Towards Scaling Robotic Manipulation: Leveraging World Knowledge as Automated Teleoperator

Paper-ID [127]

***Abstract*—The scalability of robot learning is significantly hindered by the cost of collecting real-world manipulation data. Despite recent progress in imitation learning, existing datasets are still predominantly acquired through expert teleoperation, which requires continuous low-level human control, specialized hardware, and carefully calibrated environments. This labor-intensive paradigm severely limits data throughput, generalizability, and long-term scalability. We propose PaPA: Pick-and-Place Anything, a scalable robotic data collection framework for real-world manipulation that addresses this bottleneck through autonomous trajectory generation with minimal human supervision. PaPA leverages world knowledge encoded in vision-language models to ground open-vocabulary language instructions into explicit geometric representations, enabling the autonomous generation of robot-executable manipulation trajectories. This approach reduces human supervision to lightweight post-hoc binary verification, eliminating the need for continuous teleoperation. Through extensive real-world experiments, we demonstrate that datasets collected by PaPA match the quality of expert teleoperation for imitation learning, achieving a policy success rate of 97.5%, while maintaining an efficient data collection rate of approximately two episodes per minute. These results show that high-quality robotic manipulation data can be acquired at scale without continuous expert supervision, positioning PaPA as a practical and effective alternative to teleoperation-centric data collection. More broadly, PaPA demonstrates how world knowledge can be systematically transformed into scalable real-world robotic data, offering a concrete step toward data-driven robot learning at scale.**

I. Introduction

Large-scale data-driven learning has become a central part of the modern artificial intelligence, in particular natural language processing and computer vision [55, 11, 57, 7]. Empirical evidence demonstrates that model performance improves predictably with increasing dataset size, model capacity, and training compute [38, 28]. Moreover, these successes in data scaling enable qualitatively new capabilities such as in-context learning and cross-domain generalization [68, 7]. As data scaling has emerged as a dominant paradigm, numerous fields have attempted to transfer its successes to their own domains, including robotics [43, 8, 9].

While recent work demonstrates that imitation learning can benefit from large-scale datasets [52, 6, 5], it remains a challenge to achieve robotic data at a scale comparable to vision and language domains (e.g., 10 trillion tokens of text and 5 billion images) [38, 28, 10, 59]. This difficulty stems from the fundamental nature of robotic data: unlike the web-sourced data used in vision and language domains, robotic data collection requires active environmental interaction to capture the physical dynamics and the causal relationship between actions and observations [44, 17]. In practice, active

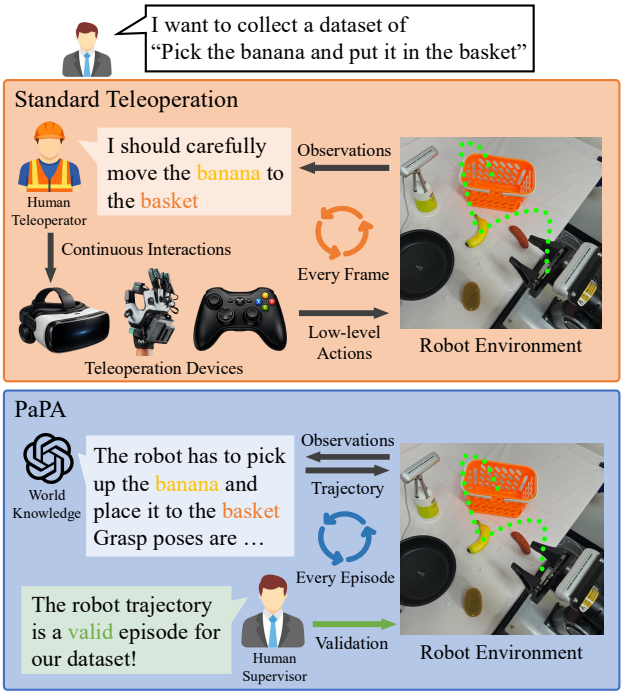


Fig. 1: **PaPA: Pick-and-Place Anything.** Comparison between standard teleoperation-based data collection and PaPA. While conventional approaches require continuous low-level control by human experts, PaPA leverages world knowledge to automatically generate grounded manipulation trajectories from visual observations and language instructions. Human involvement is reduced to lightweight trajectory validation, significantly lowering the burden of data collection.

interaction typically requires human-in-the-loop control or manually designed systems (e.g., teleoperation or simulation environments with scripted policies) [50, 56, 36]. Such cost-intensive collection methods yield datasets with limited environmental diversity, which in turn restrict the deployment of learned policies to a narrow set of environments [54, 21, 4].

We argue that the major bottleneck in scaling robotic data is rooted in the dominant data collection strategy: teleoperation. As shown in Fig. 1, teleoperation-based data collection typically requires specialized hardware, precise calibration, and continuous human labor [64, 75, 49, 73]. For even seemingly simple tasks like pick-and-place, human experts must continuously control the robot through low-level actions while observing action outcomes [14, 13, 18]. This continuous interaction requires active human involvement, preventing continuous and around-the-clock data collection. Empirically,

collecting even a minimally sufficient dataset for behavior cloning often requires over an hour of human labor. At larger scales, robotic data such as DROID [39] curate 350 hours of robotic interaction data over a year-long data collection effort involving 50 expert human operators. In other words, substantial time and human overhead in teleoperation-based data collection highlights the need for a shift away from teleoperation toward more scalable robotic data collection paradigms.

Various approaches have been proposed to shift from the standard teleoperation system. One line of work explores alternative data collection devices, such as hand-held gripper devices, which enable data collection with simpler hardware setup [15, 60]. Although these interfaces alleviate the cost of specialized teleoperation hardware, they still depend on expert demonstrations and continuous human-in-the-loop interaction [71, 20]. Another line of work investigates policy-assisted teleoperation frameworks that partially automate task execution with a motor policy [16]. However, these methods still require human intervention at unpredictable points, necessitating continuous supervision and thus limiting their scalability [18]. More recently, language-based teleoperation has been proposed to enable robotic data collection by non-expert users [48, 2]. Nevertheless, the need for repetitive human instruction in these approaches leaves the fundamental scalability bottleneck unresolved [37]. These works call for a scalable data collection framework that removes continuous human intervention while still producing high-quality robotic data.

In this paper, we propose PaPA: Pick-and-Place Anything, an autonomous robotic data collection framework designed to overcome the data scaling bottleneck. We focus on pick-and-place manipulation tasks, a constrained but practically dominant class of behaviors in real-world robotic datasets. As illustrated in Fig. 1, PaPA converts language instructions and visual observations into high-quality manipulation trajectories. The process begins with perception via world knowledge, where task-relevant semantics are extracted from open-vocabulary instructions and visual observations using off-the-shelf multimodal models. By grounding language into explicit semantic masks, PaPA leverages vast world knowledge to ensure the robot precisely understands "what" and "where" to manipulate, even in diverse, unseen environments. These masks are then passed to the trajectory planning stage, which converts semantic information into executable robot actions. This decoupled approach allows for the generation of high-quality interaction episodes while maintaining the flexibility to handle a wide range of objects. Finally, PaPA employs a lightweight human supervision step where a supervisor only provides simple binary labels for task success. This post-hoc verification significantly reduces the human effort per data point, enabling robotic data collection to scale efficiently while ensuring that only successful, high-fidelity trajectories are incorporated into the final dataset. Through this pipeline, PaPA provides an efficient and scalable approach to building high-quality manipulation datasets with minimal human effort.

