Whole-Body Mobile Manipulation using Offline Reinforcement Learning on Sub-optimal Controllers

Anonymous Authors

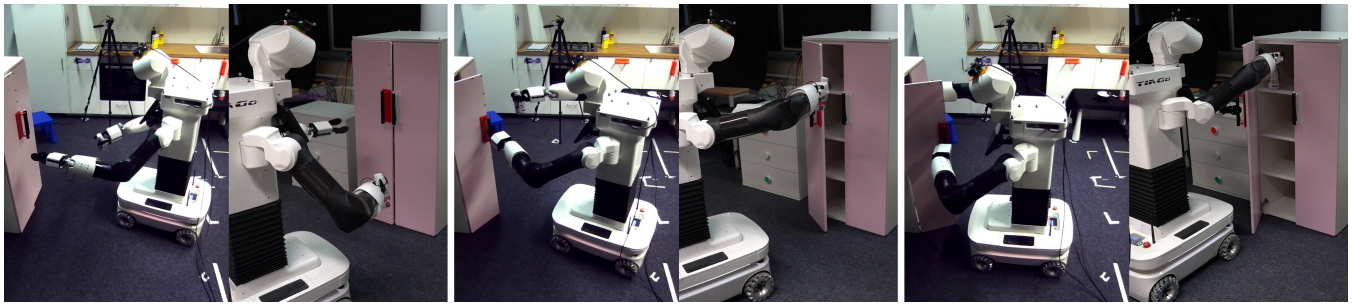

Fig. 1: WHOLE-MoMa policy on a real TIAGo++ mobile manipulator simultaneously opening a cupboard and placing an object inside it.

***Abstract*— Whole-body Mobile Manipulation (MoMa) of articulated objects – e.g., opening doors, drawers, and cupboards – demands simultaneous coordination of a robot's base and arms. Classical Whole-Body Controllers (WBCs) solve this via hierarchical optimization but require extensive tuning and remain brittle, while learning-based methods rely on expensive whole-body teleoperation data or heavy reward engineering. We observe that even a sub-optimal WBC is a powerful structural prior: it collects data in a constrained, task-relevant region of the state-action space, and its behavior can still be improved using offline RL. We propose WHOLE-MoMa, a two-stage pipeline that first generates diverse demonstrations by randomizing a lightweight WBC, and then applies offline RL to identify and stitch together improved behaviors via a reward signal. To support expressive action-chunked Diffusion Policies, we extend offline IQL with Q-chunking for chunk-level critic evaluation and advantage-weighted policy extraction. On three tasks of increasing difficulty with a TIAGo++ mobile manipulator, WHOLE-MoMa outperforms WBC, behavior cloning, and several offline RL baselines, and transfers directly to the real robot without teleoperated or real-world training data, achieving 80% success on bimanual drawer manipulation and 68% on simultaneous cupboard opening and object placement.**

I. Introduction

Mobile Manipulation (MoMa) robots are central to the vision of general-purpose home assistants, given their enhanced workspace and ability to operate in everyday environments [1], [2]. A key challenge of MoMa is coordinating the different embodiments of the robot, namely the manipulator arms and the mobile base. While many methods target mobile "pick & place" tasks [3], [4], tasks that require *simultaneous* coordination of the base and arms – such as articulating doors, drawers, and cupboards – remain under-explored. Humans perform such whole-body coordination naturally, yet few learning methods specifically target this setting, and an often overlooked aspect is performing the task in a *time-efficient, simultaneous* manner, e.g., opening a cupboard door while already placing an object inside (Figure 1).

Existing paradigms each fall short. Learning-based methods [5], [6] either rely on expensive whole-body teleoperation data [7], [8] or suffer from the curse of dimensionality in the expanded whole-body state-action space [9], [10]. Teleoperation in particular is inherently more complex in MoMa: it requires specialized equipment to simultaneously move the base and arms, and is expensive in human effort [11]. Conversely, Whole-Body Controllers (WBC) [12], [13] and MPC-based methods [14], [15] coordinate multiple embodiments via hierarchical optimization, but rely on hand-crafted cost functions and extensive tuning, and cannot plan *through* an interaction: a configuration well-suited for reaching a handle may be entirely wrong for executing the subsequent articulation.

