

Teleoperation, Simulation, or Human Video? Data Utilization Law for Robot Manipulation

Chenhao Shi^{1*} Yichen Zhu^{2*} Junjie Wen¹ Yefei Chen¹ Ziang Liu¹ Faming Fang¹✉

¹East China Normal University ²University of Toronto

{51265901076, 51255901019, 51275901014, 51265901036}@stu.ecnu.edu.cn

fmfang@cs.ecnu.edu.cn yichen.zhu@foxmail.com

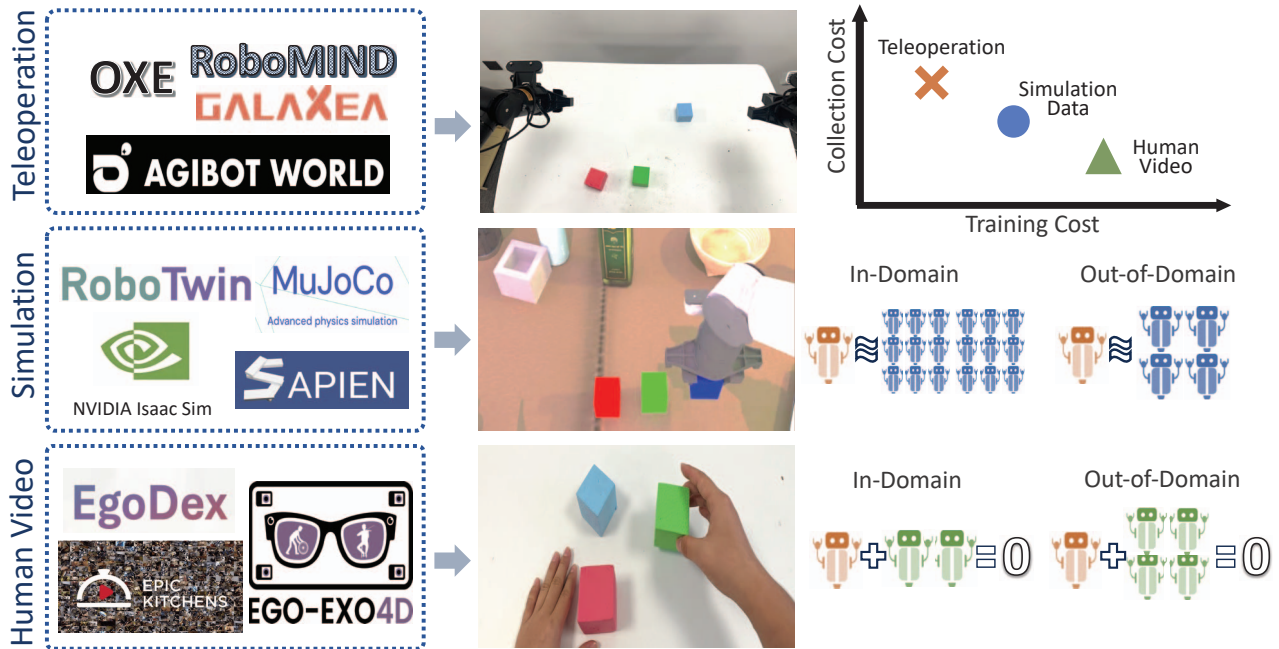


Figure 1. **Overview of robotic data sources and their quantified utilization.** We illustrate three data types—teleoperation, simulation, and human video—together with their conceptual trade-off between collection cost and training cost.

Abstract

Teleoperation, simulation, and human video represent the three primary data sources for robotic manipulation. Teleoperation data offers high quality at a high collection cost, whereas simulation and human video data are cheaper to acquire but introduce significant embodiment gaps. This trade-off has sparked a debate in the robotics community about the most effective data types for training robot policies. To address this, we introduce a data utilization law for robotic manipulation, drawing an analogy from economics to establish a formal “exchange rate” across data types. We quantify data utilization by using a single real-

world teleoperated demonstration as a base unit and then measuring the volume of other data (i.e., simulation or human video) required to achieve equivalent performance. Through a comprehensive investigation across three manipulation tasks—training Diffusion Policy and π_0 on over 10000 trajectories—we systematically analyze the interplay between teleoperation, simulated, and human video data. Our analysis reveals several key findings: 1) Simulation data generally improves model generalization, with an approximate exchange rate of 8 simulation samples providing the equivalent benefit of 1 teleoperated sample. 2) Human video tends to degrade in-domain model performance, where adding approximately 10 human video samples can negate the benefit of a single teleoperated data point. 3) Whether human video helps generalization or

*Equal Contribution.

✉Corresponding author.

simulation aids in-domain performance varies significantly across tasks. We believe that our work provides crucial insights into balancing the costs of data collection with the computational demands of model training.

“Our biggest cost is not power, or servers, or people. It’s lack of utilization. It dominates all other costs.”
— Jeff Bezos

1. Introduction

The robotics community has embraced data scaling as a path toward generalizable and robust robot policies, drawing parallels to the recent wave of success in large language models. This pursuit has primarily followed three directions: collecting targeted in-the-lab datasets via teleoperation, creating simulated environments for large-scale data generation, or leveraging unstructured in-the-wild human videos. While the community widely accepts the “more-is-better” paradigm, this heterogeneity of sources means we lack a principled understanding of the relative value of each data type. As illustrated by the data pyramid in Figure 1, roboticists must navigate a trade-off between data fidelity and scale. Teleoperation data, collected directly on a physical robot, offers the highest quality and is most aligned with the test distribution in terms of environment and embodiment. However, this data is expensive and labor-intensive to collect. Conversely, data from simulations and web-scale human videos are far easier to obtain but suffer from significant visual and embodiment gaps. This raises a crucial, yet underexplored, question regarding the true economic efficiency of each data source.

