert LoRAs on the VLM, ψ the ask-indicator head, and W_{gate} the lightweight gating network. The overall training loss is:

$$\mathcal{L}_{\text{Stage2}} = \lambda_{\text{ref}} \cdot \mathcal{L}_{\text{ref}} + \lambda_{\text{ask}} \cdot \mathcal{L}_{\text{ask}} \tag{8}$$

Here, $\mathcal{L}_{\text{ref}}$ is the standard token-level cross-entropy for decoding the reflection string across the above task mixture, and:

$$\mathcal{L}_{\text{ask}} = -\Big[y \log \hat{y} + (1-y)\log(1-\hat{y})\Big], \quad y \in \{0,1\} \tag{9}$$

is the binary cross-entropy for the ask-indicator head. $\mathcal{L}_{\text{ref}}$ updates $(\theta_{\text{ctrl}}, \theta_{\text{refl}}, W_{\text{gate}})$; $\mathcal{L}_{\text{ask}}$ updates *only* the ask head ψ. The VLM backbone and DiT parameters remain frozen. This enables explicit reflective reasoning, calibrated uncertainty recognition, and human-query triggering, without degrading the visuomotor competence established in the former stage.

IV. Experiments

We empirically evaluate CollabVLA through four questions: *(1)* Does it preserve strong multimodal understanding? *(2)* Can it detect ambiguity and execution failures and produce insightful reflections? *(3)* How well can it interpret human-in-the-loop feedback to improve performance on long-horizon, complex tasks? *(4)* To what extent can it deliver these gains with minimal human effort and latency overhead? We benchmark across simulation and real-world tasks, and compare against state-of-the-art VLA baselines and ablations.

Summary of findings. CollabVLA preserves—and often exceeds—multimodal understanding relative to strong VLA baselines (Q1). During execution, it identifies ambiguities and failures, producing concise reflections that condition the action model or trigger human queries (Q2). With brief, free-form human hints, it improves long-horizon success while keeping latency low (Q3). Compared with strong baselines—including those that rely on heavy generative detours—it delivers a better effectiveness–efficiency trade-off (Q4). These results underscore CollabVLA as a practical, collaborative assistant that unifies perception, action, and collaboration.

A. Experimental Setup

Environments and Tasks. We evaluate along three axes. (i) *Comprehensive understanding*: we report on four multimodal-understanding benchmarks (MMMU, MMStar, OCRBench, HallBench) and four VQA benchmarks (TextVQA, DocVQA, InfoVQA, RealWorldQA), plus a 500-example *ContextReflection* set constructed from AgibotWorld [27]. The ContextReflection set spans real-world executions and GenieSim scenes and is held out from reflection-tuning data. (ii) *Simulation*: we extend the Simpler setting [29] into *SimplerCollab*, a 200-task suite covering 8 task types (see Table II) focused on long-horizon control and ambiguity resolution. We lengthen horizons, add controlled ambiguity and multiple-choice subgoals, and provide each task with a hidden script specifying environment, objects, and goals. When the VLA queries for help, its question and the script are passed to an LLM simulating the human, enabling automated human-in-the-loop evaluation. (iii) *Real-world tasks*: we use a DOBOT CR5 arm equipped with a ROBOTIQ gripper and a UR5 with an AG95 gripper. The benchmark spans five task categories with four instances each; details are provided in Sec. IV-D.

Comparisons. We compare to SOTA VLAs and ablations:

