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

Galaxy Open-World Dataset and G0 Dual-System VLA Model

Tao Jiang[1,*], Tianyuan Yuan[1,2,*], Yicheng Liu[1,2,*], Chenhao Lu[1,2], Jianning Cui[1], Xiao Liu[1], Shuiqi Cheng[1], Jiyang Gao[1], Huazhe Xu[1,2], Hang Zhao[1,2]

[1]Galaxea AI [2]IIIS, Tsinghua University

[*]Equal contribution

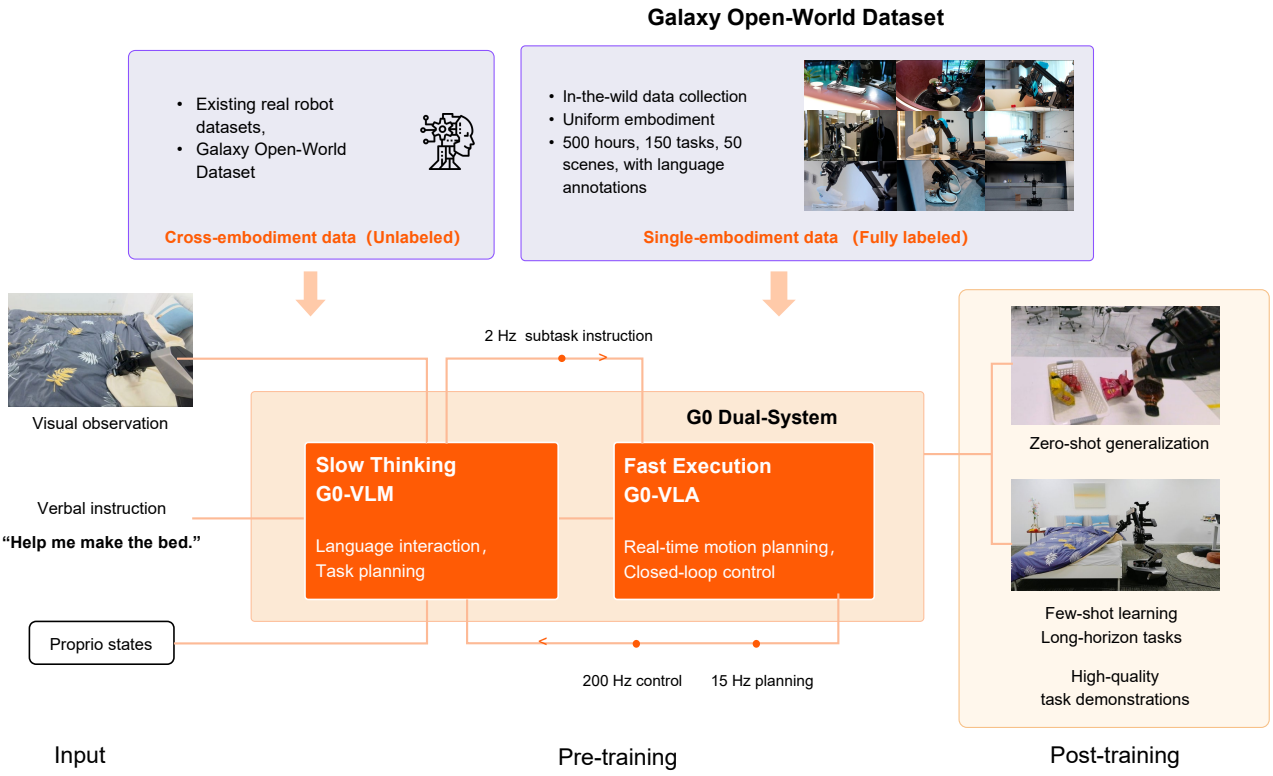


Fig. 1: We introduce Galaxy Open-World Dataset, a high-quality robot behavior dataset collected in the open world. Building on this dataset, we propose G0, a dual system which is composed of a VLM for slow thinking and a VLA model for fast execution.

***Abstract*— We present Galaxy Open-World Dataset, a large-scale, diverse collection of robot behaviors recorded in authentic human living and working environments. All demonstrations are gathered using a consistent robotic embodiment, paired with precise subtask-level language annotations to facilitate both training and evaluation. Building on this dataset, we introduce G0, a dual-system framework that couples a Vision-Language Model (VLM) for multimodal planning with a Vision-Language-Action (VLA) model for fine-grained execution. G0 is trained using a three-stage curriculum: cross-embodiment pre-training, single-embodiment pre-training, and task-specific post-training. A comprehensive benchmark—spanning tabletop manipulation, few-shot learning, and long-horizon mobile manipulation—demonstrates the effectiveness of our approach. In particular, we find that the single-embodiment pre-training stage, together with the Galaxy Open-World Dataset, plays a critical role in achieving strong performance. Dataset, code and pretrained weights will be made publicly available.**

I. Introduction

Vision-Language-Action (VLA) models have emerged as a pivotal paradigm aimed at enabling robots to autonomously perceive, reason, and perform complex tasks in the physical world. Despite significant progress, a substantial bottleneck persists due to the scarcity of large-scale, high-quality, open-world robot data. Existing datasets, exemplified by Open-X Embodiment [1], are predominantly restricted by their limited task realism and insufficient environmental richness. These limitations impair the generalization of trained models when confronted with diverse real-world contexts.

In response to this challenge, we present **Galaxy Open-World Dataset**, an extensive, meticulously curated open real-world dataset for mobile manipulation. Galaxy Open-World Dataset comprises 500 hours of high-fidelity data systematically gathered in real-world scenarios where human individuals live and work, incorporating more than 150 distinct tasks across 50 different scenes. Uniquely, Galaxy Open-World Dataset was consistently captured using a single robotic embodiment, thereby ensuring uniformity and reliability. Comprehensive data filtering and precise language annotations further enrich the dataset, facilitating the benchmarking of mobile manipulation methodologies.