We conduct a comprehensive set of experiments to evaluate PaPA as a scalable robotic data collection framework. Our evaluation focuses on three key aspects. First, we assess the quality of PaPA-collected data by evaluating imitation learning performance and trajectory quality against prior baselines. Second, we analyze data collection efficiency, showing how PaPA improves episode collection rate while minimizing human effort. Third, we perform ablation studies on trajectory generation strategies to validate our design choices. Together, these results show that PaPA yields human-like data that effectively trains imitation learning policies while improving collection efficiency and reducing human effort. Ablations further highlight the importance of trajectory diversity and lightweight supervision.

In short, our contributions are threefold:

- We propose PaPA, a scalable robotic data collection framework for real-world manipulation that overcomes the bottleneck of labor-intensive data scaling via autonomous trajectory generation with minimal human supervision.
- We design a world knowledge-driven trajectory generation pipeline that grounds open-vocabulary language instructions into explicit geometric representations, enabling autonomous generation of high-quality manipulation trajectories with post-hoc binary verification.
- We demonstrate that PaPA produces human-like, high-quality dataset, matching human teleoperation in imitation learning performance (97.5%) while achieving an efficient data collection rate of approximately two episodes per minute.

II. Related Works

A. Scalable Robotic Data Collection

Scaling robotic data remains a fundamental bottleneck due to the high cost of active interaction [43]. While large-scale datasets collected via human-in-the-loop teleoperation have improved generalization [54, 21], they are inherently limited by the need for continuous human involvement and specialized hardware [39, 8]. To reduce this burden, several approaches bootstrap data collection using an exploratory policy to autonomously generate trajectories for further training [44, 56]. However, these methods typically require a robust base policy to ensure safety and struggle to generalize to open-world scenes with unseen objects and novel tasks. Similarly, learning from human activity videos offers a web-scale supervision source [61, 72], yet converting these action sequences into precise, robot-compatible control remains a significant technical challenge [42]. Apart from data sources, another line of work focuses on improving data collection interfaces to reduce operator burden. These range from specialized hardware [70, 23, 22] and simplified handheld setups [15, 60, 20, 71] to higher-level language-based guidance [48, 2]. While these interfaces lower the entry barrier, they still necessitate iterative human interaction and often fail to scale to open-world settings where robots must interact with novel tasks without

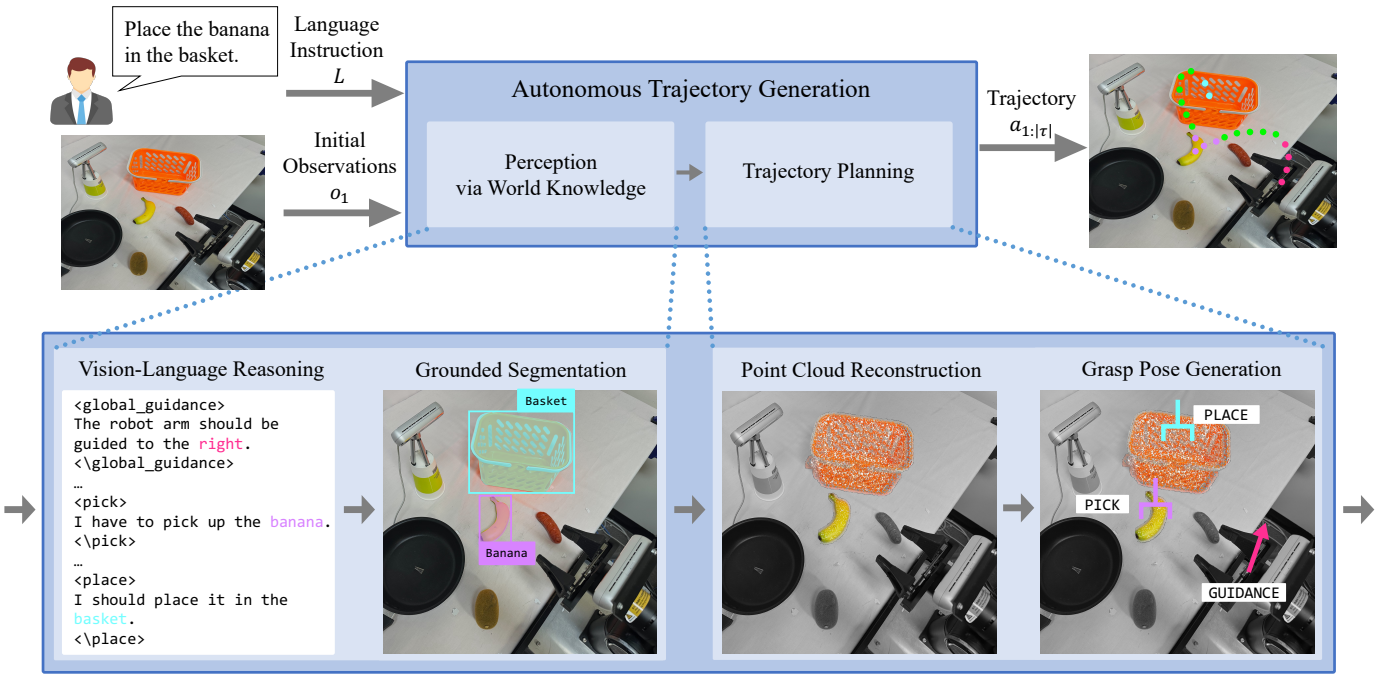


Fig. 2: **Overview of autonomous trajectory generation in PaPA. (Top)** Given a language instruction and initial visual observations, PaPA autonomously generates a grounded pick-and-place trajectory without teleoperation or task-specific supervision. The pipeline is decomposed into two stages: *Perception via World Knowledge* and *Trajectory Planning*. **(Bottom)** In *Perception via World Knowledge*, vision-language reasoning provides global semantic guidance (e.g., target object and goal direction), while segmentation grounds these semantics in the visual scene. *Trajectory Planning* then reconstructs object geometry from the grounded perception, estimates grasp poses, and generates a robot-executable manipulation trajectory.

expert guidance. In contrast, PaPA addresses the open-world challenge by autonomously generating executable trajectories for novel objects, restricting human involvement to lightweight binary success labeling.

B. Vision-Language Model in Robotic Manipulation

Recent advances in vision-language models (VLMs) have enabled the transfer of vast internet-scale knowledge to robotic manipulation [55, 51, 24]. One line of work employs vision-language reasoning to generate executable manipulation plans, represented as code-based policies [29, 31, 46]. While these methods offer high-level task planning, their robustness often depends on environment-specific assumptions on environment states or predefined interaction primitives [47, 65, 69]. Alternatively, other approaches ground instructions into explicit spatial representations, such as affordance maps [62, 33] or keypoints [45, 74, 30]. Although these methods perform explicit geometric reasoning, their performance critically depends on accurate 3D perception, making them sensitive to real-world sensing noise [32]. More recently, end-to-end models that build upon vision-language models to directly map vision-language inputs to low-level actions (i.e., vision-language-action models) have emerged [9, 52, 40]. Although these models introduce a unified framework connecting vision-language reasoning with low-level action control, their effectiveness depends on large-scale vision-language-action paired datasets, making robotic data scaling a persistent challenge [6, 5]. To mitigate this data bottleneck, VLMs have been leveraged to guide data synthesis via simulation [12, 67, 66], yet these often rely on simulation assets, necessitating additional engineering to bridge the sim-to-real gap [19, 26, 35]. By contrast, PaPA leverages world knowledge encoded in VLMs to extract task-relevant semantic priors from unconstrained, real-world language instructions and visual observations. By grounding these priors into actionable representations, PaPA enables scalable real-world robotic manipulation data collection.