Typically, the cost of data is viewed solely through the lens of collection: $\text{Cost total} = \text{Cost of data collection}$. This perspective suggests that cheaper data is always better because it scales more easily. However, this view overlooks a critical factor: data utilization. Some data types are inherently less effective than others. For instance, if a task can be solved by a model trained on 500 teleoperated trajectories, achieving the same performance might require 5,000 human video trajectories. In this scenario, the tenfold increase in training cost could easily outweigh the savings in data collection. This necessitates a more comprehensive cost function: $\text{Cost total} = \text{Cost data collection} + \text{Cost model training}$. This revised model brings the concept of data utilization to the forefront. In economics, utilization measures how fully a resource is used. Applying this here, our core idea is to establish a quantitative “exchange rate” between different data types, thereby creating a formal method to measure their relative utility and true cost-effectiveness.

To this end, we introduce the data utilization law for robotic manipulation. We operate within the imitation learning paradigm—the dominant approach for real-world

robot skill acquisition. We investigate the trade-offs between two critical data pairs: simulation versus teleoperation data and human video versus teleoperation data. To ground our study, we curate a large-scale dataset for a suite of representative manipulation tasks—Pick Dual Bottles, Hand Over Block, and Rank RGB Blocks. This data set comprises real-world teleoperated data collected with an Agilex-2.0 system, RoboTwin 2.0 simulation data [7], and human video demonstrations processed via a multi-stage pipeline combining video stabilization, MediaPipe hand landmark extraction [20], and gripper state estimation. Leveraging state-of-the-art policy models like Diffusion Policy [8] and pre-trained vision-language-action model π_0 [4], we meticulously evaluate how varying the composition of training data impacts policy generalization across two key axes: visual background and object position. By analyzing the data utilization ratio, we derive empirical data utilization laws. Our extensive investigation, backed by over 10,000 demonstrations and more than 600 real-world rollouts, reveals that:

- Simulation data reliably improves out-of-domain generalization. Although weaker per sample than real teleoperation, its effect is quantifiable: on average, about 8 simulation trajectories match the generalization benefit of 1 teleoperated trajectory, making large-scale simulation a cost-effective source of diversity.
- Naively adding human video can degrade in-domain performance due to embodiment and viewpoint mismatch. We observe a negative exchange rate in which roughly 10 human-video trajectories can offset the benefit of a single teleoperated demonstration, underscoring the hidden cost of unaligned human demonstrations.
- While human video typically hurts in-domain learning, its impact on generalization ranges from mildly helpful to detrimental, and simulation does not always benefit in-domain tasks. These nuances indicate that data mixing must be tailored to the specific task and performance objectives.

Our work offers a novel framework and a new vocabulary for the systematic analysis of robot data, providing a principled guide for mixing different data types in real-world applications.

2. Related Work

Data scaling in robotic manipulation. Scaling laws in vision and NLP show that performance improves predictably with data, model size, and compute [6, 10, 13, 25]. Robotic manipulation faces distinct scaling challenges: real-world data is expensive and task-specific, and despite progress in simulation [1, 7, 18], teleoperation [23], and human video [9, 11, 15–17], datasets remain far smaller than in vision and language. Recent studies [12, 14, 21, 28, 30] have investigated robot data scaling laws and cross-

modality transfer, but most improve learning by modifying architectures with auxiliary objectives. In contrast, we fix the policy and study how data modality composition alone affects real-world performance.

Imitation learning in robot learning. Imitation learning (IL) provides a data-efficient paradigm for training robotic agents by leveraging expert demonstrations rather than trial-and-error exploration. Early IL methods, such as behavioral cloning (BC) [24] and inverse reinforcement learning (IRL) [22] enabled policy acquisition from demonstration trajectories, but often suffered from compounding errors and poor generalization. More recent approaches combine IL with large-scale pretraining and multimodal inputs, leading to vision-language-action (VLA) models [3–5, 26, 31, 32]. These models align perception, language, and action for language-conditioned task execution and transfer. We introduce a framework for systematic robot data analysis and principled data mixing in imitation learning.

3. Approach

3.1. Data sources

We use three data sources: teleoperation, simulation, and human videos. For each task, all data types follow similar execution procedures to maintain consistency in task definition and motion patterns.

Task description. We evaluate three dual-arm manipulation tasks of increasing difficulty: *Pick Dual Bottles* (easy), *Hand Over Block* (medium), and *Rank RGB Blocks* (hard). This progression enables analysis of whether the data utilization law holds across varying task complexities.

- **Pick Dual Bottles.** The robot uses its right arm to grasp a Sprite bottle, followed by using its left arm to pick up a Coca-Cola bottle. Subsequently, both arms lift the bottles simultaneously in a coordinated manner.
- **Hand Over Block.** It requires the robot’s left arm to grasp a red block, perform a hand-over to the right arm, and place the block into the designated top-right goal location.
- **Rank RGB Blocks.** It assesses multi-step rearrangement skills by sorting three colored blocks (Red, Green, Blue) from random permutations into the canonical R–G–B order.

Teleoperation. We collected a dataset of expert demonstrations, $\mathcal{D}_{\text{tele}}$, using two real-world robotic platforms: the Aloha-Agilex-2.0 and the Dual-ARX-R5. For each of the three tasks, we gathered 200 to 500 trajectories per platform, resulting in a total of approximately 1500 teleoperation demonstrations. This range of data volume was chosen because, at such amounts, the teleoperation data can already achieve a relatively high success rate, making it a meaning-

ful and practical baseline for our study. Both systems utilize a 14-DoF action space, representing the joint angles of their dual 7-DoF arms and binary gripper states.

Simulation. To investigate the role of large-scale, low-cost data, we generated a simulated dataset, $\mathcal{D}_{\text{sim}}$, using the RoboTwin2.0 environment (Figure 2, middle). We created trajectories for both the Aloha-Agilex-2.0 (1,000 trajectories per task for three tasks) and Dual-ARX-X5 (1,000 trajectories per task for two tasks) models, yielding 5,000 simulated demonstrations in total. A key feature of this dataset is extensive **background randomization**, where we systematically varied background textures.

