AgiBot World Colosseo: Large-scale Manipulation Platform for Scalable and Intelligent Embodied Systems

Team AgiBot-World*

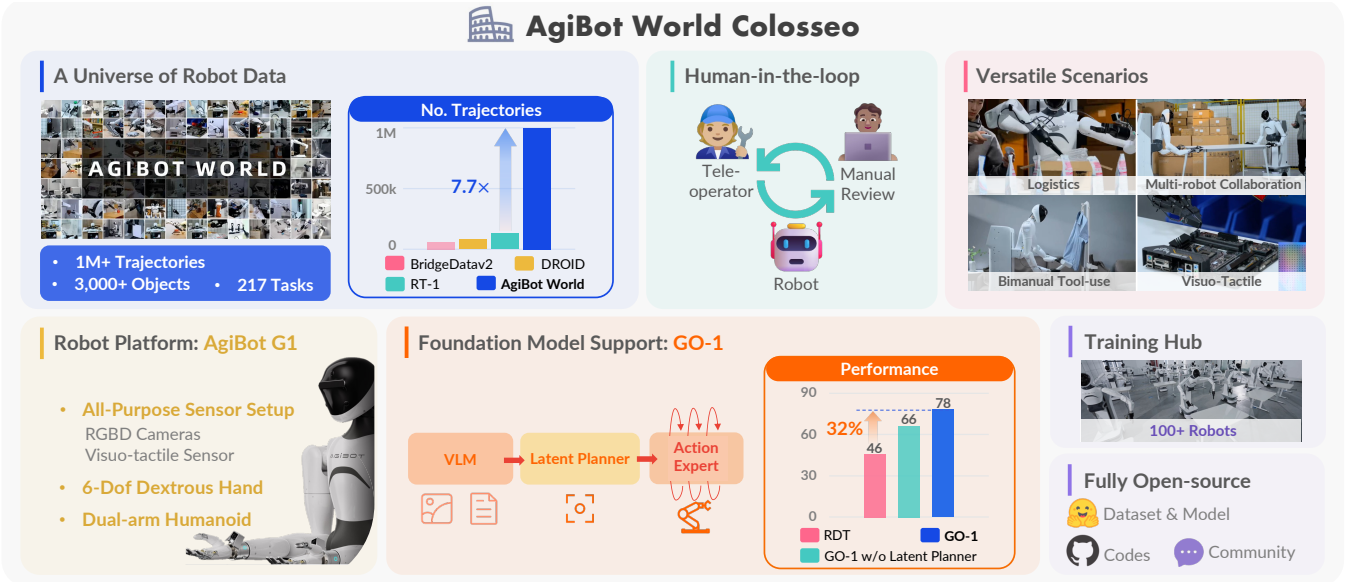

Fig. 1: Introducing **AgiBot World Colosseo**, an open-sourced large-scale manipulation platform comprising data, models, benchmarks and ecosystem. AgiBot World stands out for its unparalleled scale and diversity compared to prior counterparts. A suite of 100 dual-arm humanoid robots, namely AgiBot G1, is deployed to capture multimodal mobile manipulation demonstrations. Data quality is guaranteed by proficient teleoperators and the rigorous human-in-the-loop verification. We further propose a generalist policy, **G**enie **O**perator-1 (GO-1), with the latent action planner. It achieves unified training across diverse data corpus with an impressive scalable performance of 32% gain compared to prior arts.

***Abstract*— We explore how scalable robot data can address real-world challenges for generalized robotic manipulation. Introducing AgiBot World, a large-scale platform comprising over 1 million trajectories across 217 tasks in five deployment scenarios, we achieve an order-of-magnitude increase in data scale compared to existing datasets. Accelerated by a standardized collection pipeline with human-in-the-loop verification, AgiBot World guarantees high-quality and diverse data distribution. It is extensible from grippers to dexterous hands and visuo-tactile sensors for fine-grained skill acquisition. Building on top of data, we introduce Genie Operator-1 (GO-1), a novel generalist policy that leverages latent action representations to maximize data utilization, demonstrating predictable performance scaling with increased data volume. Policies pre-trained on our dataset achieve an average performance improvement of 30% over those trained on Open X-Embodiment, both in in-domain and out-of-distribution scenarios. GO-1 exhibits exceptional capability in real-world dexterous and long-horizon tasks, achieving over 60% success rate on complex tasks and outperforming prior RDT approach by 32%. By open-sourcing the dataset, tools, and models, we aim to democratize access to large-scale, high-quality robot data, advancing the pursuit of scalable and general-purpose intelligence.**

* This work is a collaborative effort between Shanghai AI Lab and AgiBot Inc. For detailed authorship, please visit the official website: https://agibot-world.com/colosseo-contributors

I. Introduction

Manipulation is a cornerstone task in robotics, enabling the agent to interact with and adapt to the physical world. While significant progress has been made in general-purpose foundational models for natural language processing [1] and computer vision [2], robotics lags behind due to the difficulty of (high-quality) data collection. In the controlled lab setting, simple tasks such as pick-and-place have been well studied [3], [4]. Yet for the open-set real-world setting, tasks spanning from fine-grained object interaction, mobile manipulation to collaborative tasks, remains a formidable challenge [5]. These tasks require not only physical dexterity but also the ability to generalize across diverse environment and scenarios, a merit beyond the reach of current robotic systems. The widely accepted reason is the lack of high-quality data—unlike images and text, which are abundant and standardized, robotic datasets suffer from fragmented clips due to heterogeneous hardware and unstandardized collection procedure, leading to low-quality and inconsistent outcome. In this work we ask, *how could we resolve the real-world complexity effectively by scaling up real-world robot data?*

