lılı ByteDance | Seed

GR-RL: Going Dexterous and Precise for Long-Horizon Robotic Manipulation

ByteDance Seed

Full Author List in Contributions and Acknowledgments

Abstract

We present GR-RL, a robotic learning framework that turns a generalist vision-language-action (VLA) policy into a highly capable specialist for long-horizon dexterous manipulation. Assuming the optimality of human demonstrations is core to existing VLA policies. However, we claim that in highly dexterous and precise manipulation tasks, human demonstrations are noisy and suboptimal. GR-RL proposes a multi-stage training pipeline that filters, augments, and reinforces the demonstrations by reinforcement learning. First, GR-RL learns a vision-language-conditioned *task progress*, filters the demonstration trajectories, and only keeps the transitions that contribute positively to the progress. Specifically, we show that by directly applying offline RL with sparse reward, the resulting Q-values can be treated as a robust progress function. Next, we introduce *morphological symmetry augmentation* that greatly improves the generalization and performance of GR-RL. Lastly, to better align the VLA policy with its deployment behaviors for high-precision control, we perform online RL by learning a latent space noise predictor. With this pipeline, GR-RL is—to our knowledge—the first learning-based policy that can autonomously lace up a shoe by threading shoelaces through multiple eyelets with an 83.3% success rate, a task requiring long-horizon reasoning, millimeter-level precision, and compliant soft-body interaction. We hope GR-RL provides a step toward enabling generalist robot foundations models to specialize into reliable real-world experts.

Correspondence: xiao.ma@bytedance.com
Project Page: https://seed.bytedance.com/gr_rl

1 Introduction

The emergence of large-scale vision-language-action (VLA) models has rapidly advanced the ambition of building generalist robotic agents capable of performing a broad range of tasks given visual observations and free-form natural language instructions. Recent systems [4, 7–11, 51] have demonstrated impressive generalization across objects, environments, and semantic concepts, suggesting that robotics can benefit from the scaling laws that propelled progress in vision-language models.

However, being general is not equivalent to being reliable, and current VLA policies still fall short in two fundamental aspects for real-world deployment: (1) *Dexterity with precision* – millimeter-level control over deformable objects remains unsolved. (2) *Long-horizon robustness* – errors accumulate over steps, and it gets worse when coupled with high-precision dexterous manipulation. Consider the task of threading the shoelaces: (1) the robot should be dexterous enough to handle the deformable objects, including both the shoelace and the shoe; (2) the robot needs millimeter-level control precision to thread the shoelace into the eyelets; (3) the

1
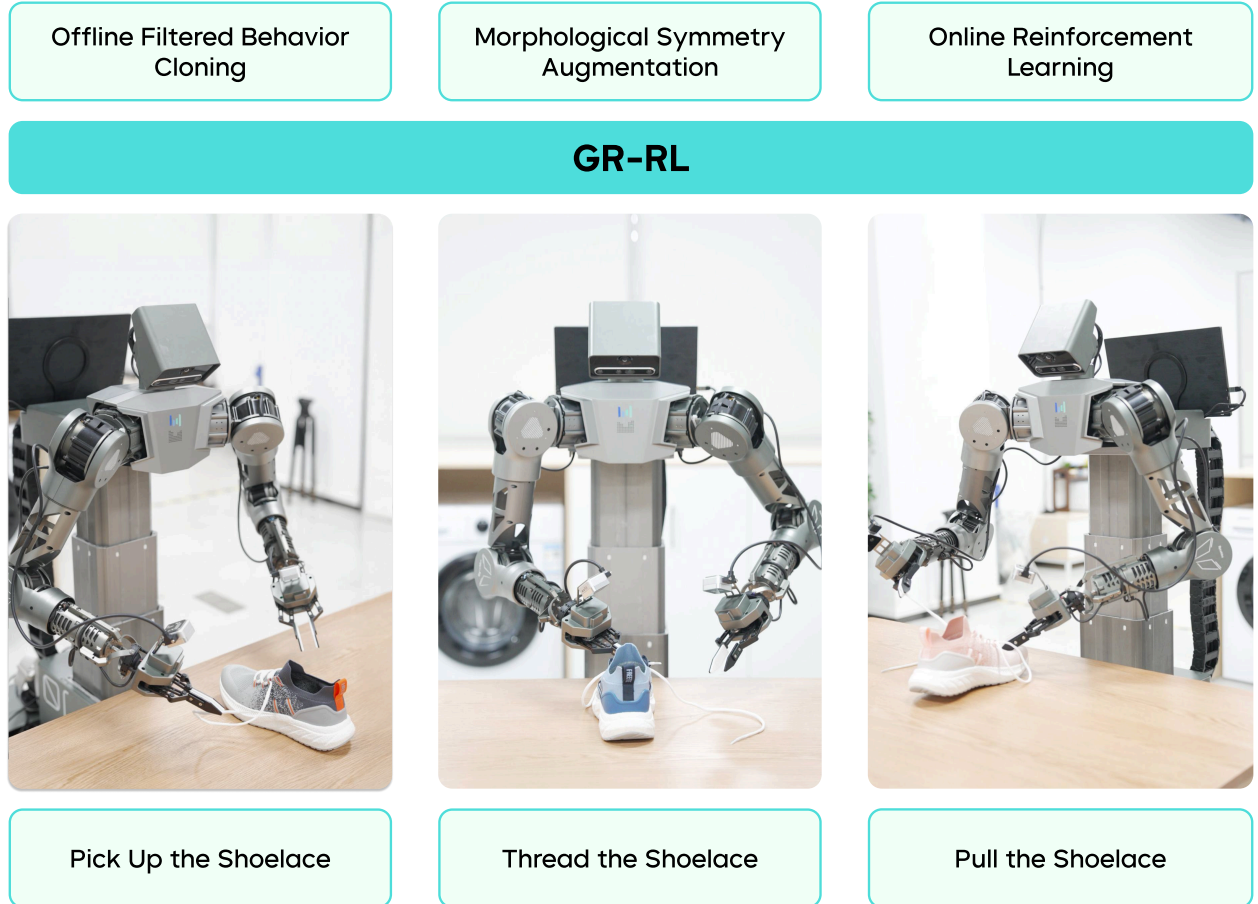
Figure 1 GR-RL performs long-horizon, dexterous, and high-precision manipulation, in the task of shoe lacing, by adopting a multi-stage training pipeline, consisting of 1) offline filtered behavior cloning with learned task progress, 2) simple yet effective action augmentation, 3) online reinforcement learning.

robot also needs long-horizon manipulation capabilities to handle diverse and unexpected scenarios.

Our starting point is GR-3 [12], a large-scale VLA policy trained from internet data, robot trajectories, and human demonstrations. Despite its strong generalization capabilities, GR-3 would fail when precision, dexterity, and long-horizon robustness matter. We observe that there are two key bottlenecks: (1) *suboptimal human demonstrations* and (2) the *demonstration and inference mismatch*. Under extreme precise and dexterous manipulation scenarios, human demonstrators would slow down, hesitate, and introduce noisy suboptimal demonstrations to the policy. In addition, during standard offline training, a VLA policy learns to mimic human demonstrators by predicting action chunks of *fixed lengths* derived from a sliding window given human demonstrations [14, 59]. However, to achieve smooth inference and control, post-smoothing to predicted trajectories (e.g., temporal ensembling [59]), asynchronous receding horizon control [8, 14, 25], and other control-level optimizations are often applied. These system-level optimization methods are necessary for the smooth execution of a learning-based policy, while inevitably causing the mismatch between model training and inference.