Complementing the dataset, we propose **G0, a dual system framework**. G0 capitalizes on System 2 (G0-VLM) for generalized multimodal planning, directing System 1 (G0-VLA) to perform precise action execution. The systems run asynchronously at different frequencies, enabling both efficient training and deployment. Importantly, we propose a 3-stage training curriculum for G0-VLA: (1) cross-embodiment pre-training on large-scale unlabeled datasets to acquire general world knowledge priors; (2) single-embodiment pre-training on our Galaxy Open-World Dataset to specialize in the perceptual-action pairs on the target platform; and (3) post-training on high-quality task demonstrations for the mastery of specific complex skills.

Finally, we develop a comprehensive benchmark spanning tabletop manipulation, device operation, and long-horizon tasks such as bed making, evaluated under both standard and few-shot learning settings. Experiments reveal that our high-quality dataset and the proposed pre-training strategy are effective in improving the dual system's performance. Notably, when there is a large embodiment gap between the pre-training platform and the target robot, the benefits of cross-embodiment pre-training diminish or can even degrade the VLA model's performance, underscoring the importance of the proposed single-embodiment pre-training stage.

II. Related Work

Dual System Designs. Our G0 model architecture builds on the foundation of hierarchical planning in robotics. In early methods such as Task and Motion Planning (TAMP) [2], high-level task planning and low-level motion control were often decoupled. The advent of VLMs has recently revitalized this paradigm. For example, SayCan [3] demonstrated that a pretrained LLM can serve as a zero-shot planner for high-level goals. Inspired by this, the community has started to adopt dual-system frameworks based on Kahneman's theory of System 1 (fast, reactive) and System 2 (deliberative, planning) [4]. This hierarchical approach, separating deliberate planning from reactive control, forms the basis of our work.

VLA as the System 1 Executor. The rise of VLA models has provided a powerful paradigm for building generalist robot policies. Within a dual-system framework, these VLAs are a natural fit for the System 1 executor: a reactive policy that translates immediate sensory inputs and simple instructions into low-level robot control actions [5]–[7].

The action generation module in these VLA models employs two prevailing paradigms: autoregressive generation [8]–[11] and diffusion generation [12]–[14]. Autoregressive models excel at transferring knowledge from pretrained VLMs but can be slow [11], [15], while diffusion models offer higher throughput but risk degrading the VLM's original capabilities [16]. Hybrid designs attempt to combine their advantages [17]; G0-VLA adopts a similar approach.

VLM as the System 2 Planner. System 2 VLMs provide high-level guidance, decomposing complex commands into subtasks for System 1 [18]–[20]. We focus on constructing and fine-tuning the VLM planner, systematically comparing open-source models against human and closed-source baselines.

Large-scale Manipulation Datasets. Dataset scale and diversity are critical for VLA performance. Prior efforts include single-platform datasets like BridgeData V2 [21] and DROID [22], and multi-embodiment datasets like Open-X Embodiment [1]. While these improve scale or diversity, most operate in controlled settings, leaving a domain gap for unstructured real-world environments [23], [24]. More recent efforts like RoboMIND [23] and AgiBot world [24], despite pushing the boundaries of scale and task complexity, still operate within this limitation.

Our Galaxy Open-World Dataset addresses this gap with large-scale real-world data. We also study pre-training paradigms, examining how cross-embodiment and target-embodiment training affect generalization, providing a high-fidelity benchmark to disentangle their contributions [12], [25]–[27].

III. Galaxy Open-World Dataset

The Galaxy Open-World Dataset is a large-scale, high-quality, fully annotated dataset. It contains 100K demonstration trajectories, covering 500 hours across 150 task categories in 50 real-world scenes, involving over 1,600 objects and 58 operational skills, from fine-grained pick-and-place to coordinated whole-body manipulation. Data are collected with a consistent embodiment, ensuring alignment of perception, action, and language annotations.

Data Collection Platform. Demonstrations use the Galaxea R1 Lite (Fig. 2a), a commercially available dual-arm mobile robot with 23 DoF: two 6-DoF arms, a 3-DoF torso, and a 6-DoF omnidirectional base (up to 1.5 m/s). Spherical wrists and parallel grippers support payloads up to 5 kg and reaches of 60 cm. Perception combines a stereo RGB head camera and dual wrist-mounted RGB-D cameras. The compact design (1280 mm height, 600 mm chassis width) enables navigation in tight spaces.

We adopt an *isomorphic teleoperation* scheme, mapping human movements directly to the robot's kinematics. This approach keeps arms within reachable postures, avoids IK failures, and removes the need for re-targeting between human and robot morphologies.

Collection Guidelines. Data collection follows three principles: (1) *Observability* — key objects remain in view; (2) *Data quantity and quality* — simple tasks use ∼100 high-quality demonstrations, complex tasks are collected after pilot

(a) Data collection platform.

- Open-world scenes including residentials, kitchens, retails, offices.
- Objects and tasks defined by the scenes.

(b) Data collected in diverse scenes.