Recent efforts, such as Open X-Embodiment (OXE) [6], have addressed by aggregating and standardizing existing datasets. Despite advancements on large-scale cross-embodiment learning, the resulting policy is constrained within naive, short-horizon tasks and can weakly generalize to out-of-domain scenarios [4]. DROID [7] collected expert data through crowd-sourcing from diverse real-life scenes. The absence of data quality assurance (with human feedback) and the reliance on a constrained hardware setup (*i.e.,* featuring fixed, single-arm robots), limit its real-world applicability and broader effectiveness. More recently, Lin *et al.* [8] explored scaling laws governing generalizability across intra-category objects and environments, albeit limited to a few simple, single-step tasks. These efforts represent a notable advancement toward developing generalist policies, moving beyond the traditional focus on single-task learning within narrow domains [9], [3]. Nevertheless, existing robot learning datasets remain constrained by their reliance on short-horizon tasks in highly controlled laboratory environments, failing to adequately capture the complexity and diversity inherent in real-world manipulation tasks. To achieve general-purpose robotic intelligence, it is essential to develop datasets that scale in size and diversity while capturing real-world variability, supported by general-purpose humanoid robots for robust skill acquisition, a standardized data collection pipeline with assured quality, and carefully curated tasks reflecting real-world challenges.

As depicted in Fig. 1, we introduce **AgiBot World Colosseo**, a full-stack large-scale robot learning platform curated for advancing bimanual manipulation in scalable and intelligent embodied systems. A full-scale 4000-square-meter facility is constructed to represent five major domains—domestic, retail, industrial, restaurant, and office environment—all dedicated to high-fidelity data collection in authentic everyday scenarios. With over 1 million trajectories collected from 100 real robots, AgiBot World offers unprecedented diversity and complexity. It spans over 100 real-world scenarios, addressing challenging tasks such as fine-grained manipulation, tool usage, and multi-robot synergistic collaboration. Unlike prior datasets, AgiBot World dataset collection is carried out with a fully standardized pipeline, ensuring high data quality and scalability, while incorporating human-in-the-loop verification to guarantee reliability. Our hardware setup includes mobile base humanoid robots with whole-body control, dexterous hands, and visuo-tactile sensors, enabling rich, multimodal data collection. Each episode is meticulously designed, featuring multiple camera views, depth information, camera calibration, and language annotations for both the overall task and each individual sub-steps. This well-rounded hardware setup, combined with various long-horizon, real-world tasks, opens new avenues for developing next-generation generalist policies and fosters diverse future research in robotics.

Our experimental results highlight the transformative potential of the AgiBot World dataset. Policies pre-trained on our dataset achieve an average success rate improvement of 30% compared to those trained on the prior large-scale robot dataset OXE [6]. Notably, even when utilizing only a fraction of our dataset—equivalent to 1/10 of the data volume in hours compared to OXE—the generalizability of pretrained policies is elevated by 18%. These findings underscore the dataset's efficacy in bridging the gap between controlled laboratory environments and real-world robotic applications. Following our dataset, to address the limitations of previous robot foundation models that heavily rely on in-domain robot datasets, we present Genie Operator-1 (GO-1), a novel generalist policy that utilizes latent action representations to enable learning from heterogeneous data and efficiently bridges general-purpose vision-language models (VLMs) with robotic sequential decision-making. Through unified pre-training on web-scale data, spanning human videos to our high-quality robot dataset, GO-1 achieves superior generalization and dexterity, outperforming prior generalist policies such as RDT [10] and our variant without latent action planner. Moreover, we demonstrate that GO-1's performance exhibits robust scalability with increasing dataset size, underscoring its potential for sustained advancement as larger datasets become available.

Beyond its immediate impact, AgiBot World lays a strong foundation for future research in robotic manipulation. By open-sourcing the dataset, toolchain, and pre-trained models, we aim to foster community-wide innovation, enabling researchers to explore more authentic and diverse applications from household assistant to industrial automation. AgiBot World is more than yet another dataset; it is a step toward scalable, general-purpose robotic intelligence, empowering robots to tackle the complexities of the real world.

Contribution. 1) We construct AgiBot World dataset, a multifarious robot learning dataset accompanied by open-source tools to advance research on policy learning at scale. As a pioneering initiative, AgiBot World employs an inclusive optimized pipeline, from scene configuration, task design, data collection, to human-in-the-loop verification, which ensures unparalleled data quality. 2) We propose GO-1, a robot foundation policy using latent action representations to unlock web-scale pre-training on heterogeneous data. Empowered by AgiBot World dataset, it outperforms prior generalist policies in generalization and dexterity, with performance scaling predictably with dataset size.

Limitation. All evaluations are conducted in real-world scenarios. We are currently developing the simulation environment, aligning with the real-world setup and aiming to reflect real-world policy deployment outcome. It would thereby facilitate fast and reproducible evaluation.

II. Related Work

Data scaling in robotics. Robot learning datasets from automated scripts or human teleoperation have enabled policy learning, with early efforts like RoboTurk [19] and BridgeData [12] offering small-scale datasets with 2.1k and 7.2k trajectories, respectively. Larger datasets, such as RT-1 [14] (130k trajectories), expand scopes yet remain limited to few environments and skills. Open X-Embodiment [6] aggregates various datasets into a unified format, growing

Dataset	Traj.	Skill	Scene	Detailed Annotation	Cam. Calibration	Arm Type	Dex. Hand	Failure Recovery	Human-in-the-loop	Collection
RoboNet [11]	162k	n/a	10	✗	✗	Single	✗	✗	✗	scripted
BridgeData [12]	7.2k	4	12	✗	✗	Single	✗	✗	✗	human teleop
BC-Z [13]	26k	3	1	✗	✗	Single	✗	✗	✗	human teleop
RT-1 [14]	130k	8	2	✗	✗	Single	✗	✗	✗	human teleop
RH20T [15]	13k	33	7	✗	✓	Single	✗	✗	✗	human teleop
RoboSet [16]	98.5k	6	11	✗	✗	Single	✗	✗	✗	30% human / 70% scripted
BridgeData V2 [17]	60.1k	13	24	✗	✗	Single	✗	✗	✗	85% human / 15% scripted
DROID [7]	76k	86	564	✗	✓	Single	✗	✗	✗	human teleop
RoboMIND [18]	55k	36	n/a	✗	✓	Single+Dual	✓	✗	✗	human teleop
Open X-Embodiment [6]	1.4M	217	311	(✗)	✗	Single+Dual	✗	✗	✗	dataset aggregation
AgiBot World Dataset	**1M+**	87	106	✓	✓	Dual	✓	✓	✓	human teleop