Our key insight is that even a simple, sub-optimal WBC acts as a strong structural *prior* over the solution space: it dramatically reduces the search space compared to random exploration and focuses data collection on a physically feasible, task-relevant region, enabling sample-efficient offline RL. Rather than requiring a perfectly tuned optimizer or expensive teleoperated data, we use a lightweight WBC with randomized parameters to generate diverse demonstrations, and then use offline RL as a mechanism to *learn from these sub-optimal demonstrations*, identifying and stitching together the best behaviors via a reward signal. To support expressive, action-chunked Diffusion Policies needed for complex whole-body coordination, we adapt offline RL with Q-chunking, enabling IQL-based critics to evaluate action chunks directly. In summary, our contributions are:

- **WHOLE-MoMa**, a simple and scalable data generation pipeline using a multi-objective hierarchical WBC to produce structured whole-body demonstrations, without any teleoperation.
- An offline-RL formulation that improves upon the sub-optimal WBC demonstrations, stitching together better behaviors from a reward signal, without any optimal teleoperated demonstrations.

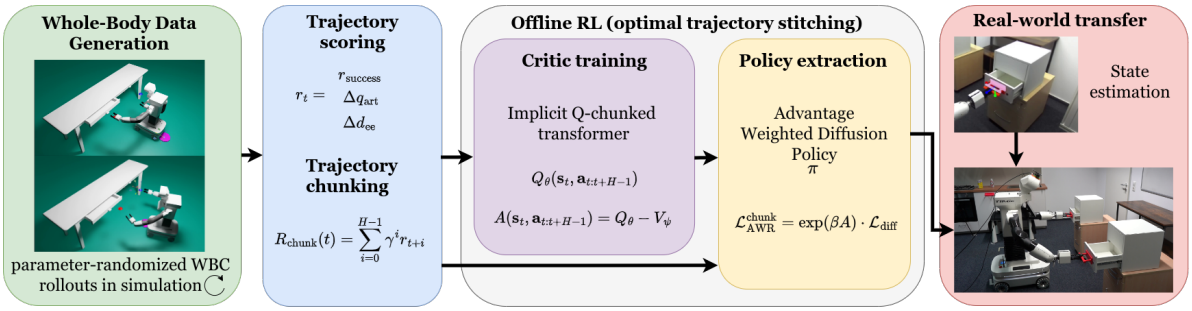

Fig. 2: **WHOLE-MoMa** pipeline. Parameter-randomized WBC rollouts produce whole-body demonstrations scored by a reward combining task success, articulation progress (Δq_{art}), and end-effector or base distance reduction ($\Delta d_{\text{ee}}, \Delta d_{\text{base}}$). Transitions are grouped into horizon-H chunks. A chunked transformer critic $Q_\theta(\mathbf{s}_t, \mathbf{a}_{t:t+H-1})$ is learned via chunked TD error and used to compute chunk advantages $A(\mathbf{s}_t, \mathbf{a}_{t:t+H-1})$. A Diffusion Policy π_ϕ is trained with the chunk-level AWR objective $\mathcal{L}_{\text{AWR}}^{\text{chunk}}$, and deployed directly on the real robot.

- An adaptation of offline RL to support expressive, action-chunked policy classes via Q-chunking, enabling IQL with Diffusion Policies for temporally consistent action-sequence prediction.
- Direct sim-to-real transfer of learned whole-body policies to a real TIAGo++ on cupboard-open-and-place and bimanual drawer manipulation tasks, without any real-world fine-tuning.

In simulation, WHOLE-MoMa achieves 98%, 80%, and 78% success on the door, drawer, and cupboard tasks respectively, outperforming WBC, imitation learning, and several offline RL baselines. On the real robot, it achieves 80% success on bimanual drawer manipulation and 68% on the hardest simultaneous cupboard-open-and-place task without any real-world fine-tuning or teleoperated data.