Human video. We collected $\mathcal{D}_{\text{human}}$: 500 demonstrations per task (1,500 total), recorded via a chest-mounted camera by an operator performing each task bimanually in the same physical environment as the teleoperation. Videos are processed through a multi-stage pipeline (details in Appendix 8.3): optical flow stabilization, hand landmark extraction via MediaPipe [20], and end-effector pose computation, yielding a 14D bimanual action vector $[\Delta p_L, \Delta r_L, g_L, \Delta p_R, \Delta r_R, g_R]$ per frame. While this end-effector delta representation differs semantically from the robot’s joint-angle action space, pre-training on human video is designed to build manipulation-aware visual representations and motion priors in the policy backbone, with the policy re-adapting to joint-angle space during teleoperation fine-tuning.

3.2. Training Strategy

Policy models and configurations. Our primary policy network is the Diffusion Policy [8] equipped with a CNN-based U-Net architecture, where visual features are extracted using a ResNet-50 encoder pretrained on ImageNet. For instruction, we embed textual inputs via DistilBERT [27]. We also utilize the state-of-the-art vision-language-action model π_0 . To isolate the effects of the model architecture from the pre-training data, we initialized all models with the PaliGemma pre-trained weights [29]. This approach allows for a fair assessment of architecture-agnostic generalization.

Training setup. We compare two distinct training protocols. In the **teleoperated-from-scratch** setting, the policy is trained exclusively on $\mathcal{D}_{\text{tele}}$ for the entire training budget, i.e., $T_{\text{tele}} = T_{\text{total}}$. In the **non-teleoperated pre-train + teleoperated fine-tune** setting, the policy is first trained for T_{nt} steps on $\mathcal{D}_{\text{nt}}$ — comprising $\mathcal{D}_{\text{sim}}$ or $\mathcal{D}_{\text{human}}$ — and then fine-tuned for T_{tele} steps on $\mathcal{D}_{\text{tele}}$. Since neither simulation-only nor human-video-only training is sufficient to accomplish the tasks, we adopt this pre-train–fine-tune procedure to leverage the complementary strengths of both data sources. To ensure computational fairness, both protocols satisfy $T_{\text{nt}} + T_{\text{tele}} = T_{\text{total}}$, guaranteeing that any performance differences arise solely from the training data

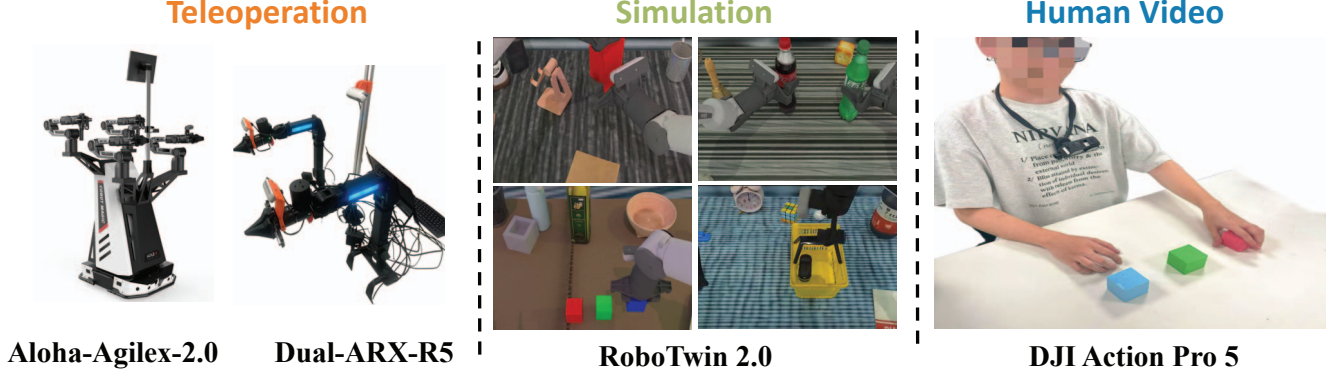


Figure 2. Overview of the three different datasets used in this study: **(left)** Teleoperation data collected via teleoperation on Aloha-Agilex-2.0 and Dual-ARX-R5; **(middle)** Simulation data generated in the RoboTwin2.0 environment with background randomization; **(right)** Human video data captured using a chest-mounted DJI Action Pro 5.

composition rather than unequal training effort.

3.3. Evaluation

We conduct all evaluations on a real-world Aloha-Agilex-2.0 platform. Each policy is evaluated for 10 trials per scenario. Performance is measured by the success rate ($\mathcal{S}$), ranging from 0 to 1, aggregating completion of critical sub-steps (e.g., reaching, grasping, placing).

We distinguish between **in-domain** and **out-of-domain** scenarios. *In-domain* refers to scenes with backgrounds and object positions seen during training, whereas *out-of-domain* includes changes to either the background or the object positions. As shown in Figure 3, we evaluate in three scenarios: **in-domain**, **unseen background**, and **unseen position**. The teleoperation-only policy is tested in all three scenarios. The simulation-pretrained policy is evaluated in the in-domain and unseen background settings, since the majority of simulation demonstrations are collected with randomized backgrounds, making background generalization the primary factor to assess. The human-video-pretrained policy is evaluated in the in-domain and unseen position settings, because human video is recorded in the same environments as teleoperation data (thus sharing the same backgrounds), but with more diverse and randomized object placements, making position generalization the relevant challenge.

3.4. Data Utilization Ratio Definition and Computation

We introduce the data utilization ratio, η , to quantify the value of non-teleoperated sources ($\mathcal{D}_{\text{sim}}$, $\mathcal{D}_{\text{human}}$) relative to $\mathcal{D}_{\text{tele}}$. A positive η means η non-teleoperated samples are equivalent to one teleoperated sample in utilization; a negative η means $|\eta|$ non-teleoperated samples hurt as much as one teleoperated sample.