TABLE I: **Comparison to existing datasets.** AgiBot World features the largest number of trajectories *to date*. We replicate real-world environment at a 1:1 scale for the industrial and retail scenarios, which are barely present before. Extensive human annotations are offered, including item, scene, skill (sub-task segmented), and task-level annotations. Notably, to expand data applicability and potential, we include imperfect data (*i.e.,* failure recovery data with annotated error states) and tasks with dexterous hands. To ensure data quality, we adopt a human-in-the-loop philosophy: the policy learning is performed on collected demonstrations. The deployment results are adopted as feedback to improve the collection protocol.

to more than 2.4 million trajectories, as a consequence it suffers from significant variability in embodiments, observation perspectives, and inconsistent data quality, limiting its overall effectiveness. More recently, DROID [7] moves towards scaling up scenes for greater diversity by crowdsourcing demonstrations yet falls short in data scale and quality control. Prior datasets above generally face limitations in data scale, task practicality, and scenario naturalness, compounded by inadequate quality assurance and hardware restrictions, which impedes generalist policy training. As shown in Tab. I, our dataset addresses these gap adequately. We build a data collection facility spanning five scenarios to reconstruct real-world diversity and authenticity. With over 1 million trajectories gathered by skilled teleoperators through rigorous verification protocols, AgiBot World utilizes humanoid robots equipped with visuo-tactile sensors and dexterous hands to enable multimodal demonstrations, setting it apart from previous efforts. Unlike Pumacay *et al.* [20], which serves as a simulation benchmark for evaluating generalization, what we propose is a full-stack platform with data, models, benchmarks, and ecosystem.

Policy learning at scale. Robotic foundation models often co-evolve with the development of dataset scale, equipping robots with escalating general-purpose capabilities through diverse, large-scale training. Several prior arts use web-scale video only to facilitate policy learning given the limited scale of action-labeled robot datasets [21], [22], [23]. Another line of work lies in the use of large, end-to-end models trained on robot trajectories with robotics data scaling up [4], [24], [14], [25]. For instance, RDT [10] employs Diffusion Transformers, initially pre-trained on heterogeneous multirobot datasets and fine-tuned on over 6k dual-arm trajectories, showcasing the benefits of pre-training on diverse sources. π_0 [26] uses a pre-trained VLM backbone and a flow-based action expert, advancing dexterous manipulation for complex tasks like laundry. LAPA [27] introduces the use of latent actions as pre-training targets; however, its latent planning capability is not preserved for downstream tasks. Building on a variety of innovative ideas from recent research, we advance the field by transferring web-scale knowledge to robotic control through the adaptation of vision-language models (VLMs) with latent actions, leveraging both human videos and robot data for scalable training. Our work demonstrates how the integration of a latent action planner enhances long-horizon task execution and enables more efficient policy learning, significantly improving upon existing generalist policies.

III. AgiBot World: Platform and Data

AgiBot World is a full-stack and open-source embodied intelligence ecosystem. Based on the hardware platform developed by us, AgiBot G1, we construct AgiBot World—an open-source robot manipulation dataset collected by more than 100 homogeneous robots, providing high-quality data for challenging tasks spanning a wide spectrum of real-life scenarios. The latest version contains 1,001,552 trajectories, with a total duration of 2976.4 hours, covering 217 specific tasks, 87 skills, and 106 scenes. We go beyond basic tabletop tasks such as *pick-and-place* in lab environments; instead, concentrate on real-world scenarios involving dual-arm manipulation, dexterous hands, and collaborative tasks. AgiBot World aims to provide an inclusive benchmark to drive the future development of advanced and robust algorithms.

We plan to release *all* resources to enable the community to build upon AgiBot World. **The dataset is available under the CC BY-NC-SA 4.0 license**, along with the model checkpoints, code for data processing and policy training.

A. Hardware: A Versatile Humanoid Robot

The hardware platform is the cornerstone of AgiBot World, determining the lower limit of its quality. The standardization of hardware is also the key to streamlining distributed data collection and ensuring reproducible results. We meticulously develop a novel hardware platform for AgiBot World, distinguished by visuo-tactile sensors, durable 6-DoF dexterous hands with humanoid configuration.

As illustrated in Fig. 1, our robotic platform features dual 7-DoF arms, a mobile chassis, and an adjustable waist. The

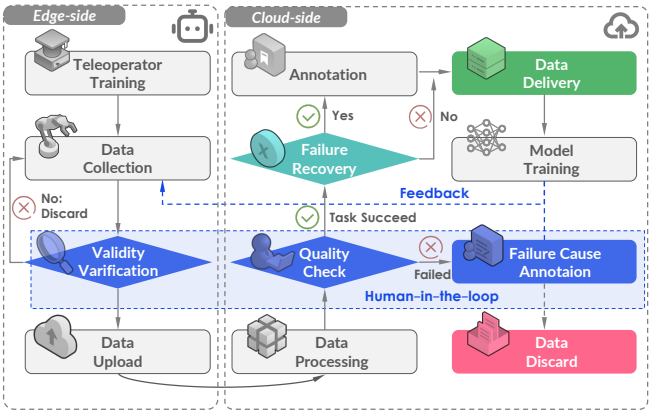

Fig. 2: **Data collection pipeline.** AgiBot World embraces a human-in-the-loop framework to ensure high quality, enriched with detailed annotations and error recovery behaviors. Human feedback plays a critical role not only in post-collection review but also in actively guiding the data collection process, which is largely overlooked in prior efforts.

end effectors are modular, allowing for the use of either a standard gripper or a 6-DoF dexterous hand, depending on task requirements. For tasks necessitating tactile feedback, a gripper equipped with visuo-tactile sensors is utilized. The robot is outfitted with eight cameras: an RGB-D camera and three fisheye cameras for the front view, RGB-D or fisheye cameras mounted on each end-effector, and two fisheye cameras positioned at the rear. Image observations and proprioceptive states, including joint and end-effector positions, are recorded at a control frequency of 30 Hz.