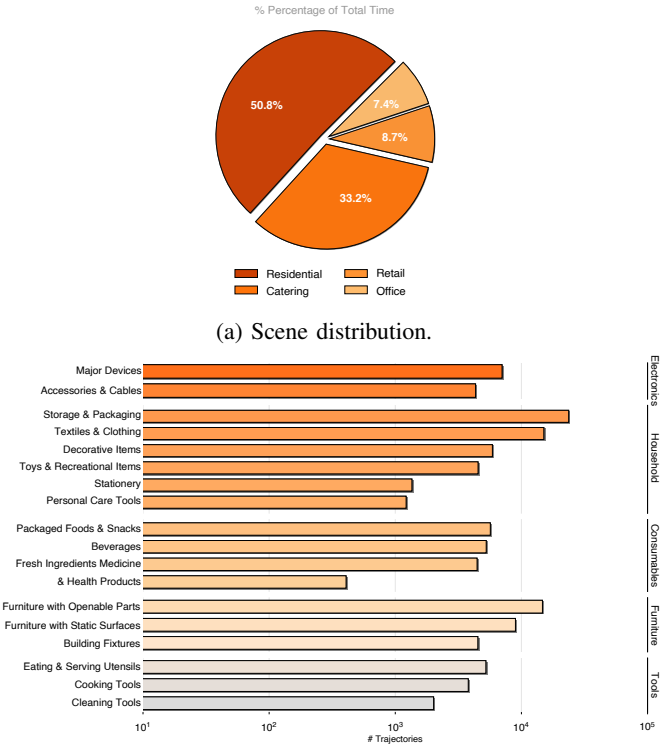
Fig. 2: **Galaxy Open-World Dataset**. Collected by a fleet of robots with identical embodiments across diverse real-world environments.

validation; (3) *Linguistic grounding* — subtasks are annotated with structured language for multi-modal alignment.

Environmental and Object Diversity. The dataset is collected at 11 physical sites—covering residential, catering, retail, and office spaces. Each site provides multiple operational zones, yielding a total of **50 unique scenes**. Object sets are sourced from real-world retail suppliers to ensure realistic visual and physical properties. For items that are unsafe or impractical to manipulate repeatedly (e.g., food), high-fidelity replicas are used to preserve visual realism while maintaining hygiene and efficiency.

Annotation Process. Episodes are segmented into atomic subtasks with a fixed annotation schema, improving speed and consistency. Rigorous quality checks remove episodes with operator errors or abnormal sensor data. Examples are shown in Fig. 5

Comparison with Existing Datasets. Compared to BridgeData [28], RT-1 [25], Open-X-Embodiment [1], and AgiBot World [24], Galaxy Open-World Dataset provides (i) *single-embodiment consistency*, (ii) fine-grained subtask annotations, and (iii) broader real-world scene diversity. These properties make it a strong benchmark for generalizable VLA models in



(a) Scene distribution.

(b) Object distribution.

Fig. 3: **Data diversity statistics. (a)** The distribution of scenes in dataset. **(b)** Trajectory counts categorized by objects, showcasing the dataset's wide range of interactive items.

unstructured environments. More statistics are listed in Fig. 3 and 4.

IV. Method

A. Dual System Overview

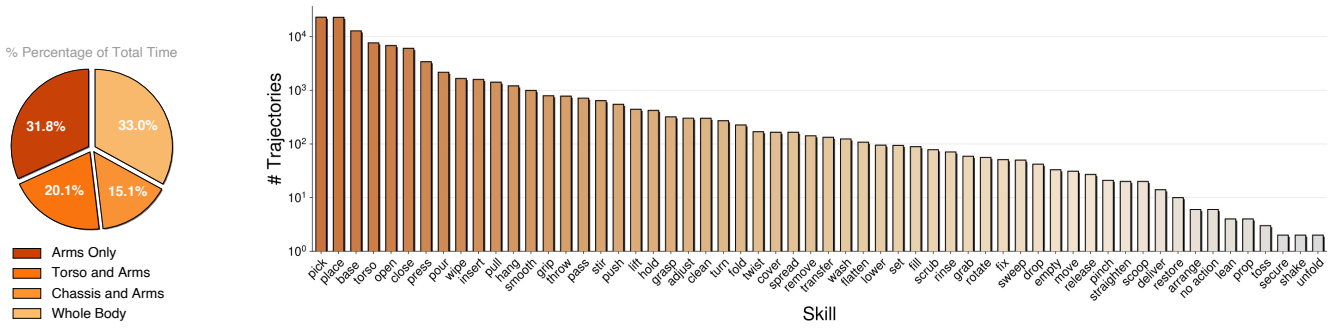
Our G0 dual system (Fig. 1) comprises a fast-response System-1 (VLA) and a deliberative System-2 (VLM). System-1 perceives the environment, interprets subtask instructions, and executes actions on a bi-manual mobile robot. At each time t, it generates an action chunk $\boldsymbol{A}_t = a_{t:t+k}$ conditioned on language l, visual observations o_t, and proprioceptive state s_t. G0-VLA embeds visual and language inputs via a pre-trained VLM, then generates continuous actions through a flow-matching action expert.

System-2 plans at a high level, processing task instructions and scene context to produce subtask directives for System-1.

Training differs for the two systems. G0-VLM is trained on image-subtask pairs from Galaxy Open-World Dataset. G0-VLA uses a 3-stage curriculum: (1) cross-embodiment pre-training on diverse robotics data to acquire general priors, (2) single-embodiment pre-training on Galaxy to specialize in the target platform, and (3) post-training on high-quality task demonstrations.

B. G0-VLA Pre-training Stage-1

Stage-1 trains only the VLM backbone using the FAST tokenizer [11] to convert continuous action chunks into



(a) Body part usage. (b) Skill distribution.

Fig. 4: **Embodied behavior statistics. (a)** Interaction time by body part, from simple 'Arms Only' manipulations to coordinated 'Whole Body' movements. **(b)** Long-tail skill distribution, covering common actions, covering both frequent actions (*e.g.*, 'pick', 'place') and diverse specialized skills.