III. PaPA: Pick-and-Place Anything

In this section, we present the problem formulation and an overview of PaPA, followed by a detailed description of each component. We first describe the perception module, which leverages world knowledge to extract task-relevant information. We then introduce the trajectory planning procedure built upon these perceptions. Finally, we explain the lightweight labeling process used to curate a high-quality dataset with PaPA.

A. Problem Formulation

PaPA is an autonomous dataset collection framework for manipulation tasks to minimize human expert teleoperation. PaPA aims to construct a dataset $\mathcal{D} = \{\mathcal{E}_n \,|\, \mathcal{E}_n = (\tau_n, L_n),\ n \in \{1, \ldots, |\mathcal{D}|\}\}$, where each episode $\mathcal{E}_n$ consists of a robot trajectory τ_n paired with a corresponding natural language instruction L_n. A trajectory $\tau_n = \{(a_t, o_t) \,|\, t \in \{1, \ldots, |\tau_n|\}\}$ is composed of robot actions a_t and multi-view visual observations $o_t = \{I_t^1, \ldots, I_t^C\}$, where t denotes the time index, C the numbers of the cameras, and I^c the image of a single camera c.



Fig. 3: **Vision-language-grounded object geometry reconstruction.** Given a textual object description and multi-view images, PaPA reconstructs an object point cloud by combining vision-language grounding with classical multi-view geometry. A pretrained vision-language model produces grounded segmentation masks, which define silhouette cones under the visual hull formulation. Intersecting these cones across camera views yields the reconstructed object point cloud.

B. Overview

For each data collection episode, PaPA takes a language instruction L and initial visual observations o_1 as input, and autonomously generates a robot-executable pick-and-place trajectory $a_{1:|\tau|}$. As illustrated in Fig. 2, PaPA first perceives the scene by leveraging the inherent world knowledge of VLMs to extract task-relevant information. This semantic information is subsequently transformed into explicit geometric representations g. Based on these geometries, key grasp poses are identified to generate a sequence of manipulation actions $a_{1:|\tau|}$. Upon execution, the resulting candidate trajectory $\tilde{\tau} = \{(a_t, o_t) \,|\, t \in \{1, \ldots, |\tilde{\tau}|\}\}$ is evaluated by a human supervisor. Successful rollouts are appended to the dataset $\mathcal{D}$ until the target size is reached, while failed attempts are discarded.

C. Perception via World Knowledge

As shown in Fig. 2, we decouple high-level semantic reasoning from low-level visual grounding by separating semantic inference from pixel-level localization: a vision-language model reasons over the language instruction and visual observations, while detection modules provide explicit semantic masks in the image.

Vision-Language Reasoning. The VLM is queried to infer two key components of task-relevant semantic information: global guidance and target descriptors. The global guidance provides a preferred approach direction for the manipulator, serving as a prior to encourage approach directions consistent with the semantic intent while facilitating data diversity. Simultaneously, the VLM generates concise textual descriptors for both the object to be grasped and the target placement region. These descriptors allow PaPA to interpret implicit instructions (e.g., "buying yellow fruits") by resolving semantic ambiguities within the specific context of the current scene.

Grounded Segmentation. Based on the textual descriptors from the VLM, a detection module (i.e., groundedSAM [58]) extracts semantic masks for the target objects across all camera views. These masks isolate task-relevant entities from the multi-view observations o_1, providing the spatial geometry g that is used to construct 3D geometric representations for trajectory generation. By masking out irrelevant scene elements, this step ensures that the subsequent trajectory generation focuses on the objects essential for task completion.

D. Trajectory Planning

During the trajectory planning stage, we project 2D segmentation masks into 3D object-centric point clouds, which we then use to estimate optimal grasp poses and key manipulation waypoints (see Fig. 2).

Point Cloud Reconstruction. To obtain object geometry in a computationally efficient yet robust manner, PaPA reconstructs 3D point clouds using a visual hull algorithm based on silhouette cone intersections [41]. Fig. 3 illustrates how segmentation masks from multiple views define silhouette cones whose intersections form the object geometry. Specifically, for each camera $c \in \{1, \ldots, C\}$ with known extrinsics T_c^w, the 2D semantic mask m_c is back-projected into the 3D world frame. Each pixel $(u, v) \in m_c$ defines a ray $r_c(d)$ that forms a view-dependent silhouette cone $\mathcal{C}_c$:

$$\mathcal{C}_c = \left\{ r_c(d) \;\middle|\; r_c(d) = T_c^w \cdot \begin{pmatrix} u \\ v \\ 1 \\ 1 \end{pmatrix} d,\ (u,v) \in m_c,\ d > 0 \right\}. \tag{1}$$

The 3D object geometry $\mathcal{V}$ is then defined as the maximal volume consistent with all observed masks:

$$\mathcal{V} = \bigcap_{c=1}^{C} \mathcal{C}_c. \tag{2}$$

These reconstructed point clouds then serve as object geometries for subsequent grasp planning.

Grasp Pose and Trajectory Generation. Given the reconstructed object point clouds, PaPA samples grasp poses to

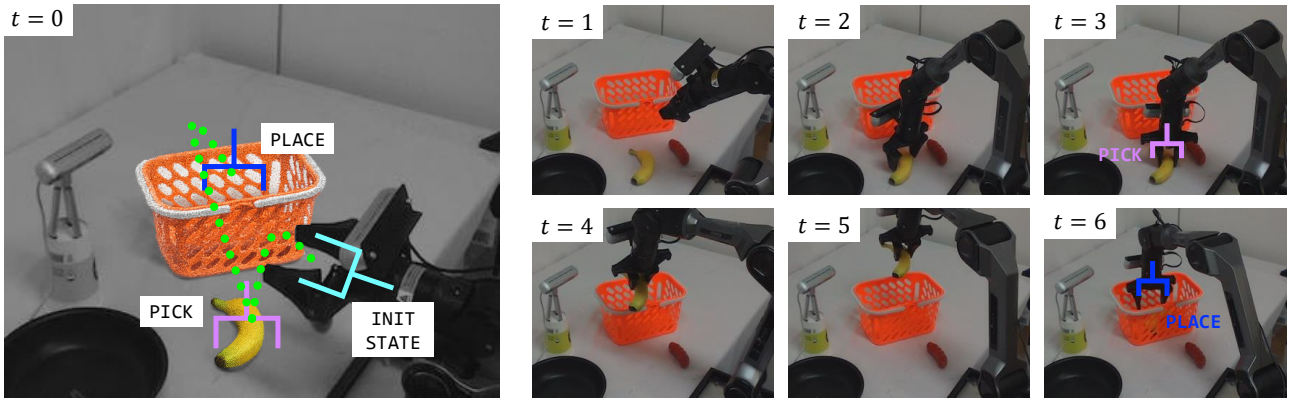


Fig. 4: **Grasp-pose-aware trajectory generation.** PaPA generates a pick-and-place trajectory by interpolating three key waypoints: the initial pose, pick pose, and place pose. **(Left)** The trajectory is constructed by sequential interpolation (initial → pick → place). **(Right)** The robot executes the trajectory using a low-level controller, closing the gripper at the pick pose and opening it at the place pose.

define key waypoints for the manipulation sequence. We build upon a PCA-based grasping method [34] to estimate the object's principal axes and orientation. Once the grasp poses are identified, a pick-and-place trajectory is constructed by connecting a sequence of key poses: the picking pose on the target object and the placing pose at the destination (see Fig. 4). These waypoints are joined via interpolation to form a continuous, robot-executable trajectory. To ensure reliability, the resulting motion is filtered for kinematic feasibility and smoothness, allowing the robot to transport the object seamlessly from the source to the target location.