We present GR-RL for long-horizon dexterous and precise manipulation. GR-RL adopts a multi-stage reinforcement-augmented training pipeline that filters, augments, and reinforces the suboptimal and mismatched human demonstrations. First, instead of directly running behavior cloning on the entire human demonstration dataset, we initialize the base GR-RL VLA policy by cloning the *filtered* trajectories. Specifically, we train a critic model on both successful and failed trajectories with offline reinforcement learning
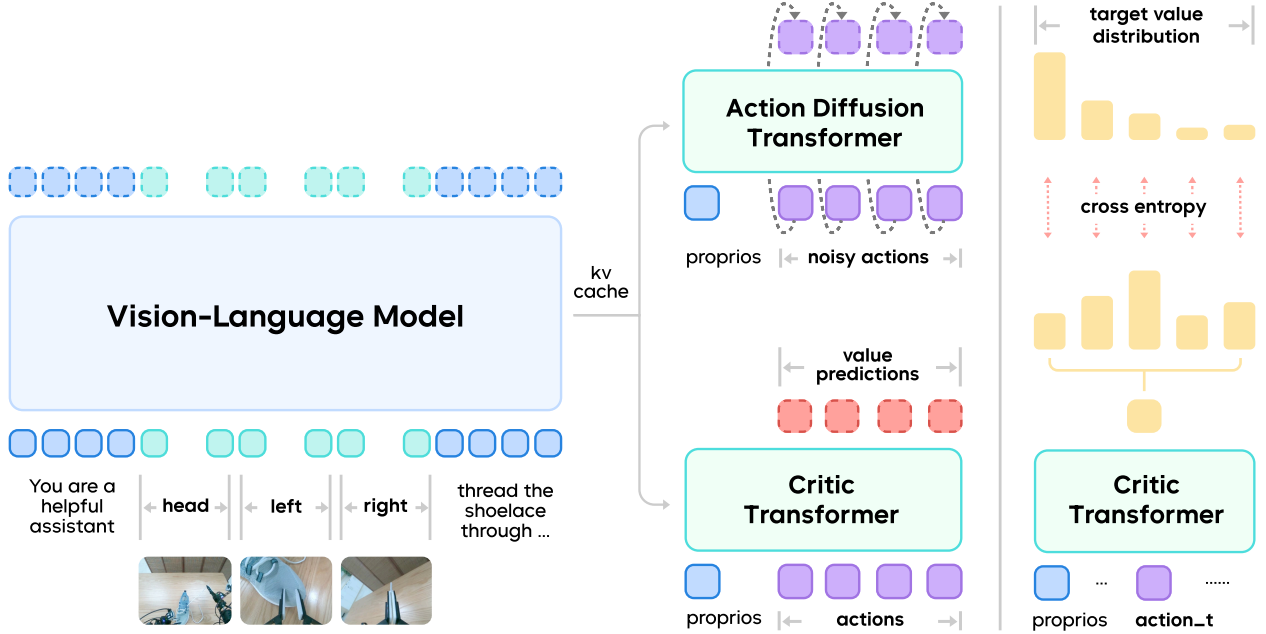
Figure 2 The GR-RL Model. GR-RL adopts a Mixture-of-Transformer (MoT) architecture. It is co-trained on robot vision-language-action trajectories via a flow-matching objective, and Temporal-Difference (TD) errors via distributional reinforcement learning.

(RL) [17]. Given a sparse reward at the end of the episode, the predicted value naturally reflects the progress of the task, which we further use to filter only transitions that contribute positively to the progress and discard the rest. We adopt distributional critics and observe that they give much more robust performance under offline sparse reward scenarios. Next, initialized from the offline pretrained checkpoints, we perform online reinforcement learning to further explore and fix the failure modes of the base policy. In particular, we achieve this by learning to steer the denoising process towards high-return regions [54]. Lastly, we devise a simple yet effective method to augment the robot actions by mirroring the robot actions and observations, with a flipped text description. Such a scheme drastically improves the overall success rate and generalization capabilities of our policy.

To the best of our knowledge, GR-RL is the first model that can thread the shoelace through multiple eyelets, achieving an overall 83.3% success rate. We hope GR-RL provides a step towards enabling generalist robot foundation models to specialize and be applicable in challenging real-world scenarios.

2   The GR-RL Model

GR-RL adopts a Mixture-of-Transformer architecture, consisting of a vision-language-action (VLA) model π_θ and a multi-task critic Q_ϕ with a total of 5B parameters.

Policy: π_θ controls a bi-manual robot with a mobile base by generating a k-length action chunk $\mathbf{a}_t = a_{t:t+k}$ conditioned on the input language instruction l, observation $\mathbf{o}_t$, and robot state $\mathbf{s}_t$, *i.e.*, $\mathbf{a}_t = \pi_\theta(l, \mathbf{o}_t, \mathbf{s}_t)$. Following the architecture design of GR-3 [12], GR-RL uses Qwen2.5-VL-3B-Instruct [3] as the Vision-Language-Model (VLM) backbone, and predicts the action chunks $\mathbf{a}_t$ with an action diffusion transformer (DiT) trained by flow matching objectives [8, 33, 35]. Specifically, we follow GR-3 and use only the KV cache from the latter half of the VLM layers for fast inference.

Critic: Similar to the policy π_θ, $Q_\phi(\mathbf{o}_t, l, \mathbf{s}_t, \mathbf{a}_t)$ is a causal transformer that evaluates each action. Specifically, we follow Q-chunking [30, 47] and predict a chunk of Q-values for each action chunk $\mathbf{a}_t$ and adopt distributional reinforcement learning [5, 16, 23, 47]. Different from unbounded regression-based policy evaluation, distributional critics treat values as a discrete distribution with upper and lower bounds. This naturally

captures uncertainty in real-world trajectories. Under sparse reward settings, distributional critics give much stronger robustness than the non-distributional ones. By setting the upper bound to 1 and the lower bound to 0, our learned critic naturally reflects the *progress* of the task, as shown in Fig. 3.

3 Training Recipe

Human demonstrations are suboptimal and noisy. In the context of long-horizon dexterous high-precision manipulation, human demonstrators tend to hesitate, make mistakes, and try to finish the task with inconsistent behaviors. In addition, inference-time optimizations, such as whole-body receding horizon control and temporal ensembling [14, 59], introduce further mismatch between training and deployment, which exacerbate the negative effect of suboptimal demonstrations.

We introduce a reinforcement-augmented training pipeline to achieve dexterous and precise robotic manipulation from human demonstrations. To prevent the policy from memorizing suboptimal behaviors during supervised learning, we learn a task progress model using offline RL and use it to filter out detrimental data (Sec. 3.1). We then augment the demonstrations to improve the robustness of the offline policy based on the symmetry in bimanual manipulation (Sec. 3.2). Finally, we perform online reinforcement learning, allowing the model to learn from trial-and-error in a closed loop, which mitigates the mismatch between training and deployment and boosts its overall performance (Sec. 3.3).