- *Vanilla VLA:* OpenVLA [4] generates action tokens autoregressively. We also build OpenVLA-Collab by fine-tuning OpenVLA on our robotic and multimodal data (following the authors' recipe) and adding only a binary ask head; reflections are decoded with the native LM head, leaving the perception–action stack unchanged.
- *Hierarchical VLAs:* π0 [39] and UniVLA [36], which explicitly decouple perception and control. We fine-tune them on our training data and add an ask head to obtain Collab variants; reflections are prompted from the LM head, and the action modules are unchanged.
- *Explicit reasoning VLAs:* ECoT [11], CoT-VLA [12], ChatVLA [9], DiVLA [30], InstrcutVLA [19], and RoboDreamer [13]. We do not build Collab variants; instead, we assess their self-generated rationales and latency under identical scenes, contrasting with CollabVLA's strategy of querying humans only when appropriate.
- *Ablations:* (i) **No-Tuning** (only Stage 1 data is used); (ii) **No-Ref** (no context reflection data); (iii) **No-FiLM** (reflections are generated but not used to condition the action model); (iv) **No-Ask** (reflection conditions the action model but question triggering is disabled); (v) **No-MoE** (the dual-expert LoRA design is removed, and a single VLM LoRA is trained and shared across Stage 1 and Stage 2); (vi) **No-MG** (no multimodal goals).

TABLE I: **Results on multimodal-understanding benchmarks, the *ContextReflection* set, and VQA.** We additionally compare against similar-sized MLLMs to quantify how much multimodal ability a VLA retains after learning robotic actions. MLLM numbers are mainly from official reports; VLA numbers from InstructVLA [19] and ChatVLA [9].

Methods	#Params	Comprehensive Understanding					VQA			
		MMMU^{Val}	MMStar	OCRB	HallB	ConRef	TextVQA	DocVQA	InfoVQA	RWQA
Multimodal Large Language Models										
Qwen2-VL [37]	2B	41.1	48.0	809	41.7	53.2	**74.9**	**88.6**	**61.4**	62.9
InternVL2.5 [31]	4B	**51.8**	58.7	**820**	**46.6**	61.4	–	–	–	–
LLaVA-OV [38]	8B	47.9	**61.9**	622	31.6	**69.4**	–	–	–	**69.9**
Magma [11]	8B	38.8	41.3	518	38.0	64.8	66.5	65.4	45.2	56.5
Vision–Language–Action Models										
DiVLA [30]	2B	17.2	21.1	294	9.0	39.2	7.5	15.2	14.7	25.2
ChatVLA [9]	2B	37.4	47.2	729	39.9	54.6	2.5	29.2	43.4	47.2
InstructVLA [19]	2B	**44.2**	**55.6**	**816**	**43.4**	**62.0**	**76.6**	**85.5**	**64.7**	**63.7**
OpenVLA [4]	7B	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ECoT [11]	7B	5.4	0.0	12	0.9	34.8	0.0	0.0	0.0	0.0
CollabVLA (**No-Tuning**)	4B	17.1	11.6	118	12.3	16.0	9.6	13.1	11.2	33.0
CollabVLA (**No-MoE**)	4B	20.3	24.7	335	18.1	29.2	10.2	18.4	16.5	58.8
CollabVLA (**No-Ref**)	4B	38.6	48.9	790	41.0	34.4	68.2	76.6	65.0	53.6
CollabVLA	4B	**49.0**	**57.2**	**814**	**43.7**	**88.6**	**77.0**	**85.7**	**65.1**	**65.3**

*Underlined values (**x**) are the best overall in each column; **bold** indicates the best within each category. A dash (−) indicates not officially reported.

TABLE II: **Results on simulation experiments.** We compare VLAs with and without explicit reasoning guidance, along with their collaborative variants. Our observations suggest that the closer an agent comes to resembling human behavior—*reading, reasoning, and interacting in a human-like manner*—the more reliable, adaptive, and effective its actions tend to be.