Local utilization ratio. We define the **local utilization ratio** at a teleoperation data count N_{tele}^* where task performance is already at a practical level. This metric measures the relative marginal contribution of teleoperation data versus non-teleoperated data (simulation or human videos) near this point. Assuming a locally linear performance model,

$$\mathcal{S} = \gamma_{\text{tele}} N_{\text{tele}} + \gamma_{\text{nt}} N_{\text{nt}}, \quad (1)$$

where N_{tele} and N_{nt} are sample counts, and γ_{tele} and γ_{nt} denote marginal performance gains per sample.

To estimate γ_{tele} and γ_{nt} , we run two controlled experiments: (i) Training with N_{tele}^* teleoperation samples only; (ii) Pre-training with N_{nt} non-teleoperated samples, followed by fine-tuning with the same N_{tele}^* teleoperation samples. These two settings yield two equations of the form in Eq. 1, from which we solve for the ratio: $\eta_{\text{local}} = \gamma_{\text{tele}}/\gamma_{\text{nt}}$, interpreted as the number of non-teleoperated samples equivalent to one teleoperation sample in terms of local contribution.

Global utilization ratio. We define the **global utilization ratio** to assess whether non-teleoperated data provides positive or negative utilization as teleoperation data varies across its full range. Since the success rate $\mathcal{S}$ is bounded in $[0, 1]$, fitting a power law directly to $\mathcal{S}$ is mathematically problematic. Instead, we model the **optimality gap** $\mathcal{G} = 1 - \mathcal{S}$, which naturally follows a power-law decay:

$$\mathcal{G}_{\text{tele}}(N_{\text{tele}}) = \alpha N_{\text{tele}}^{-\beta}, \quad \alpha, \beta > 0. \quad (2)$$

The fit is performed in log-log space via linear regression on $(\log N_{\text{tele}}, \log \mathcal{G})$, and its quality is measured by the Pearson correlation coefficient r , where $|r|$ close to 1 indicates a strong power-law relationship. This fitted curve serves as the teleoperation-only baseline. For each mixed experiment with $(N'_{\text{tele}}, N'_{\text{nt}})$, we compare the observed optimality gap

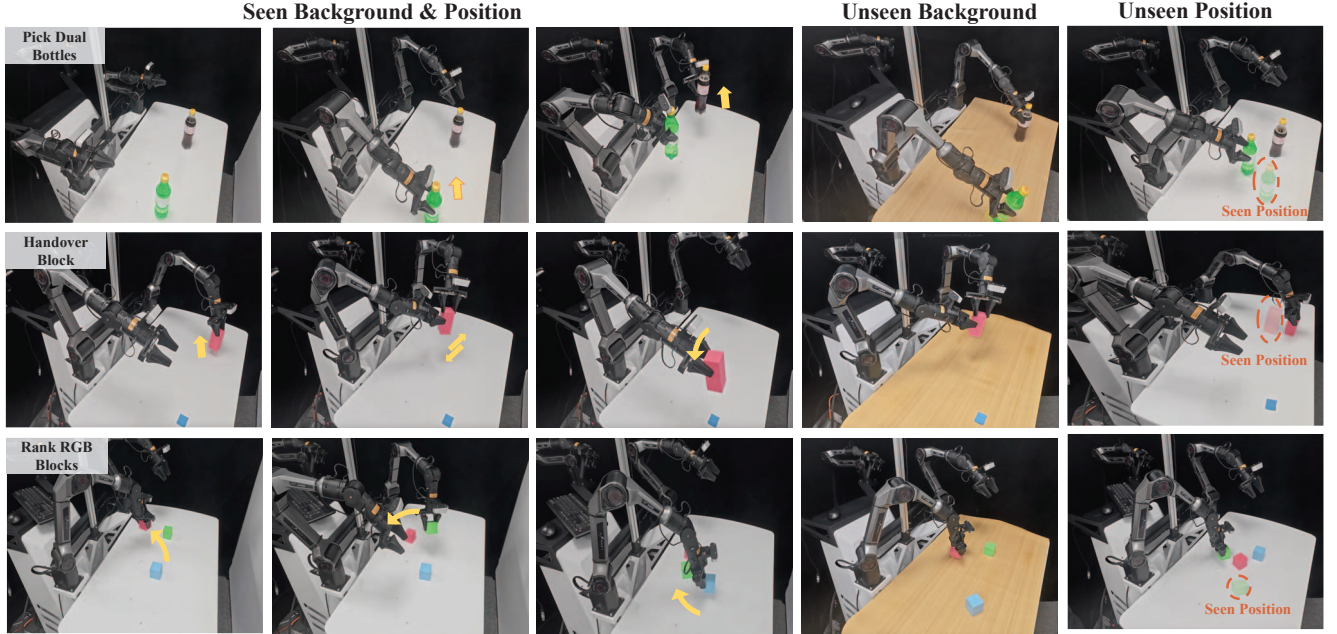


Figure 3. Our three real-world evaluation scenarios: **In-domain (left)**. The training set contains the same background and positions as the evaluation. **Unseen Background (middle)**. The background is new in the test set. **Unseen Position(right)** The object positions are not in the training set.

Task	Test Condition	$\mathcal{S}(\mathcal{D}_{\text{tele}})$	$\mathcal{S}(\mathcal{D}_{\text{sim}})$	η_{local}
Pick Dual Bottles	in-domain	0.70	0.55	-23
	unseen bg.	0.20	0.30	16
Hand Over Block	in-domain	0.70	0.60	-23
	unseen bg.	0.50	0.65	11
Rank RGB Blocks	in-domain	0.45	0.50	18
	unseen bg.	0.13	0.20	4

Table 1. Success rate $\mathcal{S}$ and local utilization ratio η_{local} for adding simulation data to teleoperation, evaluated under in-domain and unseen-background (unseen bg.) conditions. η_{local} measures the relative per-sample effectiveness between $\mathcal{D}_{\text{sim}}$ and $\mathcal{D}_{\text{tele}}$.