We employ two teleoperation systems: VR headset control and whole-body motion capture control. The VR controller maps the hand gesture to the end-effector translation and rotation, which is subsequently converted to joint angles through inverse kinematics. The thumbsticks and buttons on the controller enable robot base and body movement, while the trigger buttons control end-effector actuation. However, the VR controller restricts the dexterous hand to only a few predefined gestures. To extensively unlock our robot's capabilities, we adapt a motion capture system which records the data of human joints, including the fingers, and maps them to robot posture, enabling more nuanced control, including individual finger movements, torso pose, and head orientation. This system provides posture flexibility and execution precision that are required in achieving more complex manipulation tasks.

B. Data Collection: Protocol and Quality

The data collection session, as shown in Fig. 2, can be broadly divided into three phases. (1) Before formally commencing data collection, we first conduct preliminary data acquisition to validate the feasibility of each task and establish corresponding collection standards. (2) After feasibility validation and review of the collection standards, skilled teleoperators arrange the initial scene and formally begin data collection according to the established standards. All data undergoes an initial validity verification locally, such as verifying the absence of missing frames. Once the data is confirmed to be complete, it is uploaded to the cloud for the next phase. (3) During post-processing, the data annotators will verify whether each episode meets the collection standards established in phase 1 and provide language annotations.

Failure recovery. During data collection, teleoperators may occasionally commit errors, such as inadvertently dropping objects while manipulating the robotic arms. However, they are often able to recover from these errors and successfully complete the task without requiring a full reconfiguration of the setup. Rather than discarding such trajectories, we retain them and manually annotate each with corresponding failure reasons and timestamps. These trajectories, referred to as *failure recovery* data, constitute approximately one percent of the dataset. We consider them invaluable for achieving policy alignment [28] and failure reflection [29], essential for advancing the next generation of robot foundation models.

Human-in-the-loop. Concurrent with feedback collection from data annotators, we adopt a human-in-the-loop approach to assess and refine data quality. This process involves an iterative cycle of collecting a small set of demonstrations, training a policy, and deploying the resulting policy to evaluate data availability. Based on the policy's performance, we iteratively refine the data collection pipeline to address identified gaps or inefficiencies. For instance, during real-world deployment, the model exhibits prolonged pauses at the onset of actions, aligning with data annotator feedback highlighting inconsistent transitions and excessive idle time in the collected data. In response, we revise the data collection protocols and introduce a post-processing step to eliminate idle frames, thereby enhancing the dataset's overall utility for policy learning. This feedback-driven methodology ensures continuous improvement in data quality.

C. Dataset Statistics and Analysis: Beyond Scale

AgiBot World is developed through a large-scale data collection facility, which spans over 4,000 square meters. This extensive environment contains over 3,000 unique objects in a variety of scenes, meticulously designed to reflect real-world settings. The dataset covers a wide range of scenarios and scene setups, ensuring both scale and diversity in the pursuit of generalizable robot policy.

Reconstructing the diversity of the real world. Key statistics of our dataset are presented in Fig. 3. AgiBot World provides extensive coverage across five key domains: domestic, retail, industrial, restaurant, and office environments. Within each domain, we further define specific scene categories. For instance, the domestic domain includes detailed environments such as bedrooms, kitchens, living rooms, and balconies, while the retail domain features distinct areas like shelving units and fresh produce sections. Our dataset also features over 3,000 distinct objects, systematically categorized across various scenes. These objects span a wide range of everyday items, including food, furniture, clothing, electronic devices, and more. The distribution of object

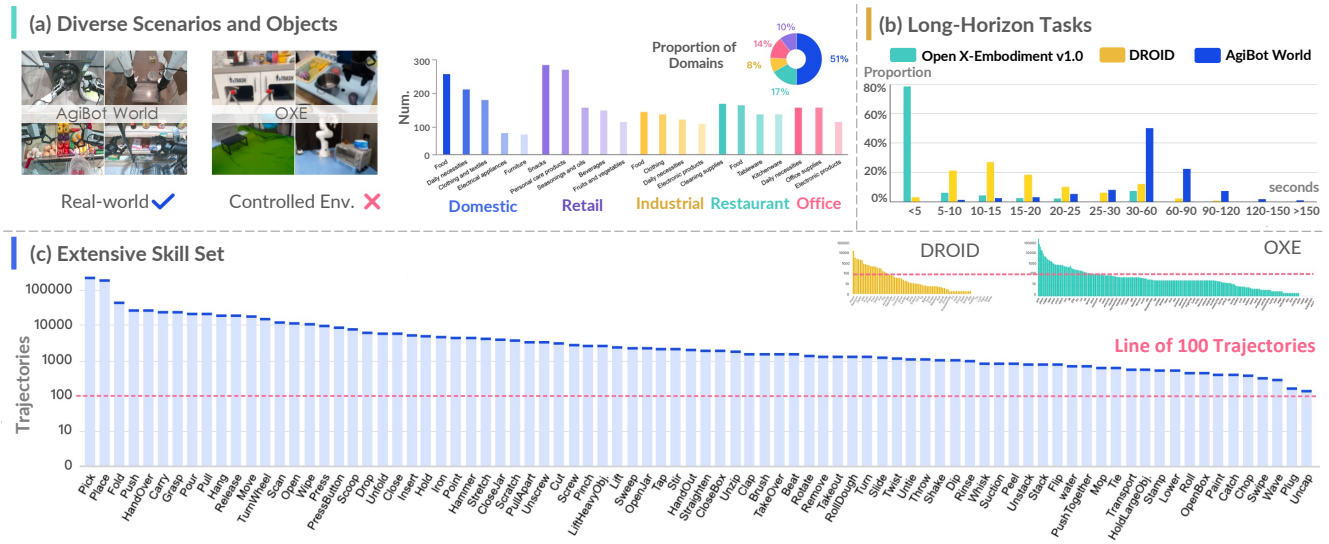

Fig. 3: **Dataset Statistics. a)** AgiBot World dataset covers the vast majority of robotic application scenarios, as well as a wide range of interactive objects. **b)** Our dataset features long-horizon tasks, with the majority of trajectories ranging from 30s to 60s. In contrast, widely used datasets, such as DROID, primarily consist of trajectories ranging from 5s to 20s, while OXE v1.0 predominantly contains trajectories within 5s. **c)** AgiBot World dataset focuses on valuable atomic skills, spanning a wide spectrum of skills, each supported by a minimum of 100 trajectories (red dashed line above).