II. WHOLE-MoMa

WHOLE-MoMa is a two-stage pipeline (Figure 2). First, a Whole-Body Controller (WBC) generates a large dataset of structured but sub-optimal demonstrations. Second, offline RL learns improved joint-velocity policies from this dataset, without any teleoperation. The WBC serves as a structural *prior* over the solution space: by constraining trajectories to physically feasible, task-relevant behaviors, it dramatically reduces the effective search space and focuses learning on a region where a reward signal can meaningfully discriminate better from worse behaviors.

WBC data generation. We use a Hierarchical Quadratic Programming (HQP) formulation [13] via the TSID library [16] to simultaneously control all robot joints while satisfying multiple objectives at different hierarchy levels. We use the following priorities (highest to lowest): *hard constraints* – self-collision avoidance; joint position, velocity, and acceleration limits; *task objectives* – end-effector and base movement tracking; *regularization* – default robot pose tracking. Higher-priority objectives are optimized first and lower-priority ones are resolved in a Pareto-optimal fashion, while multiple objectives at the same level can be weighted to induce a soft hierarchy.

For each task we design a simple state machine that sequences through stages (e.g., reach, grasp, articulate, place), with the HQP solver generating motions at each stage. We collect 3k trajectories per task, randomizing per episode a set of parameters to diversify motion styles and timing: Gaussian joint-angle noise (σ=0.1 rad), pre-grasp threshold ($[0.01, 0.25]$ m), grasp threshold ($[0.01, 0.1]$ m), articulation step size ($[0.005, 0.25]$ m), and EE / base / posture weights ($[0.1, 5.0]$, $[0.1, 5.0]$, $[0.0, 1.0]$). This amount of data is sufficient because the WBC prior already focuses sampling on a structured, task-relevant region of the state-action space; a major advantage over purely random data collection. The WBC provides reasonable data for grasping and early stages, but cannot plan *through* the interaction: a configuration well-suited for reaching the handle may be poor for the subsequent articulation. This is exactly what offline RL aims to correct by stitching together trajectories that succeed end-to-end.

Offline RL with Q-chunking. Given the WBC-generated dataset $\mathcal{D}$ with reward labels, we train an offline RL policy to identify and stitch together the best behaviors. At each time t, the policy is conditioned on a state history $\mathbf{s}_t$ of robot proprioceptive state and articulated-object joint angles, and outputs a joint-velocity action chunk $\mathbf{a}_{t:t+H-1}$, following the Diffusion Policy formulation [5]. This preserves the full expressivity of whole-body coordination at test time: the policy can learn motion styles the WBC would not produce, and is not constrained by the WBC's myopic stage-by-stage execution. The reward is $r_t = \sum_i w_i r_{i,t}$, combining task success, articulation progress (Δq_{art}), and end-effector or base distance reduction; all weights w_i=1 and the dense terms are normalized by total articulation angle or target distance so each subgoal contributes a total return of 1.0, keeping the reward simple and free of hand-tuned shaping.

To make IQL [17] compatible with chunked Diffusion Policies, we relabel $\mathcal{D}$ into chunked transitions: for each t, we construct a sample $(\mathbf{s}_t, \mathbf{a}_{t:t+H-1}, R_{\text{chunk}}(t), \mathbf{s}_{t+H})$. The bootstrap target no longer uses the immediate next state, but the horizon-shifted state reached after executing the full chunk. The critic $Q_\theta(\mathbf{s}_t, \mathbf{a}_{t:t+H-1})$ and value $V_\psi(\mathbf{s}_t)$ are trained with the standard IQL expectile regression loss, but on chunks, preserving IQL's core property: both critic and value learning remain fully in-distribution, since targets are built only from chunks already present in $\mathcal{D}$. We use a transformer critic conditioned on the state history [18], enabling it to reason over recent context rather than a single timestep. For policy extraction, we use a chunk-level

TABLE I: Simulation results for WHOLE-MoMa and all baselines. Success rates over 50 episodes with 95% confidence intervals. Time to success averaged over successful trials.