3.1 Data Filtering with a Learned Task Progress Evaluator

For high-precision manipulation with deformable objects, such as shoe lacing, collecting perfect demonstrations is extremely difficult. Even the trajectories collected by experienced teleoperators contain suboptimal fragments: erroneous attempts, hesitations, etc. Directly imitating all the data would unnecessarily introduce noisy multi-modal actions to the training and lead to a policy with suboptimal performance. However, labeling the suboptimal fragments is non-trivial and might introduce even more subjective and noisy human priors. To identify and filter out suboptimal actions, we propose to learn a task progress model using offline RL. Specifically, we train the critic using TD3+BC [17]. We adopt a sparse reward defined as

$$r(\mathbf{o}_t, l, \mathbf{s}_t, \mathbf{a}_t) = \begin{cases} \gamma^{T-t}\mathbb{I}(\tau), & t > T - k, \\ 0, & t \leq T - k, \end{cases} \tag{1}$$

where $\mathbb{I}(\cdot)$ is an indicator function to evaluate whether a trajectory τ is successful or not, T is the length of the trajectory, and γ means the discount factor. Since most of the collected trajectories end in success, we annotate the retry keyframes in each demonstration and create more failed trajectories in hindsight [1]. Suppose frames $m_i, 0 \leq i < M$ are marked as retry keyframes in a successful trajectory $\tau_{0:T}$, we can augment M failed trajectories $\tau_{0:m_i}, 0 \leq i < M$ in addition to the original successful one. With temporal difference learning over both successful and failed data, the critic Q_ϕ could function as a robust task progress evaluator. After obtaining a task progress model, we evaluate Q_ϕ and compute its mean value of the categorical distribution as progress ρ for all the transitions in the dataset,

$$\rho_t := \mathtt{mean}(Q_\phi(\mathbf{o}_t, l, \mathbf{s}_t, \mathbf{a}_t)). \tag{2}$$

An example of the predicted progress is shown in Fig. 3. We can observe a sudden drop in the progress when the teleoperator makes a mistake. We define a sample $(\mathbf{o}_t, l, \mathbf{s}_t, \mathbf{a}_t)$ at timestep t as suboptimal if there is a value drop greater than a certain threshold δ in the sequence $\rho_{t:t+k}$, and exclude all the suboptimal ones from the dataset for policy learning. We can then train π_θ simply with behavior cloning using the filtered dataset of higher quality.

### 3.2 Imitation Learning with Data Augmentation

During the offline training stage, we apply a simple yet effective *morphological symmetry augmentation* paradigm, which further boosts the policy performance. The augmentation paradigm leverages the morphological symmetry in our bimanual task settings. For image observations $\mathbf{o}_t$, we flip all the images horizontally,
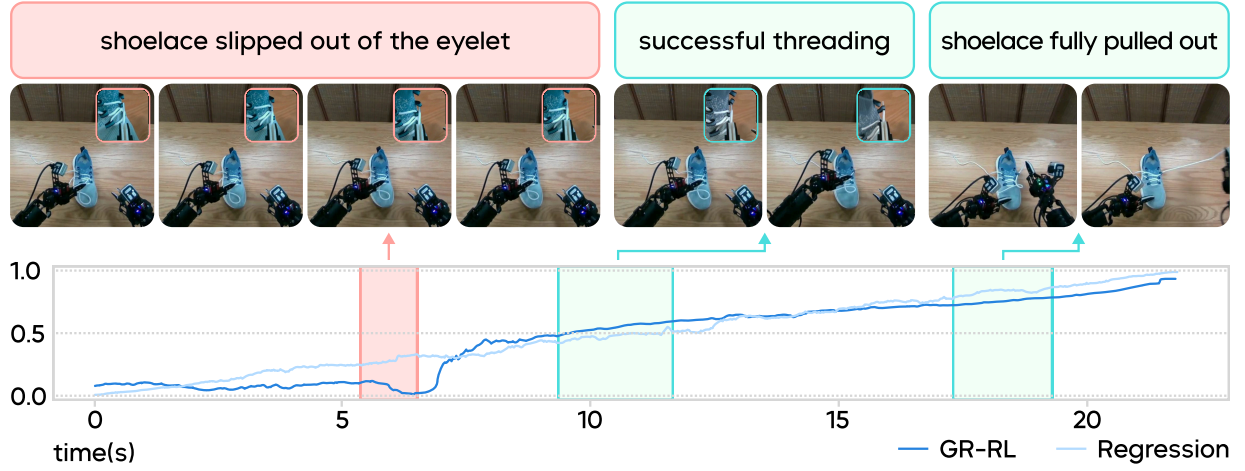
(a) GR-RL value decreases when the shoelace misses the eyelet and increases during precise threading.
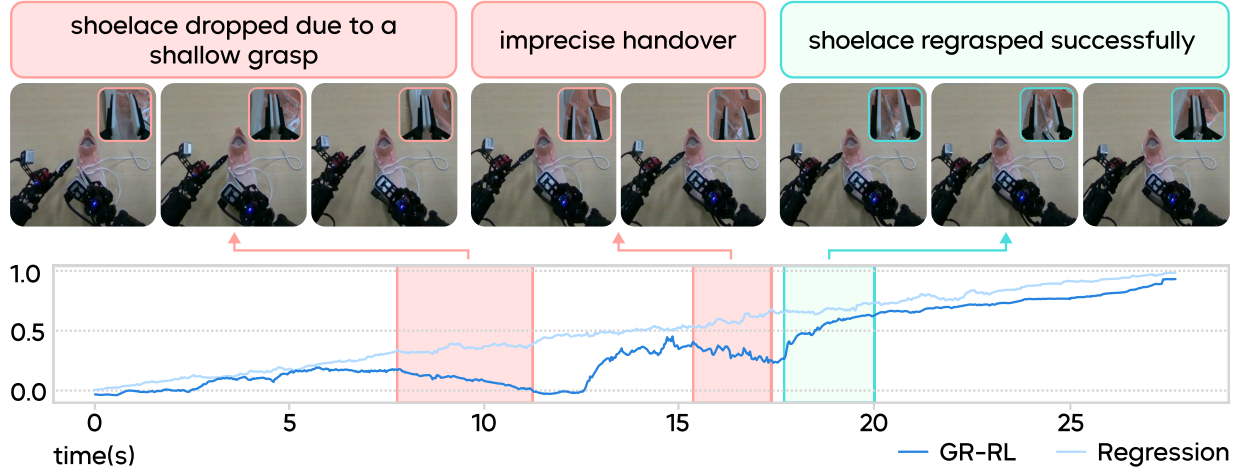
(b) GR-RL value captures multiple failed attempts during shoelace handover.
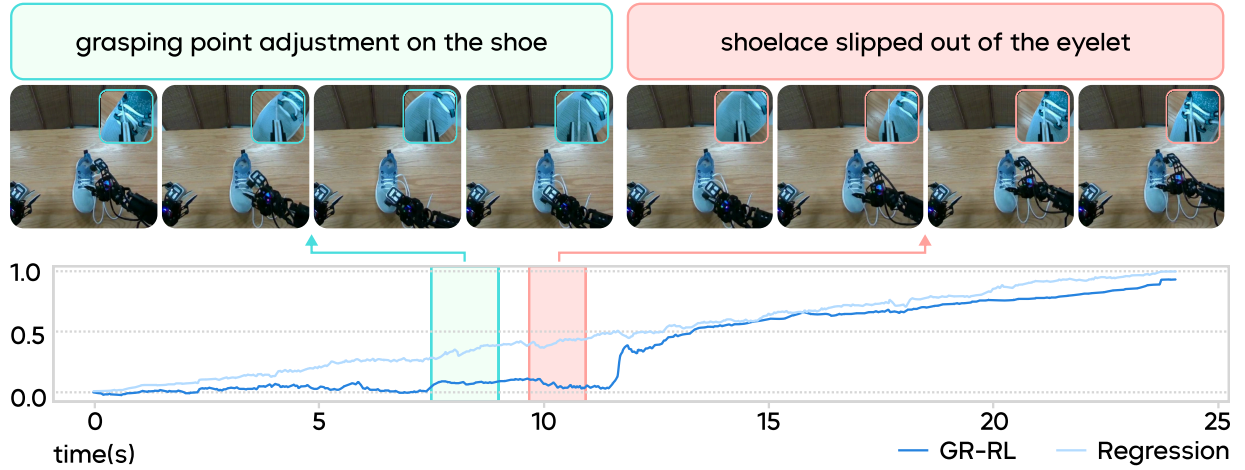
(c) GR-RL value is aware of the long-term positive effect of releasing the shoelace to adjust the grasping pose.