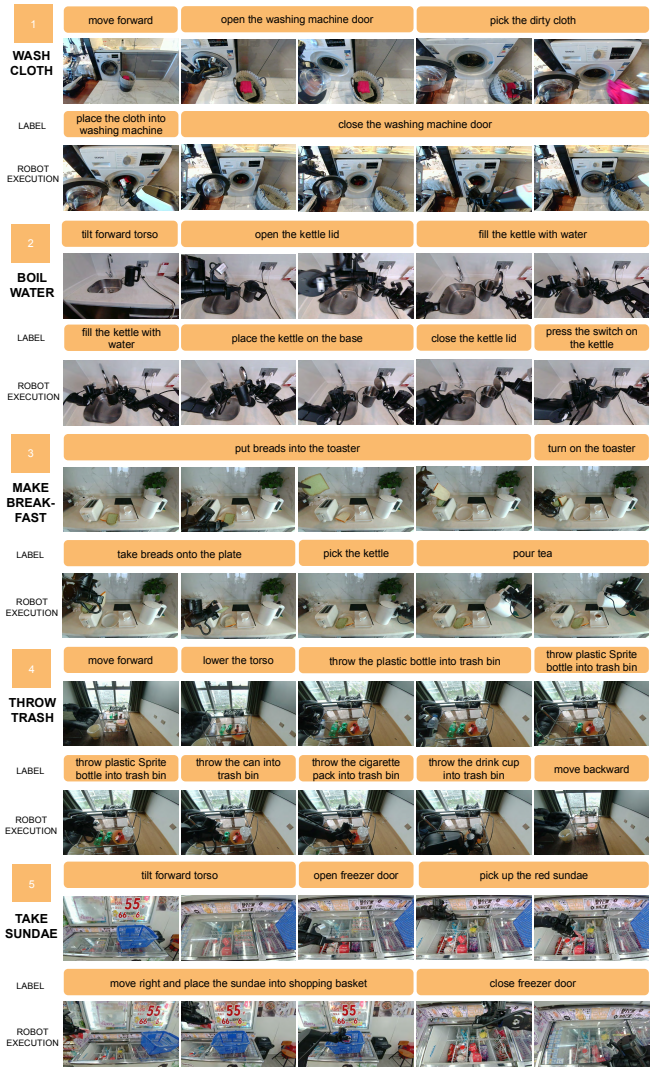


Fig. 5: **Data samples with temporal subtask annotations.** Covers diverse daily-life scenes, from dual-arm to whole-body manipulation, with high-quality, fine-grained subtask labels.

discrete tokens. The VLM predicts action tokens autoregressively. Specifically, given the image observations o_t, language instruction l_t, and proprioceptive state s_t at time t the policy is trained to model the conditional distribution over action tokens:

$$p(\mathbf{A}_t^d) = \prod_{i=1}^{N} p(a_i^d \mid a_{<i}^d, o_t, l_t, s_t),$$

where $\mathbf{A}_t^d$ denotes the N discrete action tokens a^d produced by the action tokenizer. The VLM uses PaLiGemma [29] with a SigLIP vision encoder, an MLP projector, and a Transformer to fuse visual, language, and proprioceptive embeddings.

In this stage, we train the VLM on a diverse mixture of 1,000 hours of OXE trajectories, 500 hours from the Galaxy Open-World Dataset with high-level instructions only, and 200 hours of in-house data with coarse instructions.

The motivation for training only the VLM in Stage-1 is twofold. First, the data come from multiple embodiments, with varying annotation quality and action accuracy, making it unreliable for training the action expert. Second, applying diffusion loss too early can destabilize learning before the model has acquired stable representations.

C. G0-VLA Pre-training Stage-2

Stage-2 trains the action expert on labeled Galaxy Open-World Dataset. The VLA consists of the pre-trained VLM and a newly initialized action expert. The action expert generates continuous actions conditioned on proprioceptive states and the representations generated by the VLM. Specifically, we train our VLA by maximizing the following objective:

$$\max_{\theta} \mathbb{E}_{p(\boldsymbol{A}_t, o_t, l_t, s_t)} \left[\log \pi_\theta(\boldsymbol{A}_t \mid o_t, l_t, s_t)\right]$$

with a flow-matching loss

$$\mathscr{L}_{\text{flow}}(\theta) = \mathbb{E}_{p(\boldsymbol{A}_t^\tau \mid o_t, l_t, s_t)} \left[\left\| v_\theta(\boldsymbol{A}_t^\tau, \tau, o_t, l_t, s_t) - \boldsymbol{u}(\boldsymbol{A}_t^\tau \mid \boldsymbol{A}_t) \right\|^2\right]$$

Here, $\boldsymbol{A}_t$ denotes the action chunk from time t with horizon H, o_t is the visual observation, l_t is the language instruction, and s_t is the proprioceptive state. $\boldsymbol{A}_t^\tau$ is the interpolated noisy action $\boldsymbol{A}_t^\tau = \tau \boldsymbol{A}_t + (1-\tau)\varepsilon$. $v_\theta(\cdot)$ is the flow predicted by