Methods	Fetch Robot								WidowX Robot								*Time/Dream*
	Pick Item		Move Near		Open/Close Drawer		Stack Item		Put Spoon		Put Carrot		Stack Block		Put Eggplant		
	sr	len	sr	len	sr	len	sr	len	sr	len	sr	len	sr	len	sr	len	
OpenVLA [4]	5.4	0.09	11.2	0.17	19.1	0.24	0.0	0.05	2.0	0.04	0.0	0.01	0.0	0.02	2.9	0.05	**81**/–
OpenVLA-Collab	**27.1**	**0.39**	**30.9**	**0.37**	**21.9**	**0.26**	**11.8**	**0.22**	**10.6**	**0.13**	**18.3**	**0.28**	**12.5**	**0.21**	**21.1**	**0.29**	90/**7.2**
RoboDreamer [13]	12.0	0.21	15.5	0.23	20.3	0.33	13.2	0.23	9.8	0.22	16.1	0.25	5.0	0.11	18.0	0.24	94/17.2
GR-MC [14]	20.2	0.26	26.8	0.38	22.0	0.32	14.5	0.29	10.6	0.28	17.8	0.33	9.0	0.17	21.5	0.31	74/**9.8**
MDT [18]	24.1	0.36	28.0	0.46	23.0	0.29	24.5	0.33	20.7	0.28	**29.5**	**0.48**	15.0	0.32	31.5	**0.44**	56/–
DiVLA [30]	**28.0**	**0.42**	**32.0**	**0.50**	**26.3**	**0.45**	**27.6**	**0.39**	**23.3**	**0.38**	29.4	0.43	**21.5**	**0.37**	**32.0**	0.42	**46**/–
π0 [39]	35.0	0.41	40.8	0.49	31.5	0.38	31.6	0.39	28.2	0.32	30.6	0.37	27.7	0.33	34.4	0.37	**30**/–
UniVLA [36]	34.5	0.39	38.2	0.44	29.3	0.34	30.1	0.33	33.0	0.40	31.0	0.38	28.1	0.32	34.9	0.38	36/–
π0-Collab	49.4	0.56	**57.3**	**0.63**	45.4	0.52	**38.2**	**0.43**	41.4	0.47	**42.4**	**0.47**	36.3	0.43	**45.1**	**0.49**	44/**3.6**
UniVLA-Collab	**53.2**	**0.60**	53.6	0.62	**58.1**	**0.60**	34.8	0.41	**45.0**	**0.58**	39.1	0.46	**39.6**	**0.44**	41.4	0.46	49/4.4
CollabVLA (**No-Tuning**)	18.5	0.25	23.8	0.33	15.7	0.26	12.5	0.23	13.2	0.19	11.6	0.25	10.3	0.21	19.7	0.35	53/–
CollabVLA (**No-MoE**)	23.0	0.28	28.5	0.36	21.0	0.32	16.2	0.25	18.5	0.26	24.1	0.29	18.0	0.26	29.2	0.37	49/–
CollabVLA (**No-Ref**)	27.1	0.33	29.3	0.41	22.8	0.28	16.9	0.26	20.0	0.27	23.2	0.30	20.4	0.26	29.1	0.36	37/–
CollabVLA (**No-FiLM**)	34.8	0.45	39.2	0.44	28.9	0.35	20.0	0.28	14.4	0.24	24.0	0.34	19.1	0.29	26.2	0.42	34/2.8
CollabVLA (**No-Ask**)	50.8	0.57	51.0	0.59	55.5	0.57	32.3	0.38	32.6	0.35	36.7	0.43	27.0	0.41	38.9	0.45	**32**/–
CollabVLA (**No-MG**)	55.5	0.63	59.2	0.66	60.8	0.63	40.5	0.46	37.1	0.41	44.6	0.50	31.5	0.48	47.2	0.52	38/2.3
CollabVLA	**58.5**	**0.68**	**62.2**	**0.80**	**63.8**	**0.76**	**43.5**	**0.59**	**47.1**	**0.62**	**47.5**	**0.63**	**42.5**	**0.61**	**49.2**	**0.65**	36/**1.9**

*SR denotes the success rate of tasks, and **LEN** the average completion length. **Time/Dream** is computed only on successfully completed tasks. Time is first normalized within each task across all models via clipped percentile scaling and then averaged over tasks: $\tilde{T}^{m,t} = \mathrm{clip}(T^{m,t}_{\mathrm{raw}}, p^t_5, p^t_{95})$, $T^{m,t}_{\mathrm{norm}} = \lfloor 5 + 90 \cdot \frac{\tilde{T}^{m,t} - p^t_5}{p^t_{95} - p^t_5} + 0.5 \rfloor$. Dream reports the mean number of explicit reasoning generations for non–Collab models, or human–ask calls for Collab variants.