$\mathcal{G}_{\text{mix}} = 1 - \mathcal{S}_{\text{mix}}$ against the baseline prediction $\mathcal{G}_{\text{tele}}(N'_{\text{tele}})$:

$$\eta_{\text{global}} \begin{cases} > 0 & \text{if } \mathcal{G}_{\text{mix}} < \mathcal{G}_{\text{tele}}(N'_{\text{tele}}), \\ < 0 & \text{if } \mathcal{G}_{\text{mix}} > \mathcal{G}_{\text{tele}}(N'_{\text{tele}}). \end{cases} \quad (3)$$

When the mixed experiment achieves a *lower* optimality gap than the baseline, the non-teleoperated data contributes positively ($\eta_{\text{global}} > 0$); when the gap is *higher*, the additional data is detrimental ($\eta_{\text{global}} < 0$). This qualitative sign test complements the precise local estimate near N_{tele}^* by revealing whether the direction of utilization holds consistently across data scales.

4. Experimental Analysis of Data Utilization

In this section, we apply the methodology described in Section 3.1 to empirically measure the utilization ratios of simulation and human video data. For each source, we analyze both the local ratio (η_{local}) to understand its precise marginal value and the global ratio (η_{global}) to verify whether the sign of utilization (positive or negative) holds consistently across different data scales.

4.1. Experimental Setup

We trained baseline policies from scratch on teleoperation data ($\mathcal{D}_{\text{tele}}$) for each task using varying subset sizes: Pick Dual Bottles – 50, 100, 200 demonstrations; Hand Over Block – 50, 100, 200, 300; Rank RGB Blocks – 100, 300, 500. For computing the local utilization ratio, we set $N_{\text{tele}}^* = 200$ for Pick Dual Bottles, 300 for Hand Over Block, and 500 for Rank RGB Blocks. Non-teleoperated data was fixed at 1000 simulation trajectories ($\mathcal{D}_{\text{sim}}$) or 500 human video trajectories ($\mathcal{D}_{\text{human}}$). Policies were pre-trained on simulation or human data, then fine-tuned on the same teleoperation subsets, and evaluated under both in-domain (seen background/positions) and out-of-domain (unseen background/positions) conditions. Data utilization ratios were computed as in Section 3.4.

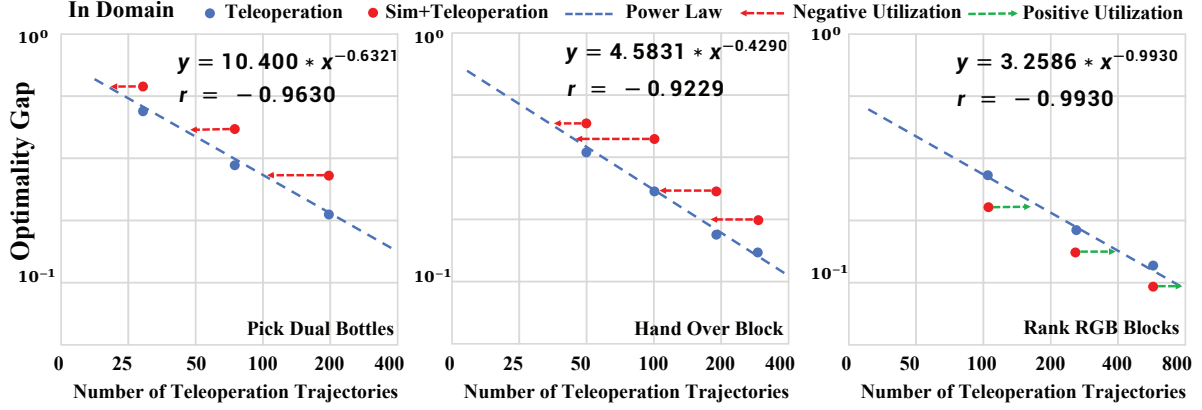


Figure 4. Global utilization analysis of simulation data in in-domain evaluation. The blue dashed line is the optimality-gap power-law baseline fitted to teleoperation-only data; r denotes the Pearson correlation coefficient of the log-log fit. Points *below* the baseline indicate positive utilization; points *above* indicate negative utilization. For simpler tasks (**left and middle**), adding simulation data is detrimental. For a difficult task (**right**), it provides a significant benefit.

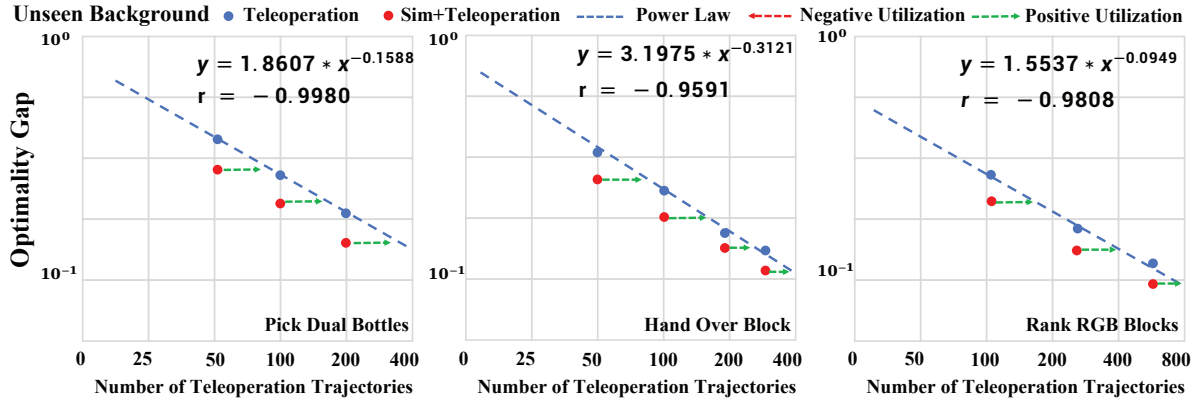


Figure 5. Global utilization analysis of simulation data under out-of-domain background. Simulation data shows consistently positive utilization across all tasks. Notably, its benefit increases with task difficulty.