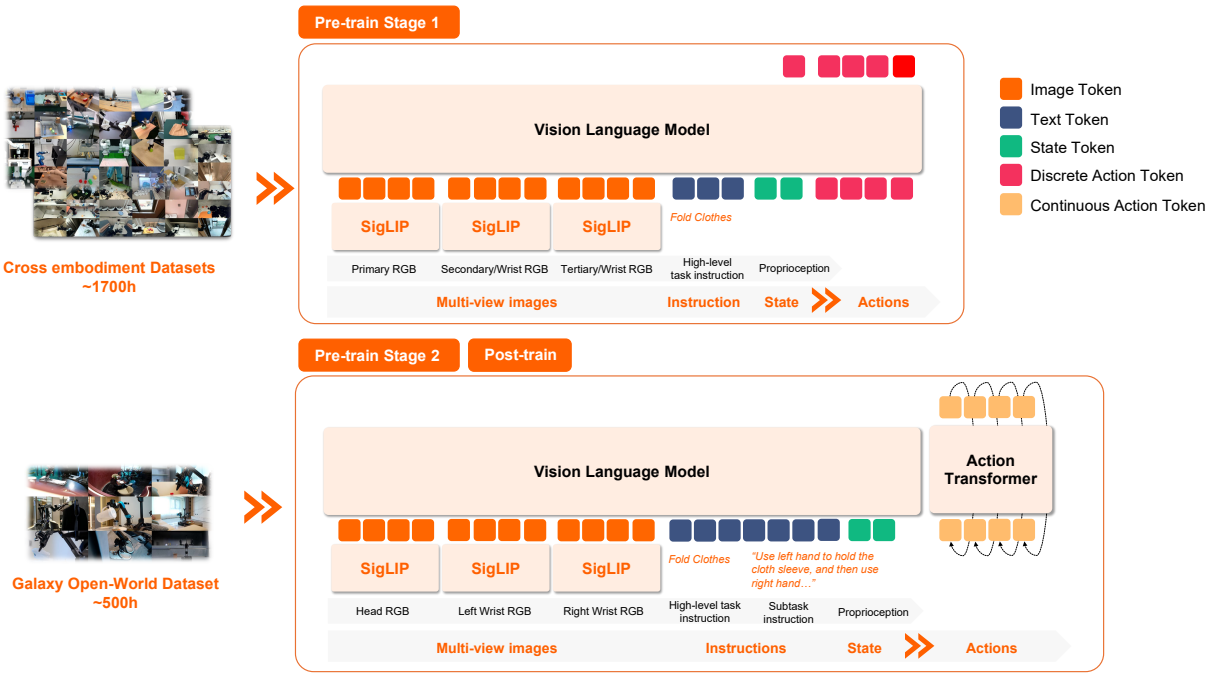


Fig. 6: G0-VLA architecture and our 3-stage training pipeline. Stage 1 pre-trains a vision-language model on cross-embodiment data in an autoregressive manner. Stage 2 and post-train share the same model structure, trained on Galaxy open-world data with embodiment-specific views and high-level and subtask instructions, by supervising the Action Transformer's action reconstruction with a flow-matching loss. Color codes indicate token modalities.

the VLA and $\boldsymbol{u}(\cdot)$ is the target flow derived from the action trajectory.

Pre-training Stage-2 focuses on improving the action precision and language grounding capabilities, enabled by two key features of the Galaxy Open-World Dataset: 1) **single embodiment**: All trajectories are collected on a single robotic platform, ensuring a consistent action space and eliminating the need for the action expert to adapt across embodiments. 2) **language-action alignment**: Instructions and trajectories are segmented at the subtask level, producing fine-grained language-action pairs. This promotes a stronger correspondence between instructions and robot actions.

D. VLA Post-training: Task-oriented Training

To test the generalization ability of pre-trained models, we fine-tune our VLA with different pre-trained weights on downstream tasks, using the same training objective as Stage-2. For each task, we limit the fine-tuning data to a maximum of 100 trajectories.

E. G0-VLM Training

G0-VLM is the high-level planner, responsible for interpreting human instructions, performing task planning, generating verbal responses, and sending atomic action directives to G0-VLA. We start from the open-source Qwen2.5-VL [30] and perform instruction tuning using Galaxy Open-World data.

To train G0-VLM in a scalable manner, we utilize human-annotated subtasks alongside synthesized human-style high-level instructions. Episodes are sampled with higher weight on key frames, such as subtask transitions or gripper changes, to emphasize task boundaries. Head-camera images and k-frame historical observations and actions at 1-second intervals are included to capture temporal context, forming D_{labeled} with task names, observations $o_{t-k}, \ldots, o_t$, and subtask instructions $l_{t-k}, \ldots, l_t$.

Fig. 7: Evaluation benchmarks.

Subsequently, we apply a reasoning LLM (DeepSeek-R1) to convert D_{labeled} into human-style high-level instructions and robot verbal responses. For each episode, the LLM receives the task name, historic and current subtasks, and the next subtask, then produces a natural verbal instruction and the robot's response (*e.g.*, *"I am going to be seated, could you help pull the chair out?"*) and the robot's verbal response to the human (*e.g.*, *"I am working on it!"*). Images are not fed to the LLM, as atomic action annotations suffice for inferring task scenarios.

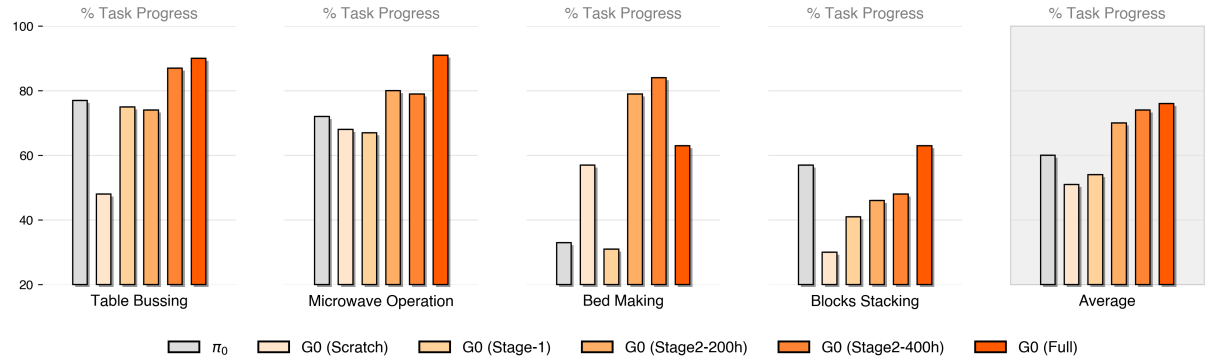


Fig. 8: **Fine-tuning benchmark results of different pre-trained VLAs.** G0 (Full) achieves the highest average progress score, excelling in object-picking tasks such as **Table Bussing**, **Microwave Operation**, and **Bed Making**. G0 (Stage-2) lead in language following, action consistency, and whole-body control. G0 (Stage-1) performs the worst among pre-trained models, highlighting the necessity of uniform-embodiment pre-training.