categories, as illustrated in Fig. 3(a), highlights the relative frequency of different object types within each scene.

Long-horizon manipulation. A distinguishing feature of the AgiBot World dataset is its emphasis on long-horizon manipulation. As shown in Fig. 3(b), prior datasets predominantly focus on tasks involving single atomic skills, with most trajectories lasting no more than 5 seconds. In contrast, AgiBot World is built upon continuous and complete tasks composed by multiple atomic skills, like "make a coffee". Trajectories in our dataset typically span approximately 30 seconds, some of which last over 2 minutes. We also provide key-frame and instruction annotation for each sub-step to facilitate policy learning in such challenging scenarios.

Comprehensive skill coverage. In terms of task design, while generic atomic skills, such as "pick-and-place", dominate the majority of tasks, we have intentionally incorporated tasks that emphasize less frequently used but highly valuable skills, such as "`chop`" and "`plug`" (as shown in Fig. 3(c)). This ensures that our dataset adequately represents a broad spectrum of skills, providing sufficient data for each to support robust policy learning.

IV. AgiBot World: Model

To effectively utilize our high-quality AgiBot World dataset and enhance the policy's generalizability, we propose a hierarchical **Vi**sion-**L**anguage-**L**atent-**A**ction (ViLLA) framework with three training stages, as depicted in Fig. 4. Compared to Vision-Language-Action (VLA) model where action is vision-language conditioned, the ViLLA model predicts latent action tokens, conditioned on the generation of subsequent robot control actions.

In Stage 1, we project consecutive images into a latent action space by training an encoder-decoder latent action model (LAM) on internet-scale heterogeneous data. This allows the latent action to serve as an intermediate representation, bridging the gap between general image-text inputs and robotic actions. In Stage 2, these latent actions act as pseudo-labels for the latent planner, facilitating embodiment-agnostic long-horizon planning and leveraging the generalizability of the pre-trained VLM. Finally, in Stage 3, we introduce the action expert and jointly train it with the latent planner to support the learning of dexterous manipulation.

A. Latent Action Model

Despite considerable advancements in gathering diverse robot demonstrations, the volume of action-labeled robot data remains limited relative to web-scale datasets. To broaden the data pool by incorporating internet-scale human videos lacking action labels and cross-embodiment robot data, we employ latent actions [30] in Stage 1 to model the inverse dynamics of consecutive frames. This approach enables the transfer of real-world dynamics from heterogeneous data sources into universal manipulation knowledge.

To extract latent actions from video frames $\{I_t, I_{t+H}\}$, the latent action model is constructed around an inverse dynamics model-based encoder $\mathbf{I}(z_t|I_t, I_{t+H})$ and a forward dynamics model-based decoder $\mathbf{F}(I_{t+H}|I_t, z_t)$. The encoder employs a spatial-temporal transformer [31] with casual temporal masks following Bruce *et al.* [30], while the decoder is a spatial transformer that takes the initial frame and discretized latent action tokens $z_t = [z_t^0, ..., z_t^{k-1}]$ as input, with k set to 4. The latent action tokens are quantized using a VQ-VAE objective [32], with a codebook of size $|C|$.

B. Latent Planner

With the aim of establishing a solid foundation for scene and object understanding and general reasoning ability, the ViLLA model harnesses a VLM pre-trained on web-scale

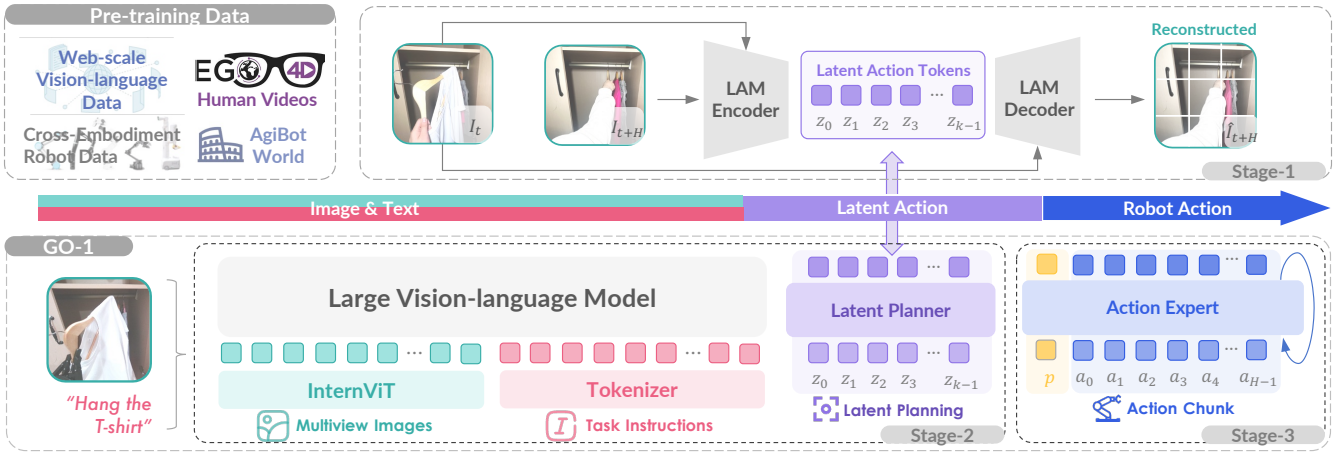

Fig. 4: **We propose GO-1**, a generalist policy featuring general reasoning and long-horizon planning capabilities. The latent action model (LAM) learns universal action representations from web-scale video data (*i.e.,* human videos from Ego4D), and quantizes them into discrete latent action tokens. The latent planner conducts temporal reasoning through latent action prediction, bridging the gap between image-text inputs and robot actions generated by the action expert.