Task	Test Condition	$\mathcal{S}(\mathcal{D}_{\text{tele}})$	$\mathcal{S}(\mathcal{D}_{\text{human}})$	η_{local}
Pick Dual Bottles	in-domain	0.70	0.45	-7
	unseen position	0.25	0.30	+12
Hand Over Block	in-domain	0.70	0.50	-6
	unseen position	0.25	0.30	+8
Rank RGB Blocks	in-domain	0.45	0.25	-2
	unseen position	0.17	0.13	-4

Table 2. Success rate $\mathcal{S}$ and local utilization ratio η_{local} when adding human video to teleoperation, evaluated under in-domain and unseen-position conditions for three tasks. η_{local} measures the relative effectiveness per sample between human video and teleoperation data.

4.2. Teleoperation vs. Simulation Utilization

We compare simulation and teleoperation data via local and global utilization ratios to evaluate how well simulation can

substitute teleoperation in improving task performance in both in-domain and unseen background tasks.

Local utilization ratio for simulation. As reported in Table 1, the local utilization of simulation data is highly task-dependent. For simpler in-domain tasks (*Pick Dual Bottles*, *Hand Over Block*), simulation exhibits a strong negative utilization ($\eta_{\text{local}} = -23$), indicating that the sim-to-real gap introduces detrimental artifacts that harm performance. Conversely, for the complex *Rank RGB Blocks* task, the ratio becomes positive ($\eta_{\text{local}} = 18$), suggesting that simulation provides valuable diversity for learning complex coordination. In the unseen background setting, simulation is consistently beneficial, yielding significant positive ratios across all tasks. This confirms that domain randomization in simulation is a highly effective strategy for improving visual robustness.

Global utilization ratio for simulation. Our global analysis confirms the sign of utilization observed locally and

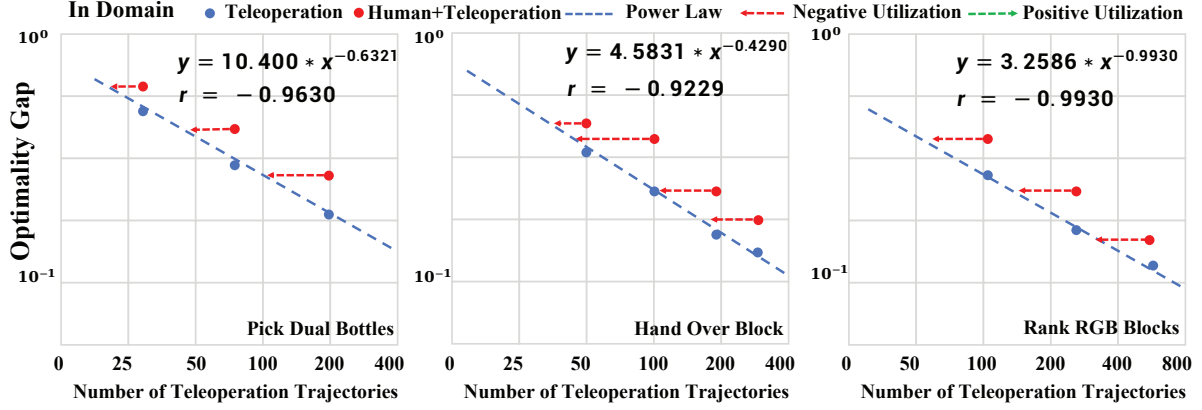


Figure 6. Global utilization analysis of human video data in in-domain evaluation. The blue dashed line is the optimality-gap baseline. All mixed-data points fall *above* the baseline, indicating consistently negative utilization across all tasks.

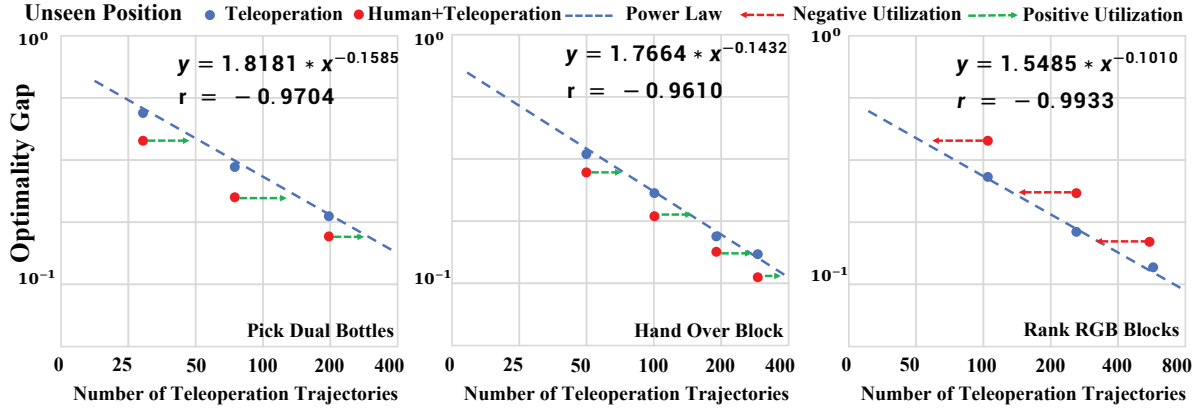


Figure 7. Global utilization analysis of human video data under unseen position. **(Left, middle)** Mixed-data points fall below the baseline, showing positive utilization for easy tasks. **(Right)** Points fall above the baseline, indicating negative utilization for the difficult task.

shows it holds across data scales. As shown in Figure 4, the blue dashed line represents the optimality-gap power-law baseline fitted to teleoperation-only data. For simple in-domain tasks, all mixed-data points fall *above* the baseline ($\eta_{\text{global}} < 0$), confirming that simulation data consistently hurts performance regardless of scale. In contrast, for the complex *Rank RGB Blocks* task, mixed-data points fall *below* the baseline ($\eta_{\text{global}} > 0$) across all scales, confirming that simulation data reliably improves performance. In the unseen background setting (Figure 5), all mixed-data points lie below the baseline for every task, indicating consistently positive utilization that grows with task difficulty.