B. Main Results

Multimodal understanding and contextual reflection. Across comprehensive understanding and VQA task, *CollabVLA* matches/exceeds strong VLAs and remains competitive with larger MLLMs (see Table I). Concretely, CollabVLA attains **49.0/57.2** on MMMU^{Val}/MMStar–near InternVL2.5 (51.8/58.7) and LLaVA-OV (47.9/61.9)–scores **814** on OCRBench (vs. 820 best) and **43.7** on HallBench (vs. 46.6 best), and achieves **88.6** on ConRef, best overall. Against the strongest VLA baseline (InstructVLA): +4.8/+1.6/−2/+0.3/+26.6 on the above metrics. On VQA, CollabVLA leads *two of four* benchmarks–TextVQA **77.0** and

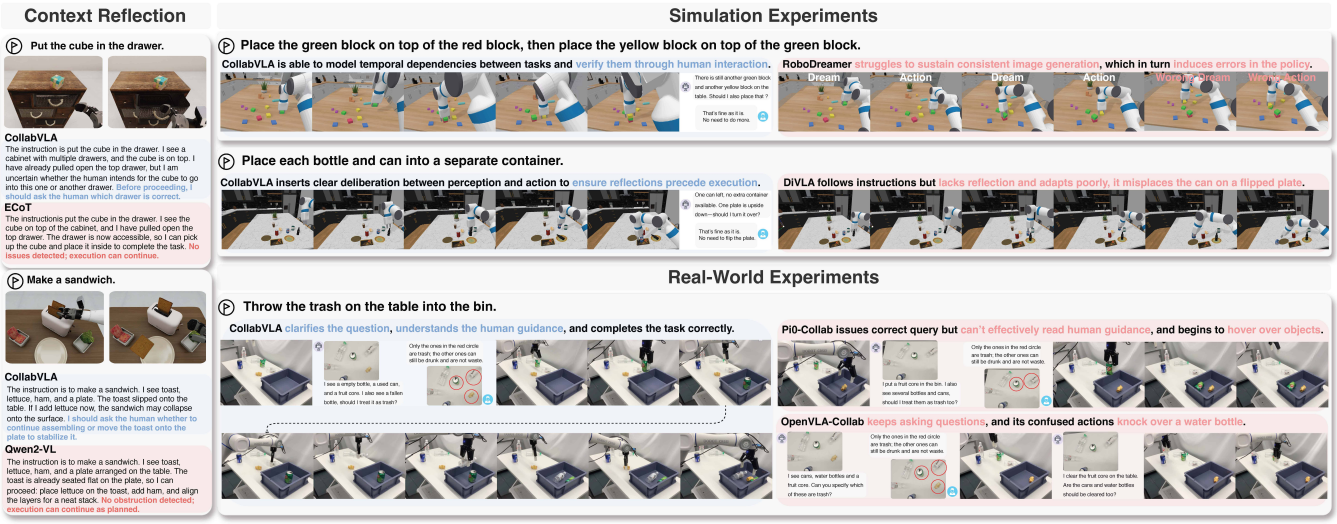

Fig. 5: **Demonstration of experiment results.** The left context-understanding task shows that CollabVLA not only handles multiple-choice reasoning better than ECoT [11], but also detects action–observation gaps to prevent long-run execution errors, outperforming general MLLMs like Qwen2-VL [37] in embodied settings. On the right, in complex, unseen stacking setting, RoboDreamer [13] struggles because its generation generalize poorly and its actions rely on inverse dynamics from generation. CollabVLA leverages self-reflection and timely oracle human guidance to refine its policy. The real-robot results further illustrate CollabVLA's superior ability to interpret multimodal human guidance and translate it into effective strategies compared with Collab-variants of $\pi 0$ [39] (which asks but struggles to translate well) and OpenVLA [4] (which over-asks).