E. Trajectory Validation

The generated trajectory is executed on the robot to perform the specified task. During execution, the robot records multi-view visual observations o_t and executed actions a_t at each timestep t, forming an episode candidate $\tilde{\tau}$. Successful executions are identified through human verification, such as lightweight success labeling, and are added to the dataset for subsequent imitation learning. Failed executions, typically caused by grasp slippage or collisions, are discarded.

IV. Experiments

We conduct a comprehensive set of experiments to evaluate PaPA. Specifically, our experiments are designed to (1) assess the quality of the collected data for training imitation learning policies, (2) demonstrate the efficiency of the proposed data acquisition framework, and (3) validate key design choices through ablation studies.

A. Experimental Setup

Hardware Setup. All experiments are conducted using an AgileX PiPER robotic arm, a 6-degree-of-freedom (6-DoF) manipulator equipped with a two-finger gripper. The 6-DoF manipulation capability provides sufficient motion flexibility to execute a wide range of trajectories within the workspace of the robot. The robotic arm is paired with a multi-view RGB perception system consisting of three cameras: one wrist-mounted camera and two fixed workspace cameras (Intel RealSense D455 and L515). Camera extrinsics are calibrated

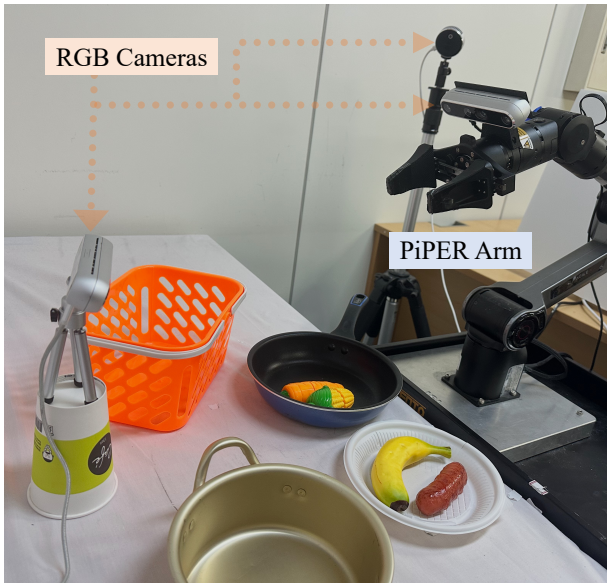


Fig. 5: **Hardware setup.** Our setup consists of an AgileX PiPER robotic manipulator and a multi-view RGB perception system with three cameras: a wrist-mounted camera, a front-facing camera, and an overhead table-view camera. A variety of everyday objects are placed within the workspace for pick-and-place data collection.

prior to data collection using AprilTag-based methods [53] and standard hand–eye calibration [63]. This aligns visual observations with robot kinematics, enabling accurate 3D reconstruction.

Baselines. We evaluate PaPA in comparison to three representative robotic data collection pipelines: (i) leader–follower teleoperation, (ii) language-based teleoperation [37], and (iii) VoxPoser [32], a semantically grounded trajectory generation framework. For the leader–follower teleoperation baseline, hereafter referred to as human teleoperation, a human expert manipulates a leader arm with an identical embodiment to the robot, which directly commands the follower arm in real time. This setup enables intuitive control and accurate trajectory transfer by leveraging identical hardware for both the leader and follower units, providing high-quality expert demonstrations. Language-based teleoperation (CLIP-RT) serves as a

TABLE I: **Task-wise behavior cloning performance.** Success counts over 10 evaluation rollouts for imitation learning policies trained on datasets collected by different methods. Higher values indicate better task execution, and bold entries denote the best performance.

Baseline	Behavior Cloning Task Success			
	Task 1	Task 2	Task 3	Task 4
Human Teleoperation	**10 / 10**	**9 / 10**	9 / 10	**10 / 10**
CLIP-RT	0 / 10	0 / 10	0 / 10	0 / 10
VoxPoser	3 / 10	1 / 10	3 / 10	1 / 10
PaPA (ours)	**10 / 10**	**9 / 10**	**10 / 10**	9 / 10

human-in-the-loop baseline that converts iterative high-level language instructions into executable low-level actions (e.g., move forward a bit: [0.05, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]). Lastly, VoxPoser represents a more autonomous trajectory generation approach, utilizing LLMs to produce 3D value maps over voxelized scene representations, which are then used to guide trajectory optimization for autonomous execution.

Tasks & Data Collection. We construct a diverse set of pick-and-place manipulation tasks to evaluate the data collection performance of each baseline. We design the task instruction to follow a pick-and-place structure (i.e., 'Pick A and place it into B'), where A $\in$ {banana, sausage, corn, carrot} and B $\in$ {plate, basket, pan, pot}. For each task, we collect 20 successful demonstrations for training imitation learning policies. We instantiate four tasks using the following object–container pairs: Task 1 (banana, plate), Task 2 (sausage, basket), Task 3 (corn, pan), and Task 4 (carrot, pot).

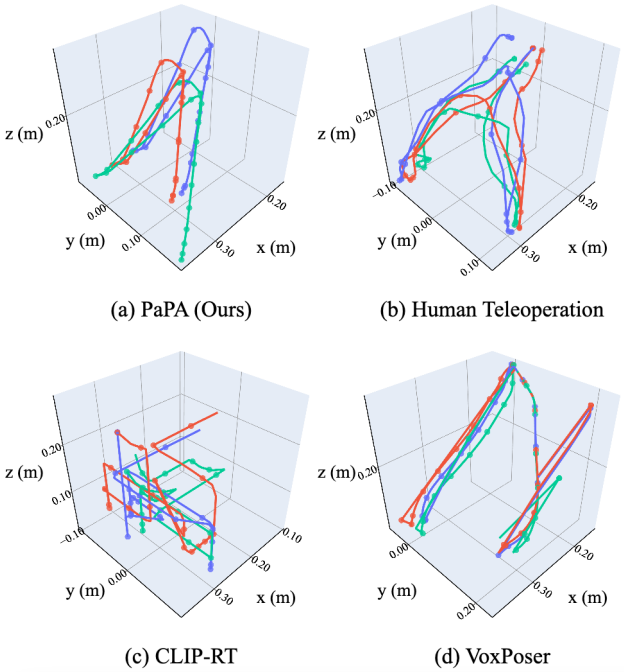
Evaluation. We evaluate PaPA and the baselines using four complementary metrics that capture task success, data quality, collection efficiency, and human effort. (1) *Behavior Cloning Task Success* counts successful task completions during behavior cloning policy rollouts. We train Diffusion Policy [14] on each dataset and evaluate the policy over 10 rollouts per task. (2) *Trajectory quality* is assessed along two axes: (i) trajectory similarity, measured by Fréchet distance [1] and Dynamic Time Warping (DTW) [3] over end-effector Cartesian trajectories (x, y, z), and (ii) smoothness and kinematic cost, quantified by jerkiness. (3) *Collection efficiency* evaluates episode collection rate and data yield. Episode collection rate measures the number of successful episodes collected, while data yield denotes the fraction of successful demonstrations among all attempted demonstrations. (4) *Human effort* is evaluated by measuring the expert workload. To capture the objective workload, we measure physical effort and cognitive effort. We also report a subjective workload by conducting surveys on five teleoperation experts.