vision-language data and incorporates a latent planner for embodiment-agnostic planning within the latent action space. We use InternVL2.5-2B [33] as the VLM backbone due to its strong transfer learning capabilities. The two-billion parameter scale has proven effective for robotic tasks in our preliminary experiments, as well as in prior studies [10], [26]. Multiview image observations are first encoded using InternViT before being projected into the language space. The latent planner consists of 24 transformer layers, which enable layer-by-layer conditioning from the VLM backbone with full bidirectional attention.

Specifically, given multiview input images $\left(I_t^h, I_t^l, I_t^r\right)$ (typically from the head, left wrist, and right wrist) at timestep t, along with a language instruction l describing the ongoing task, the latent planner predicts latent action tokens: $\mathbf{P}\left(z_t | I_t^h, I_t^l, I_t^r, l\right)$, with supervision produced by the LAM encoder based on the head view: $z_t := \mathbf{I}(I_t^h, I_{t+H}^h)$. Since the latent action space is orders of magnitude smaller than the discretized low-level actions used in OpenVLA [4], this approach also facilitates the efficient adaptation of general-purpose VLMs into robot policies.

C. Action Expert

To achieve high-frequency and dexterous manipulation, Stage 3 integrates an action expert that utilizes a diffusion objective to model the continuous distribution of low-level actions [34]. Although the action expert shares the same architectural framework as the latent planner, their objectives diverge: the latent planner generates discretized latent action tokens through masked language modeling, while the action expert regresses low-level actions via an iterative denoising process. Both expert modules are conditioned hierarchically on preceding modules, including the action expert itself, ensuring coherent integration and information flow within the dual-expert system.

The action expert decodes low-level action chunks, denoted by $A_t = [a_t, a_{t+1}, ..., a_{t+H}]$ with $H = 30$, using proprioceptive state p_t over an interval of H timesteps: $\mathbf{A}\left(A_t | I_t^h, I_t^l, I_t^r, p_t, l\right)$. During inference, the VLM, latent planner, and action expert are synergistically combined within the generalist policy GO-1, which initially predicts k latent action tokens and subsequently conditions the denoising process to produce the final control signals.

V. Experiment and Analysis

We evaluate the real-world performance of policies pre-trained on different data sources including the AgiBot World dataset, demonstrating the effectiveness credited from the GO-1 model in policy learning.

A. Experiment Setup

1) Evaluation Tasks

Here we choose a comprehensive set of tasks that span various dimensions of policy capabilities from AgiBot World for evaluation, including **tool-usage** (Wipe Table), **deformable objects manipulation** (Fold Shorts), **human-robot interaction** (Handover Bottle), **language-following** (Restock Beverage), etc. Moreover, we design 2 unseen scenarios for each task, covering position generalization, visual distractors, and language generalization, delivering thorough generalization evaluations for policies. The evaluated tasks, also partially shown in Fig. 5, are: 1) "Restock Bag": Pick up the snack from the cart and place it on the supermarket shelf; 2) "Table Bussing": Clear tabletop debris into the trash can; 3) "Pour Water": Grasp the kettle handle, lift the kettle and pour water into the cup; 4) "Restock Beverage": Pick up the bottled beverage from the cart and place it on the supermarket shelf; 5) "Fold Shorts": Fold the shorts laid flat on the table in half twice; 6) "Wipe Table": Clean water spills using the sponge. **Scoring rubrics.** The evaluation metric employs a normalized score, computed as the average across 10 rollouts per task, scenario, and method. Each episode scores 1.0 for full

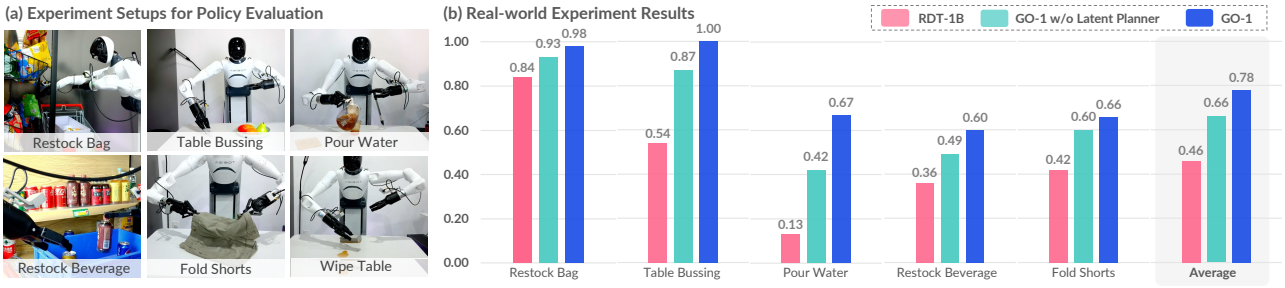

Fig. 5: **Is GO-1 a more powerful robot generalist policy?** We evaluate GO-1 against previous generalist policy RDT-1B and our baseline without the latent planner, with all policies pre-trained on AgiBot World `beta`. Across all tasks and comparisons, GO-1 outperforms baselines by a large margin. The incorporation of latent planner boosts performance on complex tasks such as "Fold Shorts" and improves generalizability in task "Restock Beverage" in great extent.