V. Evaluating G0-VLA

We build challenging real-world benchmarks (Figure 7) to evaluate G0-VLA and the impact of our dataset. Our central question is: **How does pre-training data influence VLA?** We explore this through three perspectives: (1) Does pre-training enhance fine-tuning performance on downstream tasks? How much does the pre-trained weights matter? (2) Can pre-training on a single embodiment accelerate few-shot transfer? (3) How do single-embodiment and cross-embodiment pre-training compare in embodiment-specific actions? Our benchmarks consist of the following tasks:

Table bussing: The robot is required to organize a cluttered desk by placing pens into a pen holder, picking up and hanging headphones, and moving a book onto a book stand. This task evaluates the model's capability in precise pick-and-place, coordinated dual-arm manipulation, and maintaining object stability.

Microwave operation: The robot opens a microwave door, places food onto a plate, transfers the plate into the microwave, and then closes the door to initiate heating. This task assesses the model's ability in interacting with household appliances and executing multi-step manipulation sequences.

Bed Making: The robot is asked to tidy up the messy quilt on the bed to make the quilt flat and neat. This task emphasizes whole-body control, requiring coordination of the chassis, torso, and arms for effective execution.

Blocks Stacking: The robot is asked to build the blocks to form specific words. This task tests the model's ability in language following and precise pick-and-place.

We evaluate these benchmarks by progress score, where the details of progress for each task are defined in the appendix. For reproducibility, we run each test 10 times and get the average score of each task.

A. Pre-trained Weights

In this experiment, we test the effectiveness of different pre-trained weights. We fine-tune the pre-trained models on our proposed benchmarks, using 100 training trajectories per task (each ranging from 30 seconds to 1 minute in duration). The following configurations are evaluated:

- G0 (Stage-1): VLA with only Stage-1 pre-training.
- G0 (Stage-2 200h): VLA with only Stage-2 pre-training (200 hours of data).
- G0 (Stage-2 400h): VLA with only Stage-2 pre-training (400 hours of data).
- G0 (Full): VLA with Stage-1 followed by Stage-2 pre-training (400 hours of data).
- G0 (Scratch): VLA without any action pre-training (initialized from the original VLM weights).
- π_0: π_0 [12] with officially released pre-trained weights as a baseline.

All models are fine-tuned under identical settings for four epochs. The results are shown in Figure 8. Overall, G0 (Full) achieves the highest average progress score. In particular, it demonstrates superior object-picking ability in **Table Bussing**, **Microwave Operation**, and **Bed Making**. G0 (Stage-2 400h) and G0 (Stage-2 200h) achieve the best performance in language following and action consistency, as well as the strongest whole-body control capabilities, which are further discussed in Section V-C. By contrast, G0 (Stage-1) performs the worst among all pre-trained models, underscoring the importance of single-embodiment pre-training.

We observe that Stage-1 pre-training primarily enhances VLA's ability to perform simple and universal action patterns, such as pick-and-place and push-and-pull. Meanwhile, Stage-2 pre-training grounds the model specifically to our robot platform, leading to improved action stability and instruction following.

B. Few-shot Transfer

In this part, we specifically assess the few-shot transfer capability of our VLA. We fine-tune the model using only 20 trajectories for each of two tasks: **Table Bussing** and **Microwave Operation**. Each model is fine-tuned with the same setting for 10 epochs.

As shown in Figure 9, models with Stage-2 pre-training significantly outperform those without. In addition to this quantitative improvement, we also observe that these models produce noticeably smoother and more stable actions during execution. These results suggest that single-embodiment pre-training substantially enhances few-shot generalization within

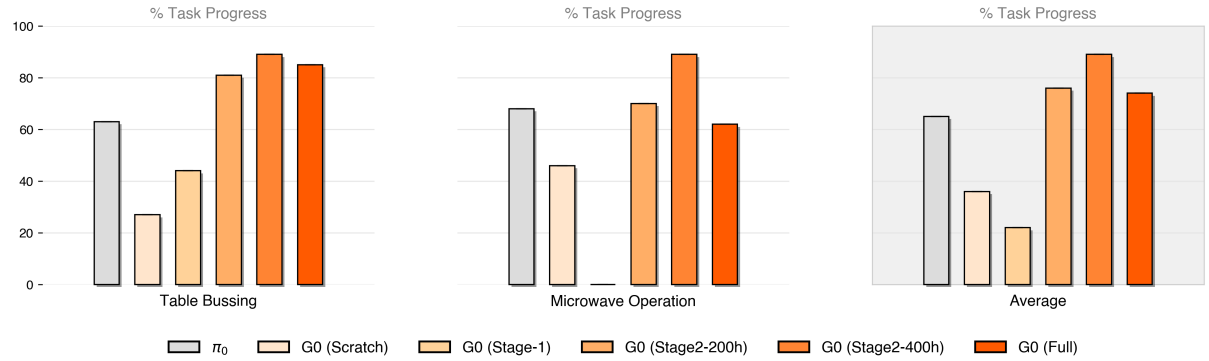


Fig. 9: **Few-shot performance of VLAs.** Few-shot transfer performance on **Table Bussing** and **Microwave Operation**. Stage-2 pre-training markedly improves success rates and execution smoothness, while Stage-1 pre-training alone offers no clear advantage over training from scratch.