InfoVQA **65.1**–and remains competitive on DocVQA (85.7 vs. 88.6, Qwen2-VL) and RWQA (65.3 vs. 69.9, LLaVA-OV). Thus, our two-stage training *does not sacrifice* multimodal competence and competes with strong VLMs; whereas training a VLM *only* for actions can catastrophically forget general skills (OpenVLA reports **0.0** across understanding/VQA). We attribute CollabVLA's gains to Stage-2 training, which teaches the model to detect uncertainty and compose evidence (e.g., scene perception, spatial/temporal resolution) with *robust, calibration-aware* reflections, yielding sharper grounding that transfers to diverse multimodal tasks.

Control and efficiency. On *Simpler–Collab*, *CollabVLA* attains the best SR in *all 8* subtasks (Table II). It also executes compactly (*Time*/*Dream*=**36**/**1.9**), while collab baselines help but remain less efficient: OpenVLA–Collab 90/7.2, $\pi 0$–Collab 44/3.6, UniVLA–Collab 49/4.4. Pure explicit–reasoning agents are slower and intervene more (RoboDreamer 94/17.2, GR–MC 74/9.8), showing that *selective* human queries are more cost–effective than long rationales. *CollabVLA* instead asks sparingly (1.9 queries/episode; $\sim$2–4$\times$ fewer than other collab variants) and runs faster: *OpenVLA* is slow (81/—) due to token–by–token action decoding; *RoboDreamer* (94/17.2) pays for image generation and inverse dynamics. Overall, *CollabVLA* yields the best success with low *Time* and minimal *Dream*, achieving a superior effectiveness–efficiency trade–off. We present several case studies in Fig. 5 that demonstrate how CollabVLA outperforms other methods.

C. Ablation Studies

Where does performance drop? The largest drop occurs for **No–Tuning** due to domain overfitting on robot data (e.g., Move: 23.8 vs. **62.2** for CollabVLA). Removing the dual–expert routing (**No–MoE**) induces task interference between perception and control and yields similarly large drops (e.g., Move 28.5 vs. **62.2**). Removing Stage 2 reflection supervision (**No–Ref**) or bypassing reflection conditioning (**No–FiLM**) degrades both SR and rollout quality (e.g., No–FiLM Open/Close 28.9 vs. **63.8**; Stack 19.1 vs. **42.5**), confirming that concise, on–policy *reflections* that condition the action are key. Text–only goals (**No–MG**) remain strong and sometimes rival hierarchical baselines (Pick 55.5, Move 59.2) but full multimodal goal grounding still provides a consistent lift. Finally, **No–Ask** achieves the fastest time (*Time* = **32**) but underperforms CollabVLA on SR (e.g., Pick 50.8 vs. **58.5**, Open/Close 55.5 vs. **63.8**), showing that judicious, sparse human queries are worth the small overhead.

TABLE III: **Results on real tasks.** Each method is evaluated for 5 trials per task on each arm; we report mean of total 10.

Methods	OpenVLA-Collab		π0-Collab		CollabVLA	
	SR	Score	SR	Score	SR	Score
Task1	1.6	45.0	**3.4**	**86.0**	3.2	84.5
Task2	0.8	34.0	2.2	55.5	**2.6**	**66.8**
Task3	1.1	33.7	2.4	61.7	**2.8**	**74.2**
Task4	0.3	15.9	1.4	34.5	**1.7**	**36.8**
Task5	0.9	20.2	1.9	36.0	**2.1**	**48.5**
Avg.	0.9	29.8	2.3	54.7	**2.5**	**62.2**

D. Real-World Experiments

We evaluate five task families on a DOBOT CR5 arm and a UR5 to assess robustness and cross-arm generalization: (i) object pick&place; (ii) open drawer & store items; (iii) open drawer & retrieve items; (iv) sort tabletop items; and (v) clear countertop waste. Each family has four difficulty tiers with

stepwise credit—*Basic* (10), *Distractors* (20), *Clarification* (30), and *Long–horizon* (40). We report *SR* as the number of fully completed instances per family (0–4) and *Score* as the summed credit normalized to $[0, 100]$.