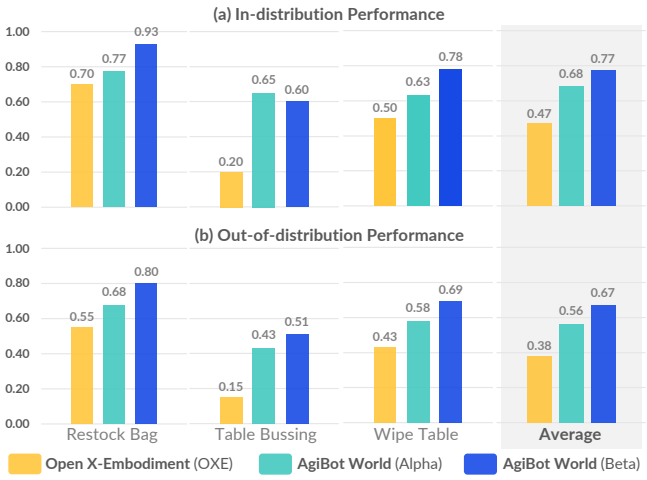

Fig. 6: **Does AgiBot World dataset improve policy performance and generalizability?** Policies pre-trained on our dataset outperform those trained on OXE in both seen (0.77 *v.s.* 0.47) and out-of-distribution scenarios (0.67 *v.s.* 0.38).

success, with fractional scores for partial success, enabling a nuanced performance assessment.

2) Implementation Details

The AgiBot World `alpha` represents the partial subset of our dataset, constituting approximately 14% of the full-version, AgiBot World `beta` (a.k.a. last row in Tab. I). Following the completion of the third-stage pre-training, the pre-trained GO-1 exhibits basic competency in task completion. Unless otherwise specified, we further enhance the model by fine-tuning it using high-quality, task-specific demonstrations, enabling adaptation to new tasks for evaluation. For GO-1, fine-tuning is conducted with a learning rate of 2e-5, a batch size of 768, and 30,000 optimization steps.

B. Does AgiBot World boost policy learning at scale?

We choose the open-source RDT [10] model to study how much the AgiBot World dataset can help policy learning. The task completion scores for three tasks are detailed in Fig. 6. Models pre-trained on the AgiBot World dataset demonstrate a significant improvement in the "Table Bussing" task, nearly tripling performance. On average, the completion

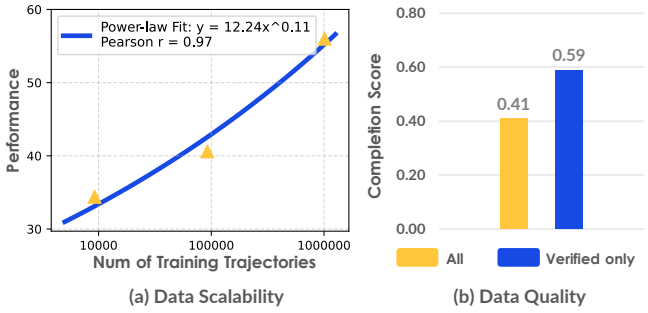

Fig. 7: **Further analysis on**: a) how model performance scales with data size, and b) the impact of filtering undesirable data through manual review on policy learning.

score increases by 0.30 and 0.29 for in-distribution and out-of-distribution setups, respectively. Notably, the AgiBot World `alpha` dataset, despite having a significantly smaller data volume than OXE (*e.g.*, 236h compared to ∼2000h), achieves a higher success rate, underscoring the exceptional data quality of our dataset.

C. Is GO-1 a more capable generalist policy?

We evaluate GO-1 on five tasks of varying complexity, categorized by their visual richness and task horizon. The results, as shown in Fig. 5, are averaged over 30 trials per task, with 10 trials conducted in a seen setup and 20 trials under variations or distractions. GO-1 significantly outperforms RDT, particularly in tasks such as "Pour Water", which demands robustness to object positions, and "Restock Beverage", which requires visual robustness and instruction-following capabilities. The inclusion of the latent planner in the ViLLA model further improves performance, resulting in an average improvement of 0.12 task completion score.

D. Does GO-1's ability scale with data size?

To investigate whether a power-law scaling relationship exists between the size of pre-training data and policy capability, we conduct an analysis using 10% subsets of the `alpha`, 100% `alpha`, and `beta` dataset, where the number of training trajectories are ranged from 9.2k to 1M. We evaluate the out-of-the-box performance of resulting policies on four seen tasks in pre-training. As shown in Fig. 7(a), the

policy's performance exhibits a predictable power-law scaling relationship with the number of trajectories, supported by a Pearson correlation coefficient of $r = 0.97$.

E. How does data quality impact policy learning?

We explore the impact of quality checks introduced in our human-in-the-loop data collection on policy learning. Specifically, we provide an ablation study by fine-tuning an RDT model pre-trained on the `alpha` dataset using both verified (528 trajectories) and unverified (482 trajectories) data from the "Wipe Table" task. As shown in Fig. 7(b), being larger in quantity does not necessarily translate to improved performance, while a smaller set of human-verified data yields a 0.18 boost in the completion score, underscoring the importance of high-quality data for policy learning.

VI. Conclusion

We introduce AgiBot World, an open-source ecosystem aimed at democratizing access to large-scale, high-quality robot learning datasets. It is complete with toolchains and foundation models to advance embodied general intelligence through community collaboration. Our dataset distinguishes itself through unparalleled scale, diversity, and quality, underpinned by carefully crafted tasks. Policy learning evaluations confirm AgiBot World's value in enhancing performance and generalizability. To further explore its impact, we develop GO-1, a generalist policy utilizing latent actions for web-scale pre-training. GO-1 excels in real-world complex tasks, outperforming existing generalist policies and demonstrating scalable performance with increased data volume. We invite the broader community to collaborate in fostering an ecosystem and maximizing the potential of our extensive dataset.

References

[1] OpenAI, "GPT-4 Technical Report," *arXiv preprint arXiv:2303.08774*, 2023.

[2] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, *et al.*, "SAM 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[3] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion Policy: Visuomotor policy learning via action diffusion," in *RSS*, 2023.