Figure 3 Examples of learned task progress.

5

then swap the images from the left wrist with those from the right wrist. All transformations in proprioception states $\mathbf{s}_t$ and actions $\mathbf{a}_t$ are converted via mirror symmetry in the world frame, and then transformed back to local wrist frames. We also flip the spatial description in the language instructions accordingly, e.g., changing "the hole on the left" to "the hole on the right". Empirically, the symmetry data augmentation can effectively enhance the performance of the policy.

3.3 Online Steering for Policy Deployment Alignment

System-level postprocessing is commonly applied to ensure smooth robot motions when deploying chunking policies, such as temporal ensembling and receding horizon control [14, 59]. However, these optimization tricks cause a mismatch between training and deployment: what the policy has seen during training (raw actions) is different from the ones actually being executed during deployment (optimized actions). In the context of dexterous and precise manipulation, the mismatch becomes non-negligible. To adapt to the discrepancy, we find it crucial for the model to explore and improve itself via closed-loop online interactions with *aligned* actions.

Performing online RL in long-horizon, precise manipulation tasks remains non-trivial, especially in exploration. Since the task requires millimeter precision to complete, adding noise to wrist poses or joint positions hardly leads to success. We instead perform structured exploration in a latent space and steer the trained flow policy [54]. Specifically, we add a *noise predictor* $\pi_{\theta'}$ after the shared VLM backbone to predict the initial noise ϵ_t for the action DiT. The number of trainable parameters in $\pi_{\theta'}$ is 51.5M. To avoid generating arbitrary actions from noise out of the offline training distribution, we penalize the noise predictor when its output diverges from the original normal distribution beyond a certain threshold β. Following [54], we also distill a Q function over noise space $Q_{\phi'}(\mathbf{o}_t, l, \mathbf{s}_t, \epsilon_t)$ to avoid back-propagating through the flow model during policy optimization. The critic in the original action space $Q_\phi(\mathbf{o}_t, l, \mathbf{s}_t, \mathbf{a}_t)$ is trained via standard TD3. The online training objectives for the noise transformer and the critic in noise space are as follows:

$$\mathcal{L}(\pi_{\theta'}) = \mathbb{E}_{(\mathbf{o}_t, l, \mathbf{s}_t) \sim \mathcal{D}} \left[-Q_{\phi'}(\mathbf{o}_t, l, \mathbf{s}_t, \epsilon_t) + c \max(\frac{1}{2} \|\epsilon_t\|^2 - \beta, 0) \right], \epsilon_t \sim \pi_{\theta'}(\mathbf{o}_t, l, \mathbf{s}_t), \tag{3}$$

$$\mathcal{L}(Q_{\phi'}) = \texttt{cross\_entropy}\left(Q_{\phi'}(\mathbf{o}_t, l, \mathbf{s}_t, \epsilon_t), Q_\phi(\mathbf{o}_t, l, \mathbf{s}_t, \pi_\theta(\mathbf{o}_t, l, \mathbf{s}_t|\epsilon_t))\right), \epsilon_t \sim \begin{cases} \mathcal{N}(\mathbf{0}, \mathbf{1}) & w.p.\ 0.5, \\ \pi_{\theta'}(\mathbf{o}_t, l, \mathbf{s}_t) & \text{otherwise.} \end{cases} \tag{4}$$

Different from the original implementation [54], to ensure a good coverage on the noise space when distilling $Q_{\phi'}$, we sample the input noise from the original normal distribution with 0.5 probability (*w.p.* 0.5), and from the noise predictor otherwise.

For sample-efficient offline to online adaptation, we maintain an off-policy buffer and an on-policy buffer, and sample batches from them evenly. Before training starts, we warm up the off-policy buffer with online rollouts of offline-trained checkpoints, similar to Warm-start RL [60]. We intentionally choose not to mix teleoperated trajectories into the buffer to prevent the policy from training on mismatched dynamics. The on-policy buffer only stores trajectories generated from the two most recent checkpoints, and the stale data is pushed into the off-policy buffer.

4 Robot & System

The robot we utilize to verify GR-RL is **ByteMini-v2**, as illustrated in Fig. 4. ByteMini-v2 is a wheeled mobile manipulation robot equipped with 7-DoF dual robotic arms, featuring a unique wrist spherical joint design [53]. The operational dexterity, stability, consistency, and usability of ByteMini-v1 robots have been fully validated in the work of GR-3 [12]. Meanwhile, drawing on the limitations identified during ByteMini-v1 usage in GR-3, we have implemented the following three key design optimizations on ByteMini-v2 robots.

Higher Load for Manipulation The maximum load ability for the 7-DoF arm is limited by the maximum output torque of the elbow actuator. We update the elbow actuator from peak output torque 17 Nm to 35 Nm. As a result, the peak load of 7-DoF arms on ByteMini-v2 robots is increased from 1.4 kg to 3.15 kg, which will significantly expand the robot's operational capabilities.
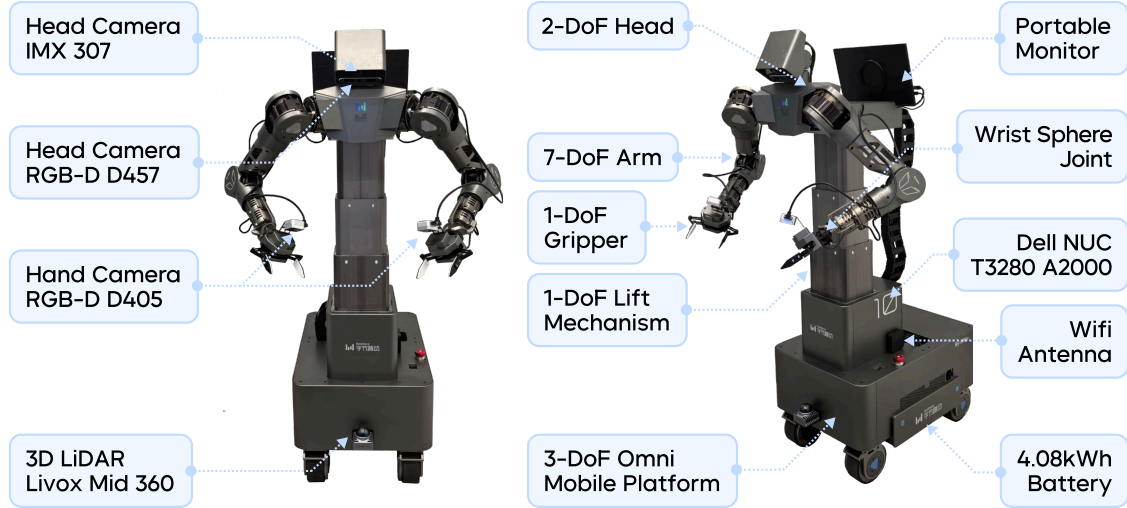
Figure 4 The ByteMini-v2 Robot. We show the robot specifications in terms of sensors, DoFs and electronic devices.