From Table III, *CollabVLA* achieves the best average SR/Score (**2.5**/**62.2**), leading in **4**/5 tasks. Compared with π0–Collab, CollabVLA improves the mean by **+0.2** SR and **+7.5** Score, with the largest gains on Tasks 2–4 ($+0.4/+11.3$, $+0.4/+12.5$, $+0.3/+2.3$). Against OpenVLA–Collab, the margins widen to **+1.6** SR and **+32.4**. These trends indicate that CollabVLA's reflect–ask/act loop with FiLM–conditioned control turns brief human hints into more reliable progress under clutter, ambiguity, and multi–step objectives, while remaining competitive on straightforward manipulation.

V. Conclusion

We introduce CollabVLA, a self-reflective VLA that reasons explicitly, reflects on uncertainty, and integrates lightweight human feedback in real time. Our two-stage recipe—grounding perception in action, then tuning reflection without harming control—yields consistent gains in success and interpretability across tasks. Looking ahead, CollabVLA could be further strengthened by integrating tactile/force sensing, improving epistemic uncertainty with a better-calibrated ask trigger, and evolving collaboration from queries to proactive task allocation and coordination. These directions target a key bottleneck for deployment: knowing when to proceed, when to verify, and when to involve a human, under real-world noise and distribution shift.

Nonetheless, we are confident that the present instantiation already constitutes a distinctive, practically impactful advance: uniting reflection with action, it delivers consistent gains and supports real-time human collaboration—moving embodied agents toward the next frontier: *robots that not only act, but reflect, adapt, and partner with humans as genuine teammates.*

References

[1] J.-B. Alayrac *et al.*, "Flamingo: a visual language model for few-shot learning," 2022. [Online]. Available: https://arxiv.org/abs/2204.14198

[2] X. Zhai *et al.*, "Sigmoid loss for language image pre-training," 2023. [Online]. Available: https://arxiv.org/abs/2303.15343

[3] A. Brohan *et al.*, "Rt-1: Robotics transformer for real-world control at scale," 2023. [Online]. Available: https://arxiv.org/abs/2212.06817

[4] M. J. Kim *et al.*, "Openvla: An open-source vision-language-action model," 2024. [Online]. Available: https://arxiv.org/abs/2406.09246

[5] X. Li, *et al.*, "Vision-language foundation models as effective robot imitators," 2024. [Online]. Available: https://arxiv.org/abs/2311.01378

[6] A. Brohan *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," 2023. [Online]. Available: https://arxiv.org/abs/2307.15818

[7] A. Liang *et al.*, "Clam: Continuous latent action models for robot learning from unlabeled demonstrations," 2025. [Online]. Available: https://arxiv.org/abs/2505.04999

[8] X. Chen *et al.*, "villa-x: Enhancing latent action modeling in vision-language-action models," 2025. [Online]. Available: https://arxiv.org/abs/2507.23682

[9] Z. Zhou *et al.*, "Chatvla: Unified multimodal understanding and robot control with vision-language-action model," 2025. [Online]. Available: https://arxiv.org/abs/2502.14420

[10] J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," 2023. [Online]. Available: https://arxiv.org/abs/2201.11903

[11] M. Zawalski *et al.*, "Robotic control via embodied chain-of-thought reasoning," 2025. [Online]. Available: https://arxiv.org/abs/2407.08693

[12] Q. Zhao *et al.*, "Cot-vla: Visual chain-of-thought reasoning for vision-language-action models," 2025. [Online]. Available: https://arxiv.org/abs/2503.22020

[13] S. Zhou *et al.*, "Robodreamer: Learning compositional world models for robot imagination," 2024. [Online]. Available: https://arxiv.org/abs/2404.12377

[14] P. Li *et al.*, "Gr-mg: Leveraging partially annotated data via multi-modal goal-conditioned policy," 2024. [Online]. Available: https://arxiv.org/abs/2408.14368