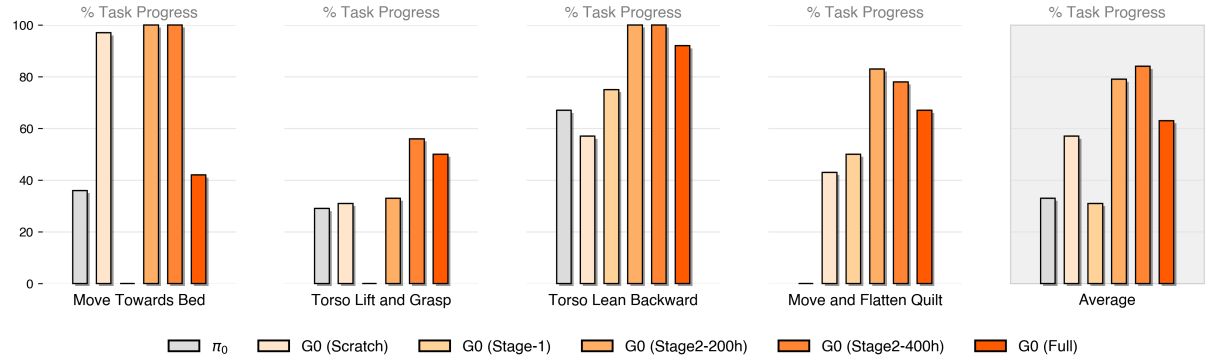


Fig. 10: **Per-skill progress scores on the Bed Making task.** Stage-2 single-embodiment pre-training substantially improves chassis, torso control, while cross-embodiment pre-training (Stage-1, π_0) yields weaker performance, in some cases worse than training from scratch.

the same embodiment, underscoring the importance of single-embodiment data in our Galaxy Open-World Dataset. Notably, models pre-trained solely with Stage-1 do not show a clear advantage over models trained from scratch. This indicates that cross-embodiment action pre-training alone may be insufficient for a model's ability to quickly adapt to a new embodiment in few-shot settings.

C. Embodiment-specific Actions

In this section, we analyze the **Bed Making** task, which is a long-horizon task requiring coordinated and precise whole-body control, including the chassis, torso and arms. These are embodiment-specific behaviors that are not represented in cross-embodiment datasets such as OXE. We report the progress scores by skill in Figure 10.

Stage-2 pre-training on single-embodiment significantly improves the model's performance on these embodiment-specific skills, suggesting that such capabilities are effectively acquired during this stage of pre-training. In contrast, models trained with cross-embodiment data (e.g., Stage-1 pre-training and π_0) demonstrate substantially weaker instruction following for chassis and less accurate torso control. In some cases, they even underperform compared to models trained from scratch. We hypothesize that the large embodiment gap between our robot and those in the OXE dataset used for Stage-1 pre-training hinders the model's ability to acquire embodiment-specific skills. These findings highlight the need to carefully design the use of cross-embodiment data in pre-training strategies.

VI. Evaluating G0-VLM

The effectiveness of G0-VLM depends on two aspects: the accuracy of command-observation alignment, which ensures the VLM correctly interprets observations, and the fidelity of action primitives, which determines whether the VLA can properly execute its commands. We therefore design our evaluation metrics to assess both aspects of VLM performance: (i) whether fine-tuning is necessary compared to using pretrained models directly, and (ii) how supervised fine-tuning (SFT) enhance the VLM's capabilities in robotic tasks, particularly in improving action-grounding accuracy. To address these, we benchmark existing LLMs, including Gemini-2.5-pro and Qwen2.5-VL [30], against our fine-tuned versions. To ensure a fair comparison, we design standardized prompts containing task-specific instructions, atomic action options, and output examples for all baseline models.

Table I demonstrates that our fine-tuned model surpasses baseline accuracy by over 50%, with task-specific tuning enabling language instructions that are directly executable by VLAs. This validates our key hypothesis that robotic

applications require not just general-purpose vision-language understanding, but precisely aligned action primitives through domain adaptation.

TABLE I: Instruction accuracy in benchmark tasks (%).

Model	Table	Microwave	Bed	Blocks
Gemini-2.5-pro	32.0	15.8	54.2	55.0
Qwen2.5-VL-72B	26.3	16.8	48.1	21.7
Qwen2.5-VL-32B	21.3	14.8	54.2	21.0
Qwen2.5-VL-7B	26.3	17.2	46.9	24.7
G0-VLM	83.3	74.2	78.2	75.6

VII. Conclusion

We have introduced Galaxy Open-World Dataset, a large-scale, high-fidelity, and richly annotated resource designed to accelerate research in robotic mobile manipulation. By pre-training on the dataset, we present G0, a dual system composed of a VLM for planning and a VLA model for execution. G0 achieves state-of-the-art performance across a diverse set of benchmarks.

References

[1] Q. Vuong, S. Levine, H. R. Walke, K. Pertsch, A. Singh, R. Doshi, C. Xu, J. Luo, L. Tan, D. Shah, *et al.*, "Open x-embodiment: Robotic learning datasets and rt-x models," in *Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition@ CoRL2023*, 2023.

[2] C. R. Garrett, R. Chitnis, R. Holladay, B. Kim, T. Silver, L. P. Kaelbling, and T. Lozano-Pérez, "Integrated task and motion planning," *Annual review of control, robotics, and autonomous systems*, vol. 4, no. 1, pp. 265–293, 2021.

[3] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," *arXiv preprint arXiv:2204.01691*, 2022.

[4] D. Kahneman, *Thinking, fast and slow*. macmillan, 2011.

[5] J. Chen, H. Guo, K. Yi, B. Li, and M. Elhoseiny, "Visualgpt: Data-efficient adaptation of pretrained language models for image captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 030–18 040.

[6] W. Shi, X. Han, C. Zhou, W. Liang, X. V. Lin, L. Zettlemoyer, and L. Yu, "Lmfusion: Adapting pretrained language models for multimodal generation," *arXiv preprint arXiv:2412.15188*, 2024.