Enhanced Mobility in Confined Spaces The projected area of the ByteMini-v1 chassis is $500\,\text{mm}$ x $720\,\text{mm}$. And now for improving the mobility in confined spaces, the projected area of ByteMini-v2 robot is reduced to $450\,\text{mm}$ x $650\,\text{mm}$ respectively. Meanwhile, the design of the servo steering wheels on the mobile platform has also been optimized accordingly, enabling the synchronous adjustment of the steering wheels' motion in both yaw and pitch directions [56] and thereby enhancing the robot's rapid direction-changing capability in confined spaces.

Higher Polished Robot Design and Enhanced Usability As shown in Fig. 4, the ByteMini-v2 robot has been designed with enhanced refinement, featuring an ID and proper encapsulation of exposed electrical wiring harnesses. The position of the portable monitor has been adjusted from the robots' chassis to the shoulder, in order to achieve a better user experience.

5 Experiments

Task Description We present GR-RL in a shoe lacing task, a challenging scenario featuring long-horizon, dexterous and precise manipulation. The observations are composed of three views of RGB images, proprioception states, and language instructions. In model inference, we incorporate a trajectory optimization module that imposes constraints on jerk and temporal continuity to refine the predicted action chunks. During RL training, we adopt a binary sparse reward setting, meaning that a positive reward of 1 is obtained only when the shoelace is threaded through the correct eyelet and put down on the table completely.

Main Results We report the effect of our multi-stage training recipe in Fig. 5. Our base model GR-3, trained with behavior cloning over all the human teleoperated data, achieves a success rate of 45.7%. After filtering the data with our learned task progress model, the success rate is boosted to 61.6%. It highlights the importance of the data cleaning mechanism when learning precise and long-horizon manipulation. With symmetry data augmentation, the success rate of a filtered behavior cloning policy is further increased to 72.7%.

The "Filtered BC + Aug." model serves as the starting point of online steering RL. To better align the offline-trained critic to the distribution of model rollout, we collect 673 trajectories generated by the offline model, then freeze the VLM backbone and continue to train the critic heads Q_ϕ, $Q_{\phi'}$ using these data. These rollout trajectories are also populated into the off-policy replay buffer in online RL to stabilize training. We then tune both the noise predictor $\pi_{\theta'}$ and the two critic heads for 50 optimization steps once we gather 12 new episodes online. The moving average of the success rate over a window size of 24 is shown as the curve on the right side of Fig. 5, and the binary success signals for all online episodes are represented as dots. In the first few iterations, the success rate decreases due to the distribution shift from offline to online RL. Later,
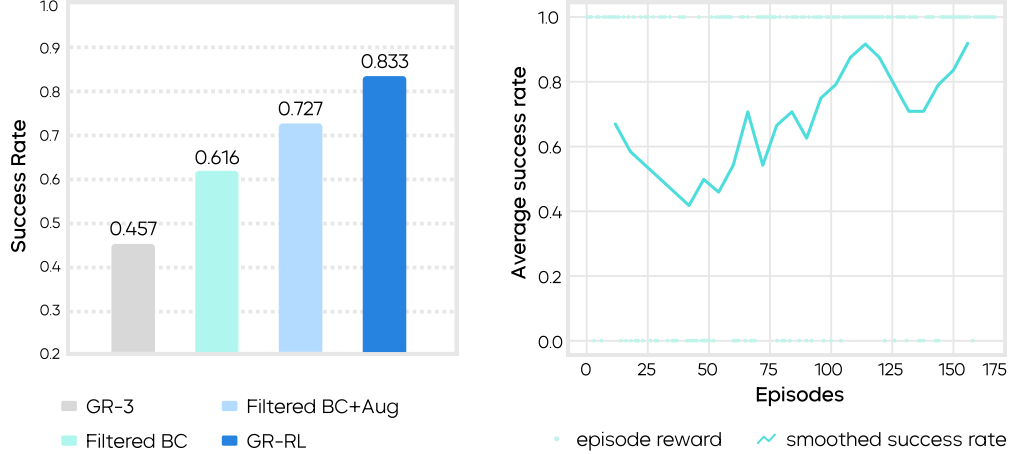
Figure 5 *Left*: the success rate of our multi-stage training recipe. Data filtering, mirror augmentation, and online tuning all contribute to the final performance. *Right*: the binary success signal per episode (dots) and the moving average of success rate (curve) during online finetuning. The performance increases rapidly after an offline-to-online adaptation phase.

the success rate rapidly recovers and grows beyond the offline performance to over 90%. The checkpoint at 500 online training steps is used for evaluation, and achieves a success rate of 83.3%.

To better understand the failure modes for different models in this long-horizon task, we evaluate whether the models succeed at several critical stages, including picking up the correct shoelace, threading the shoelace into the correct eyelet, handing it over to another gripper, and pulling the shoelace tight. The detailed success rate is illustrated in Fig. 6. The colored area denotes the success rate for completing each stage. The hatched area represents the decrease in the success rate compared to the previous stage. Data filtering and online RL can largely reduce the failure during threading. Data augmentation improves the model performance in all stages, although with a smaller magnitude.

Ablation on the Progress Evaluator We validate the effectiveness of the RL-based progress evaluator by comparing it with a regression variant. The regression baseline is trained by directly regressing the temporal progress $\frac{t}{T}$ of each state $\mathbf{s}_t$ in successful trajectories. As shown in Fig. 3, the regression-based predictor tends to overly smooth the progress prediction. It makes reasonable predictions in normal cases but is less sensitive to subtle failure (oftentimes only millimeters from success), such as failing to pull out the shoelace or imprecise insertion. Also, the regression-based predictor is not good at capturing transitions with long-term effects. There is a significant value jump in the predictions of GR-RL when the robot intentionally puts down the shoelace to adjust the grasping pose, but the regression-based prediction is almost flat during the adjustment.

We also train a non-distributional critic with the same TD3+BC algorithm to compare with our distributional critic. The models are evaluated over a high-quality successful trajectory, as illustrated in Fig. 7. Due to the long horizon and binary sparse reward in our setting, the non-distributional critic suffers from severe over-estimation, especially in earlier parts of trajectories where the reward supervision signal is weak. The value prediction of our distributional critic falls in a predefined range, thus converging to a reasonable scale more robustly and demonstrating better alignment with the true temporal order.