[15] S. Yao *et al.*, "React: Synergizing reasoning and acting in language models," 2023. [Online]. Available: https://arxiv.org/abs/2210.03629

[16] N. Shinn *et al.*, "Reflexion: Language agents with verbal reinforcement learning," 2023. [Online]. Available: https://arxiv.org/abs/2303.11366

[17] C. Fan *et al.*, "Interleave-vla: Enhancing robot manipulation with interleaved image-text instructions," 2025. [Online]. Available: https://arxiv.org/abs/2505.02152

[18] M. Reussand *et al.*, "Multimodal diffusion transformer: Learning versatile behavior from multimodal goals," 2024. [Online]. Available: https://arxiv.org/abs/2407.05996

[19] S. Yang *et al.*, "Instructvla: Vision-language-action instruction tuning from understanding to manipulation," 2025. [Online]. Available: https://arxiv.org/abs/2507.17520

[20] A. Z. Ren *et al.*, "Robots that ask for help: Uncertainty alignment for large language model planners," 2023. [Online]. Available: https://arxiv.org/abs/2307.01928

[21] K. Liang *et al.*, "Introspective planning: Aligning robots' uncertainty with inherent task ambiguity," 2025. [Online]. Available: https://arxiv.org/abs/2402.06529

[22] L. Zheng *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," 2023. [Online]. Available: https://arxiv.org/abs/2306.05685

[23] N. Sun *et al.*, "Assistantx: An llm-powered proactive assistant in collaborative human-populated environment," 2025. [Online]. Available: https://arxiv.org/abs/2409.17655

[24] H. Walke *et al.*, "Bridgedata v2: A dataset for robot learning at scale," 2024. [Online]. Available: https://arxiv.org/abs/2308.12952

[25] E. Collaboration *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models," 2025. [Online]. Available: https://arxiv.org/abs/2310.08864

[26] A. Khazatsky *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," 2025. [Online]. Available: https://arxiv.org/abs/2403.12945

[27] AgiBot-World-Contributors *et al.*, "Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems," 2025. [Online]. Available: https://arxiv.org/abs/2503.06669

[28] B. Liu, *et al.*, "Libero: Benchmarking knowledge transfer for lifelong robot learning," 2023. [Online]. Available: https://arxiv.org/abs/2306.03310

[29] X. Li *et al.*, "Evaluating real-world robot manipulation policies in simulation," 2024. [Online]. Available: https://arxiv.org/abs/2405.05941

[30] J. Wen *et al.*, "Diffusion-vla: Generalizable and interpretable robot foundation model via self-generated reasoning," 2025. [Online]. Available: https://arxiv.org/abs/2412.03293

[31] Z. Chen *et al.*, "Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling," 2025. [Online]. Available: https://arxiv.org/abs/2412.05271

[32] E. J. Hu *et al.*, "Lora: Low-rank adaptation of large language models," 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

[33] W. Peebles and S. Xie, "Scalable diffusion models with transformers," 2023. [Online]. Available: https://arxiv.org/abs/2212.09748

[34] E. Perez *et al.*, "Film: Visual reasoning with a general conditioning layer," 2017. [Online]. Available: https://arxiv.org/abs/1709.07871

[35] Q. Li *et al.*, "Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation," 2024. [Online]. Available: https://arxiv.org/abs/2411.19650

[36] Q. Bu *et al.*, "Univla: Learning to act anywhere with task-centric latent actions," 2025. [Online]. Available: https://arxiv.org/abs/2505.06111

[37] P. Wang *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," 2024. [Online]. Available: https://arxiv.org/abs/2409.12191

[38] B. Li *et al.*, "Llava-onevision: Easy visual task transfer," 2024. [Online]. Available: https://arxiv.org/abs/2408.03326

[39] K. Black *et al.*, "π_0: A vision-language-action flow model for general robot control," 2024. [Online]. Available: https://arxiv.org/abs/2410.24164