4.3. Teleoperation vs. Human Video Utilization

In this section, we compare human video and teleoperation data via local and global utilization ratios, aiming to assess how well human video can substitute teleoperation in enhancing task performance for both in-domain and unseen position tasks.

Local utilization ratio for human video. Human video

data shows limited utilization and is often detrimental. In-domain, it consistently reduces performance; On simple tasks, seven human videos cause the same degradation as removing one teleoperated demonstration. Out-of-domain, it benefits simple tasks with a local utilization ratio of 10, but harms the complex task with a ratio of -4.

Global utilization ratio for human video. The global analysis confirms the sign of local utilization and reveals its consistency across scales. For in-domain tasks (Figure 6), all mixed-data points fall above the baseline ($\eta_{\text{global}} < 0$) across all data scales, and the gap widens with task complexity, indicating that the sub-optimal and behaviorally narrow nature of the data increasingly degrades policy performance for tasks demanding long-horizon reasoning. For simpler unseen position tasks (Figure 7), mixed-data points fall below the baseline ($\eta_{\text{global}} > 0$) at small scales, but the benefit diminishes as scale grows. Overall, these results show that human video can provide limited gains in specific cases, yet may introduce biases that reduce generalization in broader settings.

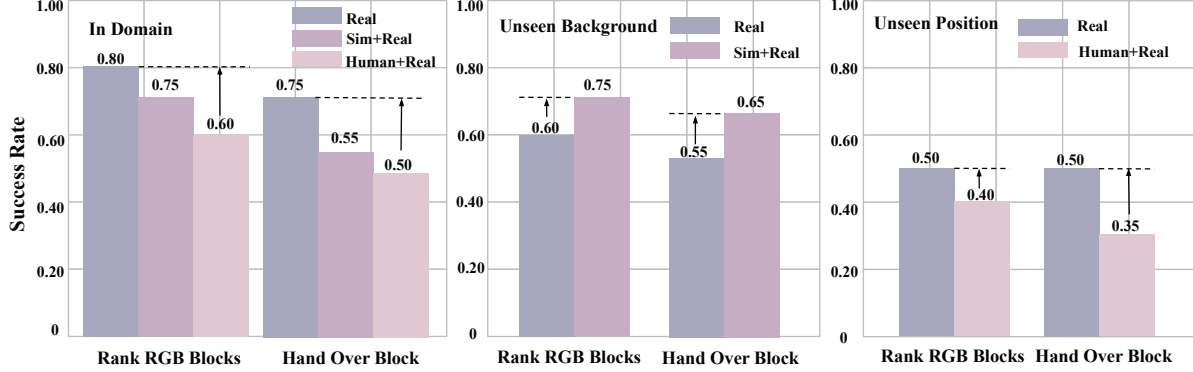


Figure 8. Verification of the data utilization law on π_0 policy and different real-world robot (ARX-R5). **(left)** Adding Simulation and Human Video in-domain both result in negative utilization. **(middle)** Adding Simulation in unseen backgrounds consistently improves performance. **(right)** Adding Human Video in unseen positions leads to negative utilization.

5. Verification Of Data Utilization Law

5.1. Data Utilization Law

To test the generality of the data utilization law beyond diffusion policy and our main platform, we evaluated a different architecture, π_0 , on the ARX-R5. Two tasks were used: *Hand Over Block* and *Rank RGB Blocks*, following the training protocol in Section 3.2.

As shown in Figure 8, in-domain experiments revealed that adding either simulation or human video data generally decreased performance, with the decline in *Rank RGB Blocks* likely due to π_0 already achieving high accuracy from teleoperation data alone. In the unseen background setting, simulation data consistently yielded positive utilization, indicating that π_0 benefits from well-aligned simulated experiences. In contrast, in the unseen position setting, human video data was uniformly detrimental, suggesting that π_0 is particularly sensitive to spatial domain mismatches introduced by human demonstrations.

5.2. Budget-Aware Collection

In the *Pick Dual Bottles* task, the local utilization ratio shows one teleoperation demo equals ~ 16 simulation demos. Since simulation produces 40 demos per hour vs. teleoperation’s 1, the speed ratio (40) exceeds the utilization ratio (16), making simulation more cost-effective in a mixed-data strategy.

We conducted a budget-constrained experiment on the Aloha-Agilex-2.0 platform, deploying a Diffusion Policy to compare two approaches over a four-hour period. The first approach, serving as the baseline, consisted solely of 200 teleoperated demonstrations collected within 4 hours. The second approach, our utilization-guided strategy, used a mix of 122 teleoperation demonstrations and 1951 simulation demonstrations, allocated using our cost-benefit ratio greedy method (Appendix 7.3). As shown in Table 3,

Real Traj.	Sim Traj.	Success Rate
200	0	0.20
122	1951	0.30

Table 3. Real-world performance on the *Pick Dual Bottles* task with unseen backgrounds, evaluated using the Diffusion Policy deployed on the Aloha-Agilex-2.0 platform. Comparison of different data collection strategies under a fixed 4-hour time budget.

our strategy achieved a success rate of 0.30, outperforming the baseline’s 0.20. This confirms that a utilization-aware data mixture yields superior real-world performance for the same data collection effort.