Visualization of the Learned Behaviors GR-RL demonstrates robust behaviors in various cases. It can handle shoes with different colors and sizes, as shown in Fig. 8. The model automatically retries when the shoelace accidentally drops down (Fig. 8b) or when the shoelace misses the eyelet (Fig. 8c). The model can actively manipulate the scene to make the task easier to solve. In Fig. 8d, the initial grasping point is far from the tip of the shoelace. The model decides to drop the shoelace on top of the deformable shoe and regrasp closer to the tip before threading. In Fig. 8e, the model first reorients the shoe from the left side to straighten it, then starts threading. Similarly, for a shoe initially positioned on the far side of the table (Fig. 8f), the robot can pull it near, adjust the position of the shoelace, and then complete the task. In cases where the two ends of
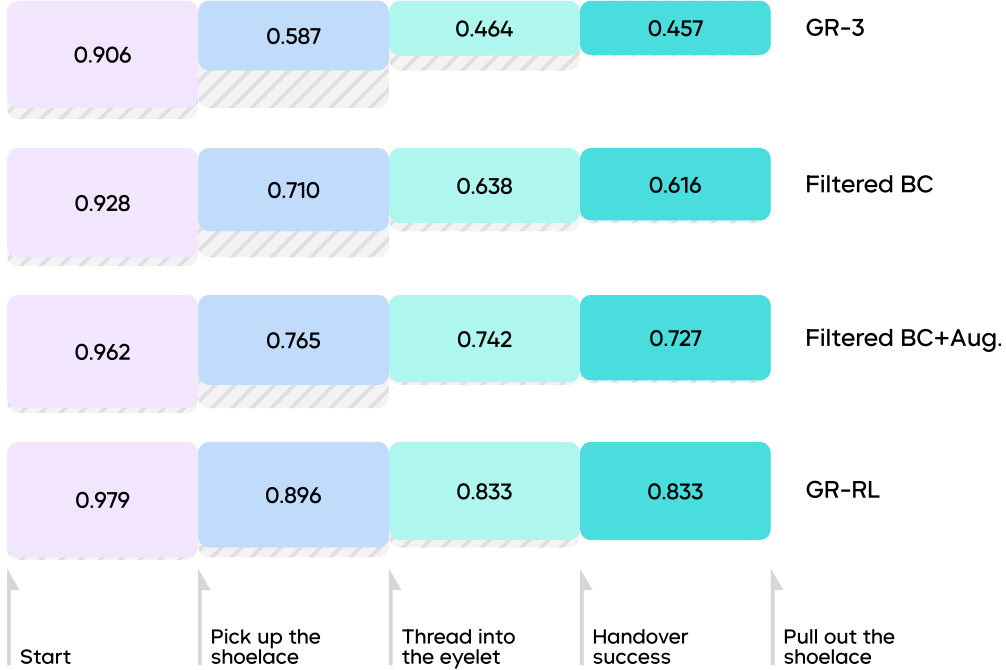
Figure 6 Detailed success rates of different models for completing intermediate stages. The height of each hatched area denotes the decrease in success rate from the previous stage to the current stage.
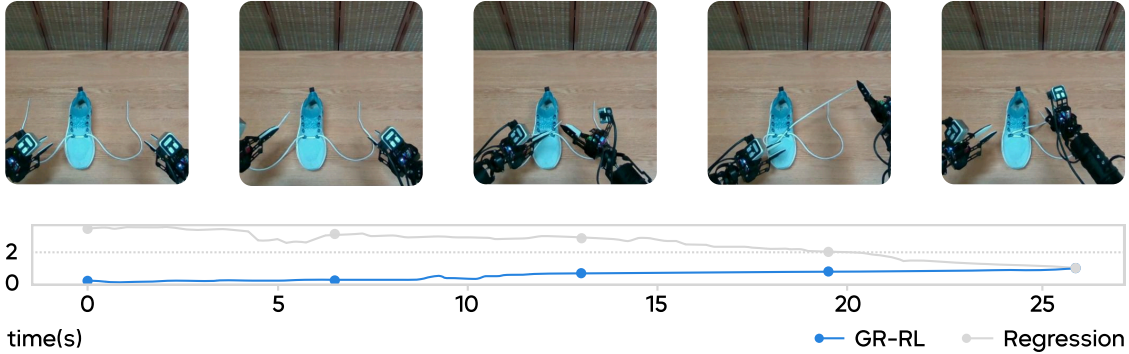

Figure 7 Comparison of progress prediction by distributional vs. non-distributional critics. Non-distributional critics are unbounded in the output range, and fail to reflect the positive progress in a successful trajectory.

the shoelace are crossed, and the end that should be grasped is underneath (Fig. 8g), the model can identify the correct one and pull it out.

6 Related Work

Generalist Robotic Foundation Policy Building generalist robotic foundation manipulation policies is a long-standing challenge for robotics research and applications [4, 6–10, 25, 27, 31, 32, 45, 50, 52]. Recent advances in building vision-language-action (VLA) models advocate adapting the vision-language models (VLMs) pretrained with web-scale data to robotic actions by adding the action modality [8, 15, 25, 27, 29, 36, 42–44, 52, 55]. The core idea is to leverage large-scale real-world robot trajectories collected by human teleoperators and generalize to potentially novel scenes and tasks. GR-RL sits upon the prior success of GR-3 [12], a generalist policy co-trained with web-scale data and human demonstrations. GR-RL further improves GR-3 by filtering high-quality data, augmenting actions, and online real-world RL, enabling it to perform long-horizon
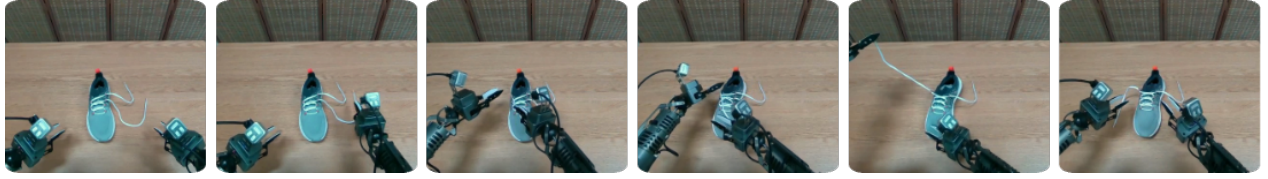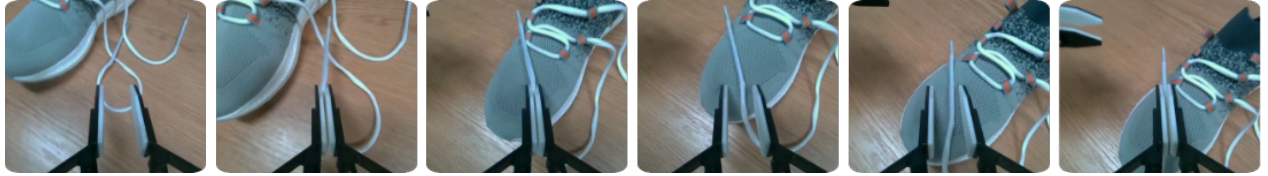
(a) Thread shoelace for a shoe with a different color.
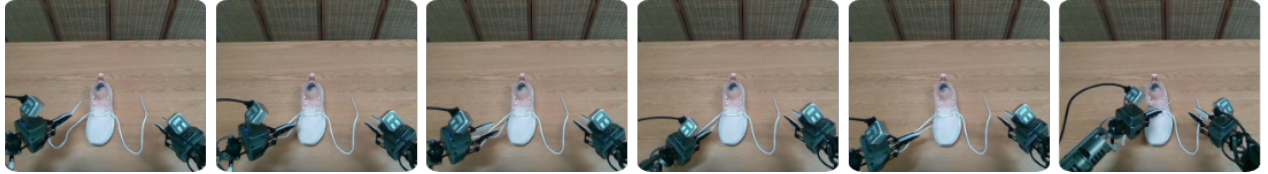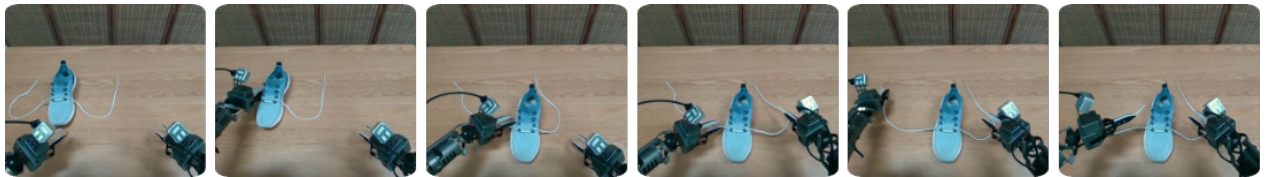