[7] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, S. XiXuan, *et al.*, "Cogvlm: Visual expert for pretrained language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 121 475–121 499, 2024.

[8] A. Szot, B. Mazoure, O. Attia, A. Timofeev, H. Agrawal, D. Hjelm, Z. Gan, Z. Kira, and A. Toshev, "From multimodal llms to generalist embodied agents: Methods and lessons," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 10 644–10 655.

[9] S. Belkhale and D. Sadigh, "Minivla: A better vla with a smaller footprint," 2024. [Online]. Available: https://github.com/Stanford-ILIAD/openvla-mini

[10] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[11] K. Pertsch, K. Stachowicz, B. Ichter, D. Driess, S. Nair, Q. Vuong, O. Mees, C. Finn, and S. Levine, "Fast: Efficient action tokenization for vision-language-action models," *arXiv preprint arXiv:2501.09747*, 2025.

[12] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, *et al.*, "π_0: A vision-language-action flow model for general robot control," *arXiv preprint arXiv:2410.24164*, 2024.

[13] S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu, H. Su, and J. Zhu, "Rdt-1b: a diffusion foundation model for bimanual manipulation," *arXiv preprint arXiv:2410.07864*, 2024.

[14] Q. Li, Y. Liang, Z. Wang, L. Luo, X. Chen, M. Liao, F. Wei, Y. Deng, S. Xu, Y. Zhang, *et al.*, "Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation," *arXiv preprint arXiv:2411.19650*, 2024.

[15] J. Wen, Y. Zhu, J. Li, M. Zhu, Z. Tang, K. Wu, Z. Xu, N. Liu, R. Cheng, C. Shen, *et al.*, "Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation," *IEEE Robotics and Automation Letters*, 2025.

[16] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, *et al.*, "$pi_{0.5}$: a vision-language-action model with open-world generalization," *arXiv preprint arXiv:2504.16054*, 2025.

[17] J. Liu, H. Chen, P. An, Z. Liu, R. Zhang, C. Gu, X. Li, Z. Guo, S. Chen, M. Liu, *et al.*, "Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model," *arXiv preprint arXiv:2503.10631*, 2025.

[18] C. Cui, P. Ding, W. Song, S. Bai, X. Tong, Z. Ge, R. Suo, W. Zhou, Y. Liu, B. Jia, H. Zhao, S. Huang, and D. Wang, "Openhelix: A short survey, empirical analysis, and open-source dual-system vla model for robotic manipulation," 2025. [Online]. Available: https://arxiv.org/abs/2505.03912

[19] L. X. Shi, B. Ichter, M. Equi, L. Ke, K. Pertsch, Q. Vuong, J. Tanner, A. Walling, H. Wang, N. Fusai, *et al.*, "Hi robot: Open-ended instruction following with hierarchical vision-language-action models," *arXiv preprint arXiv:2502.19417*, 2025.

[20] C. Gao, Z. Liu, Z. Chi, J. Huang, X. Fei, Y. Hou, Y. Zhang, Y. Lin, Z. Fang, Z. Jiang, *et al.*, "Vla-os: Structuring and dissecting planning representations and paradigms in vision-language-action models," *arXiv preprint arXiv:2506.17561*, 2025.

[21] H. R. Walke, K. Black, T. Z. Zhao, Q. Vuong, C. Zheng, P. Hansen-Estruch, A. W. He, V. Myers, M. J. Kim, M. Du, *et al.*, "Bridgedata v2: A dataset for robot learning at scale," in *Conference on Robot Learning*. PMLR, 2023, pp. 1723–1736.

[22] A. Khazatsky, K. Pertsch, S. Nair, A. Balakrishna, S. Dasari, S. Karamcheti, S. Nasiriany, M. K. Srirama, L. Y. Chen, K. Ellis, *et al.*, "Droid: A large-scale in-the-wild robot manipulation dataset," *arXiv preprint arXiv:2403.12945*, 2024.

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

[24] Q. Bu, J. Cai, L. Chen, X. Cui, Y. Ding, S. Feng, S. Gao, X. He, X. Hu, X. Huang, *et al.*, "Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems," *arXiv preprint arXiv:2503.06669*, 2025.

[25] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Conference on Robot Learning*. PMLR, 2023, pp. 2165–2183.

[26] S. Zhang, Z. Xu, P. Liu, X. Yu, Y. Li, Q. Gao, Z. Fei, Z. Yin, Z. Wu, Y.-G. Jiang, *et al.*, "Vlabench: A large-scale benchmark for language-conditioned robotics manipulation with long-horizon reasoning tasks," *arXiv preprint arXiv:2412.18194*, 2024.

[27] M. Shi, L. Chen, J. Chen, Y. Lu, C. Liu, G. Ren, P. Luo, D. Huang, M. Yao, and H. Li, "Is diversity all you need for scalable robotic manipulation?" *arXiv preprint arXiv:2507.06219*, 2025.

[28] F. Ebert, Y. Yang, K. Schmeckpeper, B. Bucher, G. Georgakis, K. Daniilidis, C. Finn, and S. Levine, "Bridge data: Boosting generalization of robotic skills with cross-domain datasets," *arXiv preprint arXiv:2109.13396*, 2021.

[29] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello, T. Unterthiner, D. Keysers, S. Koppula, F. Liu, A. Grycner, A. Gritsenko, N. Houlsby, M. Kumar, K. Rong, J. Eisenschlos, R. Kabra, M. Bauer, M. Bošnjak, X. Chen, M. Minderer, P. Voigtlaender, I. Bica, I. Balazevic, J. Puigcerver, P. Papalampidi, O. Henaff, X. Xiong, R. Soricut, J. Harmsen, and X. Zhai, "PaliGemma: A versatile 3B VLM for transfer," *arXiv preprint arXiv:2407.07726*, 2024.

[30] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.