6. Conclusion

We revisit the “more is better” assumption in robot learning by jointly considering collection and training costs, introducing a data utilization law that quantifies exchange rates between data sources. Across representative dual-arm manipulation tasks, simulation data robustly supports generalization—about eight simulation trajectories match the benefit of one teleoperation trajectory—whereas naive mixing of human video can degrade in-domain performance, with roughly ten video trajectories offsetting the gain from a single high-quality teleoperation demonstration. These effects are goal-dependent: simulation generally improves out-of-domain transfer but yields mixed in-domain benefits, while human video’s value hinges on proper alignment and robustness goals. Our framework reframes data mixing as an economic design choice, encouraging a teleoperation core, scalable simulation augmentation, and selective use of human video. As simulation platforms expand to support richer task families, future work will extend this analysis to deformable-object manipulation as well as broader, high-dimensional, and long-horizon challenges typical of advanced robot manipulation.

References

- [1] Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, et al. Learning dexterous in-hand manipulation. In *The International Journal of Robotics Research (IJRR)*, 2020. 2
- [2] Hongzhe Bi, Lingxuan Wu, Tianwei Lin, Hengkai Tan, Zhizhong Su, Hang Su, and Jun Zhu. H-rdt: Human manipulation enhanced bimanual robotic manipulation. *arXiv preprint arXiv:2507.23523*, 2025. 3
- [3] 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. 3
- [4] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. $\pi 0$: A vision-language-action flow model for general robot control. corr, abs/2410.24164, 2024. doi: 10.48550. *arXiv preprint ARXIV.2410.24164*. 2
- [5] Anthony Brohan, Noah Brown, Yevgen Chebotar, Krzysztof Choromanski, and et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022. 3
- [6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1877–1901, 2020. 2
- [7] Tianxing Chen, Zhanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Qiwei Liang, Zixuan Li, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025. 2
- [8] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023. 2, 3
- [9] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022. 2
- [10] 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. 2
- [11] Ryan Hoque, Peide Huang, David J Yoon, Mouli Sivapuru, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025. 2
- [12] Yingdong Hu, Fanqi Lin, Pingyue Sheng, Chuan Wen, Jiacheng You, and Yang Gao. Data scaling laws in imitation learning for robotic manipulation. *arXiv preprint arXiv:2410.18647*, 2024. 2
- [13] 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. 2
- [14] Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13226–13233. IEEE, 2025. 2
- [15] Marion Lepert, Jiaying Fang, and Jeannette Bohg. Masquerade: Learning from in-the-wild human videos using data-editing. *arXiv preprint arXiv:2508.09976*, 2025. 2
- [16] Haoyun Li, Ivan Zhang, Runqi Ouyang, Xiaofeng Wang, Zheng Zhu, Zhiqin Yang, Zhentao Zhang, Boyuan Wang, Chaojun Ni, Wenkang Qin, et al. Mimicdreamer: Aligning human and robot demonstrations for scalable vla training. *arXiv preprint arXiv:2509.22199*, 2025.
- [17] Qixiu Li, Yu Deng, Yaobo Liang, Lin Luo, Lei Zhou, Chengtang Yao, Lingqi Zeng, Zhiyuan Feng, Huizhi Liang, Sicheng Xu, et al. Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. *arXiv preprint arXiv:2510.21571*, 2025. 2
- [18] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791, 2023. 2
- [19] Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th international joint conference on Artificial intelligence*, pages 674–679, 1981. 2
- [20] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Ubaweja, Michael Hays, Fan Zhang, Chuoling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 2, 3
- [21] Abhiram Maddukuri, Zhenyu Jiang, Lawrence Yunliang Chen, Soroush Nasiriany, Yuqi Xie, Yu Fang, Wenqi Huang, Zu Wang, Zhenjia Xu, Nikita Chernyadev, et al. Sim-and-real co-training: A simple recipe for vision-based robotic manipulation. *arXiv preprint arXiv:2503.24361*, 2025. 2
- [22] Andrew Y Ng and Stuart J Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 663–670, 2000. 3
- [23] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024. 2
- [24] Dean A Pomerleau. Alvin: An autonomous land vehicle

- in a neural network. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 1989. [3](#)
- [25] 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 *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. [2](#)
 - [26] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022. [3](#)
 - [27] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. [3](#), [2](#)
 - [28] Modi Shi, Li Chen, Jin Chen, Yuxiang Lu, Chiming Liu, Guanghui Ren, Ping Luo, Di Huang, Maoqing Yao, and Hongyang Li. Is diversity all you need for scalable robotic manipulation? *arXiv preprint arXiv:2507.06219*, 2025. [2](#)
 - [29] Andreas Steiner, André Susano Pinto, Michael Tschanen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, et al. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*, 2024. [3](#), [2](#)
 - [30] Hengkai Tan, Xuezhou Xu, Chengyang Ying, Xinyi Mao, Songming Liu, Xingxing Zhang, Hang Su, and Jun Zhu. Manibox: Enhancing spatial grasping generalization via scalable simulation data generation. *arXiv preprint arXiv:2411.01850*, 2024. [2](#)
 - [31] Junjie Wen, Yichen Zhu, Jinming Li, Zhibin Tang, Chaomin Shen, and Feifei Feng. Dexvla: Vision-language model with plug-in diffusion expert for general robot control. *arXiv preprint arXiv:2502.05855*, 2025. [3](#)
 - [32] Junjie Wen, Yichen Zhu, Minjie Zhu, Zhibin Tang, Jinming Li, Zhongyi Zhou, Xiaoyu Liu, Chaomin Shen, Yaxin Peng, and Feifei Feng. Diffusionvla: Scaling robot foundation models via unified diffusion and autoregression. In *Forty-second International Conference on Machine Learning*, 2025. [3](#)