(b) Regrasp the shoelace when it drops.
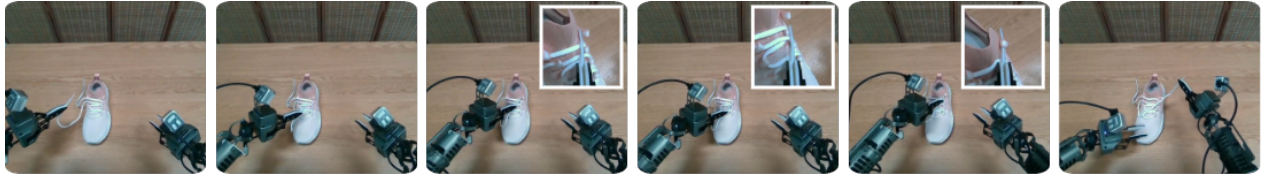
(c) Re-attempt when the shoelace is not threaded precisely through the eyelet.
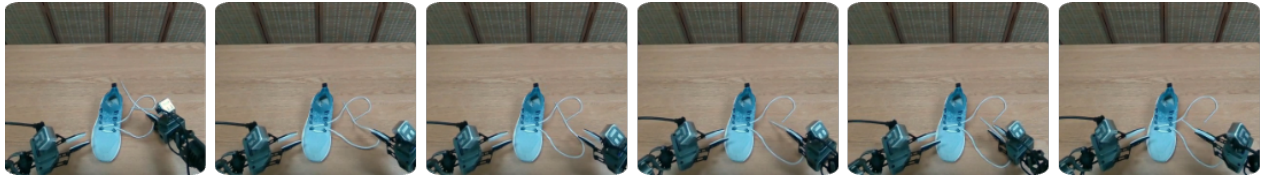

(d) Adjust the grasp pose intentionally on the surface of the shoe.

(e) Reorient the shoe before threading.

(f) Adjust the positions of the shoe and the shoelace before threading.

(g) Pull out the shoelace underneath another one from the crossed laces.

Figure 8 Robust behavior of GR-RL in various cases.

dexterous and precise manipulation.

Real-World Reinforcement Learning A central limitation of pure imitation learning is its susceptibility to compounding errors and its inability to exceed the performance of demonstrations. To address these issues, a significant body of work explores online data collection and real-world reinforcement learning (RL) to improve manipulation robustness beyond supervised training [2, 26, 28, 30, 38–40, 48, 49]. In the context of VLAs, some recent works aim to perform policy improvement via on-policy RL [13, 34, 37, 46, 58] in simulation. However, transferring their success to real-world scenarios remains difficult because real-world interactions are sample inefficient and noisy. Another line of work learns world models and performs on-policy RL interacting with the learned world model [18–22, 41, 57, 61]. World models alleviate the issue for real-robot interaction, but introduce further issues given inaccurate visual predictions. GR-RL follows [38, 39, 48] and focuses on real-world off-policy RL. This allows us to efficiently utilize past trajectories and improve the sample efficiency. Treating the noisy real-world reward as a distribution, we show that distributional critics significantly improve the robustness compared with standard regression-based critic models. Concurrent to our work, $\pi_{0.6}^*$ presents a real-world RL pipeline for high-precision manipulation [24]. Similar to $\pi_{0.6}^*$, we both adopt distributional critics that learn the progress of the task. However, instead of performing advantage-conditioned denoising, we directly perform filtered behavior cloning and also observe a strong performance boost. Given the stronger base offline policy, it helps reduce the search space during online exploration. We hope GR-RL reveals certain insights to the community on building capable specialist agents from generalist policies and helps to push the boundaries of deployable robotics research.

7   Limitations & Conclusions

Limitations Despite the strong performance of GR-RL in long-horizon high-precision dexterous tasks, it still has clear limitations. One of the major issues of our current pipeline is the behavior-drifting problem. Given a sparse and noisy reward, our policy behavior could be unstable during online RL. This is possibly due to the limited capacity of the lightweight noise predictor, or the challenging credit assignment issue in the large latent action space. Furthermore, distilling the improved policy into the base VLA could be a potential direction for obtaining both capable and general manipulation policies. We leave these issues for future study.

Conclusion We introduce GR-RL, a robotic learning framework for building specialist VLA policies for long-horizon dexterous and precise manipulation. The key insight of GR-RL is that the mismatch between data collection and policy inference needs online alignment. GR-RL learns an RL-based evaluator by treating the critic value learned from sparse reward as task progress prediction, and uses it to filter high-quality transitions to train a robust base policy. During this process, we also introduce a simple yet effective morphological symmetry augmentation method to improve the overall performance. Last, we perform online RL that aligns the rollout behavior with our training signals. To the best of our knowledge, GR-RL is the first learning-based policy capable of lacing up shoes. We hope GR-RL can be a small step towards capable real-world specialist robot policies.

References

[1] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 30, 2017.

[2] Lars Ankile, Zhenyu Jiang, Rocky Duan, Guanya Shi, Pieter Abbeel, and Anusha Nagabandi. Residual off-policy rl for finetuning behavior cloning policies. arXiv preprint arXiv:2509.19301, 2025.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923, 2025.

[4] Jose Barreiros, Andrew Beaulieu, Aditya Bhat, Rick Cory, Eric Cousineau, Hongkai Dai, Ching-Hsin Fang, Kunimatsu Hashimoto, Muhammad Zubair Irshad, Masha Itkina, et al. A careful examination of large behavior models for multitask dexterous manipulation. arXiv preprint arXiv:2507.05331, 2025.

[5] Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In International conference on machine learning, pages 449–458. PMLR, 2017.

[6] 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. In 2024 IEEE International Conference on Robotics and Automation (ICRA), pages 4788–4795. IEEE, 2024.

[7] 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.

[8] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π0: A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164, 2024.

[9] 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.

[10] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. arXiv preprint arXiv:2307.15818, 2023.

[11] Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, et al. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. arXiv preprint arXiv:2410.06158, 2024.

[12] Chilam Cheang, Sijin Chen, Zhongren Cui, Yingdong Hu, Liqun Huang, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Xiao Ma, et al. Gr-3 technical report. arXiv preprint arXiv:2507.15493, 2025.

[13] Kang Chen, Zhihao Liu, Tonghe Zhang, Zhen Guo, Si Xu, Hao Lin, Hongzhi Zang, Quanlu Zhang, Zhaofei Yu, Guoliang Fan, et al. Online rl fine-tuning for flow-based vision-language-action models. arXiv preprint arXiv:2510.25889, 2025.

[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] Ria Doshi, Homer Walke, Oier Mees, Sudeep Dasari, and Sergey Levine. Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation. arXiv preprint arXiv:2408.11812, 2024.

[16] Jesse Farebrother, Jordi Orbay, Quan Vuong, Adrien Ali Taïga, Yevgen Chebotar, Ted Xiao, Alex Irpan, Sergey Levine, Pablo Samuel Castro, Aleksandra Faust, et al. Stop regressing: Training value functions via classification for scalable deep rl. arXiv preprint arXiv:2403.03950, 2024.

[17] Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. Advances in neural information processing systems, 34:20132–20145, 2021.

[18] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. arXiv preprint arXiv:1912.01603, 2019.

[19] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. arXiv preprint arXiv:2010.02193, 2020.

[20] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. arXiv preprint arXiv:2301.04104, 2023.

[21] Nicklas Hansen, Xiaolong Wang, and Hao Su. Temporal difference learning for model predictive control. arXiv preprint arXiv:2203.04955, 2022.

[22] Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. arXiv preprint arXiv:2310.16828, 2023.

[23] Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.