B. Data Quality

Behavior Cloning Performance. Policies trained on trajectories collected by PaPA achieve behavior cloning performance comparable to those trained on expert human teleoperation data. As shown in Table I, PaPA-trained policies attain an average task success of 9.75/10, matching the performance of policies trained on human teleoperation demonstrations

TABLE II: **Trajectory Quality Comparison.** Fréchet distance and DTW quantify similarity to human demonstrations while jerkiness measures motion smoothness. Lower values indicate better performance across all metrics, and bold entries denote the best results.

Baseline	Fréchet distance (cm)	DTW (cm)	Jerkiness (cm/s^3)
Human Teleoperation	–	–	38.0 $\pm$ 3.7
CLIP-RT	14.6 $\pm$ 3.0	4.5 $\pm$ 0.8	90.7 $\pm$ 9.1
VoxPoser	20.9 $\pm$ 0.7	5.4 $\pm$ 0.1	76.0 $\pm$ 1.2
PaPA (ours)	**9.2 $\pm$ 1.1**	**2.9 $\pm$ 0.4**	**9.7 $\pm$ 0.6**



(a) PaPA (Ours) (b) Human Teleoperation

(c) CLIP-RT (d) VoxPoser

Fig. 6: **Visualization of Sampled Trajectories**. The figure shows three sampled trajectories from dataset collected for Task 1 by (a) PaPA (Ours), (b) Human Teleoperation, (c) CLIP-RT, and (d) VoxPoser.

(9.75/10). In contrast, policies trained on datasets collected using CLIP-RT (0.0/10) and VoxPoser (2.0/10) perform substantially worse, indicating insufficient demonstration quality for effective imitation learning. Given that imitation learning fundamentally aims to reproduce human-like trajectories, data collected via human teleoperation serve as a strong reference for high-quality behavior cloning. Notably, the comparable performance achieved by PaPA demonstrates that it can serve as a practical and reliable data collection framework for imitation learning.

Trajectory Quality. Table II reports trajectory quality across data collection methods using Fréchet distance, DTW, and jerkiness. For Fréchet distance and DTW, each trajectory is compared against the set of trajectories collected via human teleoperation, which serves as a reference for expert behavior. For each trajectory, we compute the minimum distance to the closest human demonstration and report the mean and standard deviation of these minimum distances across the dataset. PaPA achieves the lowest Fréchet distance (9.2) and DTW (2.9), compared to CLIP-RT (14.6, 4.5) and VoxPoser (20.9, 5.4). In terms of smoothness, PaPA also records the lowest jerkiness

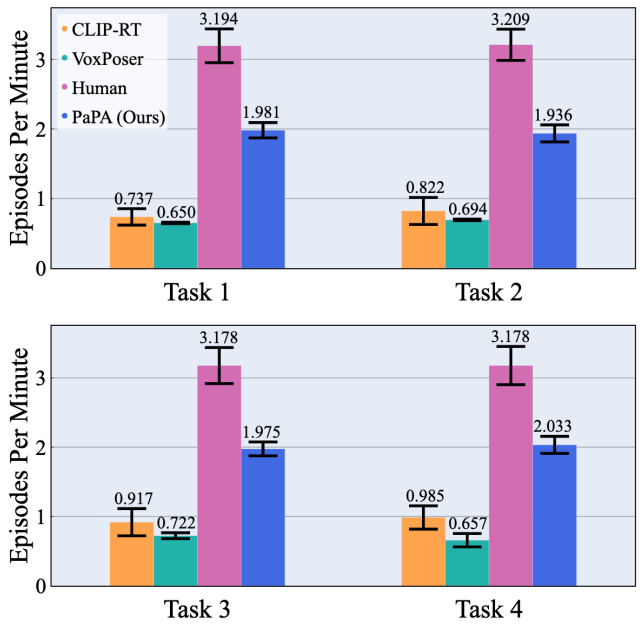


Fig. 7: **Data Collection Efficiency.** Average number of episodes collected per minute across four tasks for each method. Higher values indicate more efficient data collection.

(9.7), while human teleoperation yields 38.0, and CLIP-RT (90.7) and VoxPoser (76.0) show substantially higher values.

Qualitative Analysis. For a qualitative analysis, we sample three trajectories from each dataset of the same task and visualize them in Fig. 6. Compared to smooth and diverse patterns shown in human teleoperation data, CLIP-RT demonstrations are constrained to stiff and piecewise motions due to the restricted action space imposed by its language-based interface. In addition, trajectories generated by VoxPoser show limited diversity and are often straight-line, owing to its greedy optimization. We also observe that low-level subtasks decomposed by VoxPoser heavily rely on context prompts and lack kinematic constraints, often demonstrating unfavorable sudden movements. By contrast, PaPA generates trajectories with diverse approach behaviors toward the target, while maintaining clear convergence to precise grasp and placement poses. These properties stem from robust target localization and interpolation between key waypoints.

The behavior cloning results and trajectory quality analyses provide a coherent picture of data quality across methods. While behavior cloning performance measures downstream utility of the collected data, trajectory quality metrics explain the underlying causes of these results. This explains why PaPA, through its generation of high-quality human-like trajectories, can serve as a reliable data source for imitation learning, achieving performance comparable to human teleoperation. In contrast, CLIP-RT and VoxPoser, despite being capable of executing the tasks, do not provide demonstrations suitable for effective robotic data collection.

C. Efficiency in Data Collection

Data Collection Rate. We evaluate data collection rate by measuring the number of episodes recorded per minute. Fig. 7 shows that PaPA consistently achieves 1.9–2.1 episodes per minute across all tasks. In comparison, human teleoperation records episodes at a higher rate of 3.1–3.3 episodes per minute. Conversely, CLIP-RT and VoxPoser exhibit substantially lower throughput, achieving only 0.7–0.9 and 0.6–0.7 episodes per minute, respectively. The lower episode collection throughput of CLIP-RT and VoxPoser can be attributed to the structure of their data-generation pipelines. CLIP-RT requires a human operator to iteratively provide language instruction for incremental motion control. These iterative language inputs for fine-grained motion introduce significant interaction overhead. Meanwhile, VoxPoser necessitates repeated perception and planning cycles before execution, slowing down a single rollout. In contrast, although slower than human teleoperation, PaPA achieves efficient and stable data collection by restricting vision-language reasoning to a single perception step and generating complete trajectories via lightweight geometric planning.

TABLE III: **Data Yield Evaluation.** Data yield measured as the ratio of successful demonstrations to total trials when collecting 20 successes per task. Higher values indicate more reliable data collection with fewer failed attempts, and bold entries denote the best results.