[4] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, *et al.*, "OpenVLA: An open-source vision-language-action model," in *CoRL*, 2024.

[5] J. Cui and J. Trinkle, "Toward next-generation learned robot manipulation," in *Science Robotics*, 2021.

[6] A. Padalkar, A. Pooley, A. Jain, A. Bewley, A. Herzog, A. Irpan, A. Khazatsky, A. Rai, A. Singh, A. Brohan, *et al.*, "Open X-Embodiment: Robotic learning datasets and RT-X models," in *ICRA*, 2024.

[7] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, *et al.*, "DROID: A large-scale in-the-wild robot manipulation dataset," in *RSS*, 2024.

[8] F. Lin, Y. Hu, P. Sheng, C. Wen, J. You, and Y. Gao, "Data scaling laws in imitation learning for robotic manipulation," in *ICLR*, 2025.

[9] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," in *RSS*, 2023.

[10] S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu, H. Su, and J. Zhu, "RDT-1B: a diffusion foundation model for bimanual manipulation," in *ICLR*, 2025.

[11] S. Dasari, F. Ebert, S. Tian, S. Nair, B. Bucher, K. Schmeckpeper, S. Singh, S. Levine, and C. Finn, "RoboNet: Large-scale multi-robot learning," in *CoRL*, 2019.

[12] F. Ebert, Y. Yang, K. Schmeckpeper, B. Bucher, G. Georgakis, K. Daniilidis, C. Finn, and S. Levine, "Bridge data: Boosting generalization of robotic skills with cross-domain datasets," in *RSS*, 2022.

[13] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn, "BC-Z: Zero-shot task generalization with robotic imitation learning," in *CoRL*, 2022.

[14] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu, *et al.*, "RT-1: Robotics transformer for real-world control at scale," in *RSS*, 2023.

[15] H.-S. Fang, H. Fang, Z. Tang, J. Liu, J. Wang, H. Zhu, and C. Lu, "RH20T: A robotic dataset for learning diverse skills in one-shot," in *RSS Workshops*, 2023.

[16] H. Bharadhwaj, J. Vakil, M. Sharma, A. Gupta, S. Tulsiani, and V. Kumar, "RoboAgent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking," in *ICRA*, 2024.

[17] H. R. Walke, K. Black, T. Z. Zhao, Q. Vuong, C. Zheng, P. Hansen-Estruch, A. W. He, V. Myers, M. J. Kim, M. Du, *et al.*, "BridgeData v2: A dataset for robot learning at scale," in *CoRL*, 2023.

[18] K. Wu, C. Hou, J. Liu, Z. Che, X. Ju, Z. Yang, M. Li, Y. Zhao, *et al.*, "RoboMIND: Benchmark on multi-embodiment intelligence normative data for robot manipulation," *arXiv preprint arXiv:2412.13877*, 2024.

[19] A. Mandlekar, Y. Zhu, A. Garg, J. Booher, M. Spero, A. Tung, J. Gao, J. Emmons, A. Gupta, E. Orbay, S. Savarese, and L. Fei-Fei, "RoboTurk: A crowdsourcing platform for robotic skill learning through imitation," in *CoRL*, 2018.

[20] W. Pumacay, I. Singh, J. Duan, R. Krishna, J. Thomason, and D. Fox, "The colosseum: A benchmark for evaluating generalization for robotic manipulation," in *RSS*, 2024.

[21] Y. Du, S. Yang, B. Dai, H. Dai, O. Nachum, J. Tenenbaum, D. Schuurmans, and P. Abbeel, "Learning universal policies via text-guided video generation," in *NeurIPS*, 2024.

[22] K. Black, M. Nakamoto, P. Atreya, H. R. Walke, C. Finn, A. Kumar, and S. Levine, "Zero-shot robotic manipulation with pre-trained image-editing diffusion models," in *ICLR*, 2024.

[23] Q. Bu, J. Zeng, L. Chen, Y. Yang, G. Zhou, J. Yan, P. Luo, H. Cui, Y. Ma, and H. Li, "Closed-loop visuomotor control with generative expectation for robotic manipulation," in *NeurIPS*, 2024.

[24] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, *et al.*, "RT-2: Vision-language-action models transfer web knowledge to robotic control," in *CoRL*, 2023.

[25] D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, T. Kreiman, C. Xu, J. Luo, *et al.*, "Octo: An open-source generalist robot policy," in *RSS*, 2024.

[26] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, *et al.*, "A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

[27] S. Ye, J. Jang, B. Jeon, S. Joo, J. Yang, B. Peng, A. Mandlekar, R. Tan, Y.-W. Chao, B. Y. Lin, *et al.*, "Latent action pretraining from videos," in *ICLR*, 2025.

[28] Z. Zhang, K. Zheng, Z. Chen, J. Jang, Y. Li, C. Wang, M. Ding, D. Fox, and H. Yao, "GRAPE: Generalizing robot policy via preference alignment," *arXiv preprint arXiv:2411.19309*, 2024.

[29] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: Language agents with verbal reinforcement learning," *NeurIPS*, 2023.

[30] J. Bruce, M. D. Dennis, A. Edwards, J. Parker-Holder, Y. Shi, E. Hughes, M. Lai, A. Mavalankar, R. Steigerwald, C. Apps, *et al.*, "Genie: Generative interactive environments," in *ICML*, 2024.

[31] M. Xu, W. Dai, C. Liu, X. Gao, W. Lin, G.-J. Qi, and H. Xiong, "Spatial-temporal transformer networks for traffic flow forecasting," *arXiv preprint arXiv:2001.02908*, 2020.

[32] A. Van Den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *NeurIPS*, 2017.

[33] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu, *et al.*, "Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling," *arXiv preprint arXiv:2412.05271*, 2024.

[34] Q. Bu, H. Li, L. Chen, J. Cai, J. Zeng, H. Cui, M. Yao, and Y. Qiao, "Towards synergistic, generalized, and efficient dual-system for robotic manipulation," *arXiv preprint arXiv:2410.08001*, 2024.