[24] Physical Intelligence, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Kevin Black, Ken Conley, Grace Connors, James Darpinian, Karan Dhabalia, Jared DiCarlo, Danny Driess, Michael Equi, Adnan Esmail, Yunhao Fang, Chelsea Finn, Catherine Glossop, Thomas Godden, Ivan Goryachev, Lachy Groom, Hunter Hancock, Karol Hausman, Gashon Hussein, Brian Ichter, Szymon Jakubczak, Rowan Jen, Tim Jones, Ben Katz, Liyiming Ke, Chandra Kuchi, Marinda Lamb, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Yao Lu, Vishnu Mano, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Charvi Sharma, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, Will Stoeckle, Alex Swerdlow, James Tanner, Marcel Torne, Quan Vuong, Anna Walling, Haohuan Wang, Blake Williams, Sukwon Yoo, Lili Yu, Ury Zhilinsky, and Zhiyuan Zhou. $\pi^*0.6$: a vla that learns from experience. arXiv preprint arXiv:2511.14759, 2025.

[25] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi0.5$: a vision-language-action model with open-world generalization. arXiv preprint arXiv:2504.16054, 2025.

[26] 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.

[27] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246, 2024.

[28] 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.

[29] Qixiu Li, Yaobo Liang, Zeyu Wang, Lin Luo, Xi Chen, Mozheng Liao, Fangyun Wei, Yu Deng, Sicheng Xu, Yizhong Zhang, et al. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. arXiv preprint arXiv:2411.19650, 2024.

[30] Qiyang Li, Zhiyuan Zhou, and Sergey Levine. Reinforcement learning with action chunking. arXiv preprint arXiv:2507.07969, 2025.

[31] Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, et al. Vision-language foundation models as effective robot imitators. arXiv preprint arXiv:2311.01378, 2023.

[32] Xinghang Li, Peiyan Li, Minghuan Liu, Dong Wang, Jirong Liu, Bingyi Kang, Xiao Ma, Tao Kong, Hanbo Zhang, and Huaping Liu. Towards generalist robot policies: What matters in building vision-language-action models. arXiv preprint arXiv:2412.14058, 2024.

[33] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. arXiv preprint arXiv:2210.02747, 2022.

[34] Jijia Liu, Feng Gao, Bingwen Wei, Xinlei Chen, Qingmin Liao, Yi Wu, Chao Yu, and Yu Wang. What can rl bring to vla generalization? an empirical study. arXiv preprint arXiv:2505.19789, 2025.

[35] Qiang Liu. Rectified flow: A marginal preserving approach to optimal transport. arXiv preprint arXiv:2209.14577, 2022.

[36] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. arXiv preprint arXiv:2410.07864, 2024.

[37] Guanxing Lu, Wenkai Guo, Chubin Zhang, Yuheng Zhou, Haonan Jiang, Zifeng Gao, Yansong Tang, and Ziwei Wang. Vla-rl: Towards masterful and general robotic manipulation with scalable reinforcement learning. arXiv preprint arXiv:2505.18719, 2025.

[38] Jianlan Luo, Zheyuan Hu, Charles Xu, You Liang Tan, Jacob Berg, Archit Sharma, Stefan Schaal, Chelsea Finn, Abhishek Gupta, and Sergey Levine. Serl: A software suite for sample-efficient robotic reinforcement learning. In 2024 IEEE International Conference on Robotics and Automation (ICRA), pages 16961–16969. IEEE, 2024.

[39] Jianlan Luo, Charles Xu, Jeffrey Wu, and Sergey Levine. Precise and dexterous robotic manipulation via human-in-the-loop reinforcement learning. arXiv preprint arXiv:2410.21845, 2024.

[40] Lei Lv, Yunfei Li, Yu Luo, Fuchun Sun, Tao Kong, Jiafeng Xu, and Xiao Ma. Flow-based policy for online reinforcement learning. arXiv preprint arXiv:2506.12811, 2025.

[41] Xiao Ma, Siwei Chen, David Hsu, and Wee Sun Lee. Contrastive variational reinforcement learning for complex observations. In Conference on robot learning, pages 959–972. PMLR, 2021.

[42] 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.

[43] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. arXiv preprint arXiv:2501.09747, 2025.

[44] Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. arXiv preprint arXiv:2501.15830, 2025.

[45] Moritz Reuss, Ömer Erdinç Yağmurlu, Fabian Wenzel, and Rudolf Lioutikov. Multimodal diffusion transformer: Learning versatile behavior from multimodal goals. arXiv preprint arXiv:2407.05996, 2024.

[46] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017.

[47] Younggyo Seo and Pieter Abbeel. Coarse-to-fine q-network with action sequence for data-efficient robot learning. arXiv preprint arXiv:2411.12155, 2024.

[48] Younggyo Seo, Jafar Uruç, and Stephen James. Continuous control with coarse-to-fine reinforcement learning. arXiv preprint arXiv:2407.07787, 2024.

[49] Archit Sharma, Ahmed M Ahmed, Rehaan Ahmad, and Chelsea Finn. Self-improving robots: End-to-end autonomous visuomotor reinforcement learning. arXiv preprint arXiv:2303.01488, 2023.

[50] Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, et al. Smolvla: A vision-language-action model for affordable and efficient robotics. arXiv preprint arXiv:2506.01844, 2025.

[51] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. arXiv preprint arXiv:2503.20020, 2025.

[52] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. arXiv preprint arXiv:2405.12213, 2024.

[53] Jiawen Tian, Liqun Huang, Zhongren Cui, Jingchao Qiao, Jiafeng Xu, Xiao Ma, and Zeyu Ren. Bytewrist: A parallel robotic wrist enabling flexible and anthropomorphic motion for confined spaces. arXiv preprint arXiv:2509.18084, 2025.

[54] Andrew Wagenmaker, Mitsuhiko Nakamoto, Yunchu Zhang, Seohong Park, Waleed Yagoub, Anusha Nagabandi, Abhishek Gupta, and Sergey Levine. Steering your diffusion policy with latent space reinforcement learning. arXiv preprint arXiv:2506.15799, 2025.

[55] Lirui Wang, Xinlei Chen, Jialiang Zhao, and Kaiming He. Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers. Advances in neural information processing systems, 37:124420–124450, 2024.

[56] 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, 47(8):1087–1102, 2023.

[57] Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In Conference on robot learning, pages 2226–2240. PMLR, 2023.

[58] Wenke Xia, Yichu Yang, Hongtao Wu, Xiao Ma, Tao Kong, and Di Hu. Robotic policy learning via human-assisted action preference optimization. arXiv preprint arXiv:2506.07127, 2025.

[59] 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.

[60] Zhiyuan Zhou, Andy Peng, Qiyang Li, Sergey Levine, and Aviral Kumar. Efficient online reinforcement learning fine-tuning need not retain offline data. arXiv preprint arXiv:2412.07762, 2024.

[61] Fangqi Zhu, Zhengyang Yan, Zicong Hong, Quanxin Shou, Xiao Ma, and Song Guo. Wmpo: World model-based policy optimization for vision-language-action models. arXiv preprint arXiv:2511.09515, 2025.

Contributions and Acknowledgments

Authors are listed in alphabetical order.

Core Contributors Yunfei Li, Xiao Ma, Jiafeng Xu

Contributors Yu Cui, Zhongren Cui, Zhigang Han, Liqun Huang, Tao Kong, Yuxiao Liu, Hao Niu, Wanli Peng, Jingchao Qiao, Zeyu Ren, Haixin Shi, Zhi Su, Jiawen Tian, Yuyang Xiao, Shenyu Zhang, Liwei Zheng

Supervisors Hang Li, Yonghui Wu