Baseline	Data Yield Rate (%)			
	Task 1	Task 2	Task 3	Task 4
VoxPoser	74.1	**80.0**	50.0	54.1
PaPA (ours)	**90.9**	**80.0**	**90.9**	**83.3**

Data Yield. Data collection efficiency is also influenced by data yield, defined as the fraction of successful demonstrations among all attempted trials. To evaluate data yield, we measure the percentage of successful rollouts over all attempts, as reported in Table III. We compare PaPA and VoxPoser, since human teleoperation and CLIP-RT rely on human experts and rarely fail during execution. PaPA achieves a data yield of 90.9% on Task 1, indicating that approximately 91 out of 100 attempted rollouts result in valid demonstrations. In contrast, VoxPoser attains a lower yield of 74.1%, requiring more failed attempts to collect the same number of successful trajectories. Across all evaluated tasks, PaPA consistently achieves higher data yield than VoxPoser, with gains ranging up to 40 percentage points. These results indicate that PaPA provides a more efficient and reliable trajectory generation pipeline for autonomous data collection.

Human Effort. We analyze human effort, as data collection efficiency is closely tied to expert workload. We consider both physical and cognitive effort during data collection. For physical effort, we quantify the amount of direct human interaction required by each method (e.g., leader arm control or keyboard inputs). Human teleoperation requires continuous control of the leader arm throughout each episode, demanding full-time expert involvement and thus incurring the highest physical effort. For CLIP-RT, we measure the number of iterative language instructions issued per episode, reporting an

TABLE IV: **Ablation Studies.** We ablate key design components of PaPA, including global guidance and human supervision. Both ablations demonstrate their importance across multiple evaluation metrics.

Ablations	**Task Success**	**Fréchet distance** (cm)	**DTW** (cm)
w/o global guidance	3 / 10	10.5 ± 2.0	3.2 ± 0.2
w/o human supervision	8 / 10	9.6 ± 0.5	3.1 ± 0.1
PaPA (ours)	**10 / 10**	$\mathbf{9.2 \pm 1.1}$	$\mathbf{2.9 \pm 0.4}$

average of 36.61 instructions across tasks. In contrast, PaPA and VoxPoser require only lightweight human involvement in the form of post-hoc success verification, averaging 1.16 and 1.61 interventions per episode, respectively. To assess cognitive effort, we adopt a dual-task paradigm based on the Paced Auditory Serial Addition Test (PASAT) [25], which evaluates mental load through an auditory task without interfering with visual or motor control. PaPA exhibits the lowest cognitive load, achieving 0.33 correct responses per second, followed by VoxPoser (0.30), human teleoperation (0.28), and CLIP-RT (0.17). We further report perceived workload using NASA-TLX ratings [27], following the protocol of [18], which show a consistent trend: PaPA imposes the lowest mental workload, followed by VoxPoser, human teleoperation, and CLIP-RT. Further details are shown in Appendix.

The results on collection rate, data yield, and human effort highlight clear differences in data collection efficiency across methods. Although human teleoperation achieves the highest raw throughput, it relies on continuous expert involvement, limiting scalability. In contrast, CLIP-RT and VoxPoser exhibit substantially lower efficiency due to iterative human inputs or repeated perception and planning overhead, which reduce both collection rate and data yield. PaPA, on the other hand, achieves a favorable balance by generating complete trajectories from a single perception step, resulting in stable collection rates, consistently high data yield, and minimal human effort. These characteristics suggest that PaPA provides a more scalable and practical data collection pipeline for real-world robotic manipulation.

D. Ablation Studies

Effect of Global Guidance. Removing global guidance from the vision-language reasoning stage leads to a pronounced degradation in imitation learning performance (Table IV). Without global guidance, trajectories are generated solely by interpolating key waypoints, lacking consistent approach directions and motion diversity. As a result, behavior cloning success drops from 10/10 to 3/10 compared to PaPA. Trajectory similarity metrics also degrade, with Fréchet distance increasing from 9.2 to 10.5 and DTW from 2.9 to 3.2. These results indicate that global guidance is not merely a heuristic for task completion, but a critical component for inducing human-like and diverse motion patterns. Although successful demonstrations can still be collected, the reduced trajectory diversity significantly harms generalization in imitation learning.

Zero Human Supervision. We study an ablation without human supervision, where all rollouts are collected regardless of success. In this setting, we collect 20 episodes and train the policy on a dataset containing 18 successful and 2 failed demonstrations. As shown in Table IV, task success decreases to 8/10, and trajectory similarity also degrades, with Fréchet distance increasing from 9.2 to 9.6 and DTW from 2.9 to 3.1. Although the drop in task success is modest, removing human supervision causes noticeable qualitative degradation, such as oscillatory and corrective motions. This highlights that minimal post-hoc human verification is crucial for maintaining data quality in imitation learning.

V. Discussion, Limitations and Future Work

A. Discussion

We presented PaPA: Pick-and-Place Anything, a scalable robotic data collection framework that reduces the cost of teleoperation-based data collection. By decomposing the data collection paradigm into autonomous trajectory generation and lightweight human supervision, PaPA enables efficient real-world robotic data collection without continuous expert interaction, while maintaining data quality suitable for imitation learning. Through extensive experiments, we find that world knowledge encoded in pretrained vision-language models is sufficient for robust perception in open-world pick-and-place tasks. Moreover, we observe that smooth trajectories and motion diversity are critical factors for constructing high-quality datasets, and that lightweight yet sufficient human supervision (i.e., post-hoc verification) further improves data reliability. Additionally, we investigate how PaPA can extend as a fully automated robotic data collection framework (see Appendix).

B. Limitations and Future Work

While PaPA performs well on pick-and-place tasks that constitute a large portion of everyday manipulation, it remains challenging to extend the framework to more complex and dexterous tasks, such as cloth folding. This limitation highlights the need to extend PaPA to a tighter integration of grounded scene perception and planning for sophisticated dexterous manipulation. Future work could augment the existing PaPA framework with richer grasp and contact modeling, and task-aware motion primitives, enabling the autonomous collection of data for fine-grained manipulation tasks. Addressing these challenges would expand the task coverage of PaPA beyond pick-and-place and move toward scalable high-quality data collection for more dexterous real-world manipulation.

VI. Conclusion

We introduced PaPA, a scalable robotic data collection framework that leverages vision-language world knowledge to generate high-quality manipulation trajectories without continuous teleoperation. PaPA matches expert teleoperation in imitation learning performance of 97.5% while achieving efficient data collection at approximately two episodes per minute. Overall, PaPA offers a practical pathway toward scalable real-world robotic data collection and serves as a strong foundation for autonomous robot learning in open-world settings.

References

[1] Helmut Alt and Michael Godau. Computing the fréchet distance between two polygonal curves. *International Journal of Computational Geometry & Applications*, 5 (01n02):75–91.

[2] Suneel Belkhale, Tianli Ding, Ted Xiao, Pierre Sermanet, Quon Vuong, Jonathan Tompson, Yevgen Chebotar, Debidatta Dwibedi, and Dorsa Sadigh. Rt-h: Action hierarchies using language, 2024. URL https://arxiv.org/abs/2403.01823.

[3] Donald J. Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, AAAIWS'94, page 359–370. AAAI Press, 1994.

[4] Homanga Bharadhwaj, Jay Vakil, Mohit Sharma, Abhinav Gupta, Shubham Tulsiani, and Vikash Kumar. Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking, 2023.

[5] 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.

[6] 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. *arXiv preprint arXiv:2410.24164*, 2024.

[7] Rishi Bommasani. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[8] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[9] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alex Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *arXiv preprint arXiv:2307.15818*, 2023.

[10] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

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

[12] 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.

[13] Xuxin Cheng, Jialong Li, Shiqi Yang, Ge Yang, and Xiaolong Wang. Open-television: Teleoperation with immersive active visual feedback. *arXiv preprint arXiv:2407.01512*, 2024.

[14] 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*, 2024.

[15] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024.

[16] Henry M Clever, Ankur Handa, Hammad Mazhar, Kevin Parker, Omer Shapira, Qian Wan, Yashraj Narang, Iretiayo Akinola, Maya Cakmak, and Dieter Fox. Assistive tele-op: Leveraging transformers to collect robotic task demonstrations. *arXiv preprint arXiv:2112.05129*, 2021.

[17] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning. *arXiv preprint arXiv:1910.11215*, 2019.

[18] Shivin Dass, Karl Pertsch, Hejia Zhang, Youngwoon Lee, Joseph J Lim, and Stefanos Nikolaidis. Pato: Policy assisted teleoperation for scalable robot data collection. *arXiv preprint arXiv:2212.04708*, 2022.

[19] 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.

[20] Haritheja Etukuru, Norihito Naka, Zijin Hu, Seungjae Lee, Julian Mehu, Aaron Edsinger, Chris Paxton, Soumith Chintala, Lerrel Pinto, and Nur Muham-

mad Mahi Shafiullah. Robot utility models: General policies for zero-shot deployment in new environments. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8275–8283. IEEE, 2025.

[21] Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 653–660. IEEE, 2024.

[22] Hongjie Fang, Chenxi Wang, Yiming Wang, Jingjing Chen, Shangning Xia, Jun Lv, Zihao He, Xiyan Yi, Yunhan Guo, Xinyu Zhan, Lixin Yang, Weiming Wang, Cewu Lu, and Hao-Shu Fang. Airexo-2: Scaling up generalizable robotic imitation learning with low-cost exoskeletons. *arXiv preprint arXiv:03081*, 2025.

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

[24] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, Steven Bohez, Konstantinos Bousmalis, Anthony Brohan, Thomas Buschmann, Arunkumar Byravan, Serkan Cabi, Ken Caluwaerts, Federico Casarini, Oscar Chang, Jose Enrique Chen, Xi Chen, Hao-Tien Lewis Chiang, Krzysztof Choromanski, David D'Ambrosio, Sudeep Dasari, Todor Davchev, Coline Devin, Norman Di Palo, Tianli Ding, Adil Dostmohamed, Danny Driess, Yilun Du, Debidatta Dwibedi, Michael Elabd, Claudio Fantacci, Cody Fong, Erik Frey, Chuyuan Fu, Marissa Giustina, Keerthana Gopalakrishnan, Laura Graesser, Leonard Hasenclever, Nicolas Heess, Brandon Hernaez, Alexander Herzog, R. Alex Hofer, Jan Humplik, Atil Iscen, Mithun George Jacob, Deepali Jain, Ryan Julian, Dmitry Kalashnikov, M. Emre Karagozler, Stefani Karp, Chase Kew, Jerad Kirkland, Sean Kirmani, Yuheng Kuang, Thomas Lampe, Antoine Laurens, Isabel Leal, Alex X. Lee, Tsang-Wei Edward Lee, Jacky Liang, Yixin Lin, Sharath Maddineni, Anirudha Majumdar, Assaf Hurwitz Michaely, Robert Moreno, Michael Neunert, Francesco Nori, Carolina Parada, Emilio Parisotto, Peter Pastor, Acorn Pooley, Kanishka Rao, Krista Reymann, Dorsa Sadigh, Stefano Saliceti, Pannag Sanketi, Pierre Sermanet, Dhruv Shah, Mohit Sharma, Kathryn Shea, Charles Shu, Vikas Sindhwani, Sumeet Singh, Radu Soricut, Jost Tobias Springenberg, Rachel Sterneck, Razvan Surdulescu, Jie Tan, Jonathan Tompson, Vincent Vanhoucke, Jake Varley, Grace Vesom, Giulia Vezzani, Oriol Vinyals, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Fei Xia, Ted Xiao, Annie Xie, Jinyu Xie, Peng Xu, Sichun Xu, Ying Xu, Zhuo Xu, Yuxiang Yang, Rui Yao, Sergey Yaroshenko, Wenhao Yu, Wentao Yuan, Jingwei Zhang, Tingnan Zhang, Allan Zhou, and Yuxiang Zhou. Gemini robotics: Bringing ai into the physical world, 2025. URL https://arxiv.org/abs/2503.20020.

[25] D M Gronwall. Paced auditory serial-addition task: a measure of recovery from concussion. *Percept Mot Skills*, 44(2):367–373, April 1977.

[26] Huy Ha, Pete Florence, and Shuran Song. Scaling up and distilling down: Language-guided robot skill acquisition, 2023. URL https://arxiv.org/abs/2307.14535.

[27] Sandra G. Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In Peter A. Hancock and Najmedin Meshkati, editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland, 1988. doi: https://doi.org/10.1016/S0166-4115(08)62386-9. URL https://www.sciencedirect.com/science/article/pii/S0166411508623869.

[28] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[29] Yingdong Hu, Fanqi Lin, Tong Zhang, Li Yi, and Yang Gao. Look before you leap: Unveiling the power of gpt-4v in robotic vision-language planning, 2023. URL https://arxiv.org/abs/2311.17842.

[30] Haoxu Huang, Fanqi Lin, Yingdong Hu, Shengjie Wang, and Yang Gao. Copa: General robotic manipulation through spatial constraints of parts with foundation models, 2024. URL https://arxiv.org/abs/2403.08248.

[31] Siyuan Huang, Zhengkai Jiang, Hao Dong, Yu Qiao, Peng Gao, and Hongsheng Li. Instruct2act: Mapping multi-modality instructions to robotic actions with large language model, 2023. URL https://arxiv.org/abs/2305.11176.

[32] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023.

[33] Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation, 2024. URL https://arxiv.org/abs/2409.01652.

[34] Zhizhou Jia, Zhengchao Lai, Yuetao Li, Zhuo Yang, Qun Rao, and Shaohui Zhang. An efficient pca-based target pose estimation algorithm for low-precision point clouds. In *2023 IEEE 11th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, volume 11, pages 1352–1356, 2023. doi: 10.1109/ITAIC58329.2023.10408792.

[35] Yixiang Jin, Dingzhe Li, Yong A, Jun Shi, Peng Hao, Fuchun Sun, Jianwei Zhang, and Bin Fang. Robotgpt: Robot manipulation learning from chatgpt, 2023. URL https://arxiv.org/abs/2312.01421.

[36] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen,

Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on robot learning*, pages 651–673. PMLR, 2018.

[37] Gi Cheon Kang, Junghyun Kim, Kyuhwan Shim, Jun Ki Lee, and Byoung-Tak Zhang. Clip-rt: Learning language-conditioned robotic policies from natural language supervision. *arXiv preprint arXiv:2411.00508*, 2024.

[38] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[39] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, Vitor Guizilini, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Muhammad Zubair Irshad, Donovon Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. Droid: A large-scale in-the-wild robot manipulation dataset. 2024.

[40] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[41] A. Laurentini. The visual hull concept for silhouette-based image understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(2):150–162, 1994. doi: 10.1109/34.273735.

[42] Marion Lepert, Jiaying Fang, and Jeannette Bohg. Phantom: Training robots without robots using only human videos, 2025. URL https://arxiv.org/abs/2503.00779.

[43] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016.

[44] Sergey Levine, Peter Pastor, Alex Krizhevsky, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection, 2016. URL https://arxiv.org/abs/1603.02199.

[45] Yi Li, Yuquan Deng, Jesse Zhang, Joel Jang, Marius Memmel, Raymond Yu, Caelan Reed Garrett, Fabio Ramos, Dieter Fox, Anqi Li, Abhishek Gupta, and Ankit Goyal. Hamster: Hierarchical action models for open-world robot manipulation, 2025. URL https://arxiv.org/abs/2502.05485.

[46] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control, 2023. URL https://arxiv.org/abs/2209.07753.

[47] Zefu Lin, Rongxu Cui, Chen Hanning, Xiangyu Wang, Junjia Xu, Xiaojuan Jin, Chen Wenbo, Hui Zhou, Lue Fan, Wenling Li, and Zhaoxiang Zhang. Embodiedcoder: Parameterized embodied mobile manipulation via modern coding model, 2025. URL https://arxiv.org/abs/2510.06207.

[48] Huihan Liu, Alice Chen, Yuke Zhu, Adith Swaminathan, Andrey Kolobov, and Ching-An Cheng. Interactive robot learning from verbal correction, 2023.

[49] Corey Lynch and Pierre Sermanet. Language conditioned imitation learning over unstructured data. *arXiv preprint arXiv:2005.07648*, 2020.

[50] Ajay Mandlekar, Yuke Zhu, Animesh Garg, Jonathan Booher, Max Spero, Albert Tung, Julian Gao, John Emmons, Anchit Gupta, Emre Orbay, et al. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. In *Conference on Robot Learning*, pages 879–893. PMLR, 2018.

[51] NVIDIA, :, Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, Liang Feng, Francesco Ferroni, Rama Govindaraju, Jinwei Gu, Siddharth Gururani, Imad El Hanafi, Zekun Hao, Jacob Huffman, Jingyi Jin, Brendan Johnson, Rizwan Khan, George Kurian, Elena Lantz, Nayeon Lee, Zhaoshuo Li, Xuan Li, Maosheng Liao, Tsung-Yi Lin, Yen-Chen Lin, Ming-Yu Liu, Xiangyu Lu, Alice Luo, Andrew Mathau, Yun Ni, Lindsey Pavao, Wei Ping, David W. Romero, Misha Smelyanskiy, Shuran Song, Lyne Tchapmi, Andrew Z. Wang, Boxin Wang, Haoxiang Wang, Fangyin Wei, Jiashu Xu, Yao Xu, Dinghao Yang, Xiaodong Yang, Zhuolin Yang, Jingxu Zhang, Xiaohui Zeng, and Zhe Zhang. Cosmos-reason1: From physical common sense to embodied reasoning, 2025. URL https://arxiv.org/abs/2503.15558.

[52] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Charles Xu, Jianlan Luo, Tobias Kreiman, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

[53] Edwin Olson. Apriltag: A robust and flexible visual fiducial system. In *2011 IEEE International Conference on Robotics and Automation*, pages 3400–3407, 2011. doi: 10.1109/ICRA.2011.5979561.

[54] Open X-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Andrey Kolobov, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Felipe Vieira Frujeri, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guangwen Yang, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homanga Bharadhwaj, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jay Vakil, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi "Jim" Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Muhammad Zubair Irshad, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi, Patrick "Tree" Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Mart'in-Mart'in, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shubham Tulsiani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vikash Kumar, Vincent Vanhoucke, Vitor Guizilini, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiangyu Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yansong Pang, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yongqiang Dou, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, Zipeng Fu, and Zipeng Lin. Open X-Embodiment: Robotic learning datasets and RT-X models. https://arxiv.org/abs/2310.08864, 2023.

[55] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[56] Lerrel Pinto and Abhinav Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours, 2015. URL https://arxiv.org/abs/1509.06825.

[57] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[58] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint*

arXiv:2401.14159, 2024.

[59] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

[60] Nur Muhammad Mahi Shafiullah, Anant Rai, Haritheja Etukuru, Yiqian Liu, Ishan Misra, Soumith Chintala, and Lerrel Pinto. On bringing robots home. *arXiv preprint arXiv:2311.16098*, 2023.

[61] Kenneth Shaw, Shikhar Bahl, and Deepak Pathak. Videodex: Learning dexterity from internet videos, 2022. URL https://arxiv.org/abs/2212.04498.

[62] Grace Tang, Swetha Rajkumar, Yifei Zhou, Homer Rich Walke, Sergey Levine, and Kuan Fang. Kalie: Fine-tuning vision-language models for open-world manipulation without robot data, 2024. URL https://arxiv.org/abs/2409.14066.

[63] R.Y. Tsai and R.K. Lenz. A new technique for fully autonomous and efficient 3d robotics hand/eye calibration. *IEEE Transactions on Robotics and Automation*, 5 (3):345–358, 1989. doi: 10.1109/70.34770.

[64] Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning (CoRL)*, 2023.

[65] Beichen Wang, Juexiao Zhang, Shuwen Dong, Irving Fang, and Chen Feng. Vlm see, robot do: Human demo video to robot action plan via vision language model, 2025. URL https://arxiv.org/abs/2410.08792.

[66] Wensheng Wang and Ning Tan. Hybridgen: Vlm-guided hybrid planning for scalable data generation of imitation learning, 2025. URL https://arxiv.org/abs/2503.13171.

[67] Yufei Wang, Zhou Xian, Feng Chen, Tsun-Hsuan Wang, Yian Wang, Katerina Fragkiadaki, Zackory Erickson, David Held, and Chuang Gan. Robogen: Towards unleashing infinite data for automated robot learning via generative simulation, 2023.

[68] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[69] Jimmy Wu, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser. Tidybot: Personalized robot assistance with large language models. *Autonomous Robots*, 2023.

[70] Philipp Wu, Yide Shentu, Zhongke Yi, Xingyu Lin, and Pieter Abbeel. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators, 2023.

[71] Yue Xu, Litao Wei, Pengyu An, Qingyu Zhang, and Yong-Lu Li. exumi: Extensible robot teaching system with action-aware task-agnostic tactile representation. *arXiv preprint arXiv:2509.14688*, 2025.

[72] Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Liden, Kimin Lee, Jianfeng Gao, Luke Zettlemoyer, Dieter Fox, and Minjoon Seo. Latent action pretraining from videos, 2024. URL https://arxiv.org/abs/2410.11758.

[73] 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.

[74] Enshen Zhou, Qi Su, Cheng Chi, Zhizheng Zhang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, and He Wang. Code-as-monitor: Constraint-aware visual programming for reactive and proactive robotic failure detection, 2025. URL https://arxiv.org/abs/2412.04455.

[75] Yifeng Zhu, Abhishek Joshi, Peter Stone, and Yuke Zhu. Viola: Imitation learning for vision-based manipulation with object proposal priors. In *Conference on Robot Learning*, pages 1199–1210. PMLR, 2023.