Metric / Method	WBC	BC (Diff. Pol.)	RL (TD3)	IQL+DDPG_BC	IDQL	RISE	WHOLE-MoMa
Door task							
Success %	86 [73.8, 93.0]	78 [64.8, 87.2]	88 [76.2, 94.4]	86 [73.8, 93.0]	90 [78.6, 95.7]	92 [81.2, 96.8]	**98** [89.5, 99.6]
Time (s)	7.8	11.3	10.9	11.0	11.0	10.9	10.6
Drawer task							
Success %	68 [54.2, 79.2]	70 [56.2, 80.9]	44 [31.2, 57.7]	64 [50.1, 75.9]	72 [58.3, 82.5]	70 [56.2, 80.9]	**80** [67.0, 88.8]
Time (s)	14.4	15.0	19.3	20.5	18.7	18.5	17.4
Cupboard task							
Success %	52 [38.5, 65.2]	48 [34.8, 61.5]	0 [0.0, 7.1]	6 [2.1, 16.2]	64 [50.1, 75.9]	64 [50.1, 75.9]	**78** [64.8, 87.2]
Partial %	80 [67.0, 88.8]	78 [64.8, 87.2]	42 [29.4, 55.8]	54 [40.4, 67.0]	88 [76.2, 94.4]	90 [78.6, 95.7]	**100** [92.9, 100]
Time (s)	14.4	19.2	–	21.1	19.6	19.1	18.7

Door task | Drawer task | Cupboard task | RealDrawerOpenOneCloseAnother | RealCupboardOpenAndPlace

Fig. 3: Simulated and real whole-body tasks. Simulated, at increasing complexity: level 1 Door (push open and pass through), level 2 Drawer (close one and open another bimanually), level 3 Cupboard (open door while placing object inside). Real: RealDrawerOpenOneCloseAnother and RealCupboardOpenAndPlace on a TIAGo++ holonomic mobile manipulator.

Advantage-Weighted Regression (AWR) objective:

$$A(\mathbf{s}_t, \mathbf{a}_{t:t+H-1}) = Q_\theta(\mathbf{s}_t, \mathbf{a}_{t:t+H-1}) - V_\psi(\mathbf{s}_t), \quad (1)$$

$$\mathcal{L}_{\text{AWR}}^{\text{chunk}}(\phi) = \mathbb{E}\big[\exp\big(\beta A(\mathbf{s}_t, \mathbf{a}_{t:t+H-1})\big) \cdot \mathcal{L}_{\text{diff}}(\phi; \mathbf{s}_t, \mathbf{a}_{t:t+H-1})\big] \quad (2)$$

The critic identifies promising chunks, and the diffusion policy places more probability mass on those while remaining anchored to the demonstration distribution. We prefer AWR over alternatives (DDPG+BC [19], IDQL [20], RISE [21]) for stability in this high-dimensional action space: it does not require online action optimization or large candidate sets, and most WBC behaviors are already reasonable and only require precision improvements rather than major behavioral changes.

III. Experiments

Since whole-body MoMa of articulated objects with simultaneous base-arm coordination lacks an established benchmark, we design three tasks of increasing difficulty using a TIAGo++ holonomic mobile manipulator in Isaac Sim, with articulated objects from the GAPartNet dataset [22] (Figure 3). The level 1 **Door task** requires pushing open a door and navigating through without colliding. The level 2 **Drawer task** requires simultaneously closing one drawer and opening another, with handles spaced far apart so that simultaneous base translation and twist are needed for one arm to pull while the other pushes. The level 3 **Cupboard task** requires opening a cupboard door with one arm while *simultaneously* placing a held object inside with the other, demanding continuous base movement and sustained two-arm coordination throughout.

The state is $2 \times (7 \text{ arm} + 2 \text{ gripper})$ joints + 3 base coordinates $(x, y, \theta) = 21$ robot state dimensions, plus task-specific articulated joints, yielding a 22-dimensional state for Door and Cupboard and 23-dimensional for Drawer. Actions are 21-dim joint-velocity commands. Control runs at 40 Hz in both simulation and on the real robot. Episodes are capped at 600 steps for Door and 900 for Drawer and Cupboard (15 s and 22.5 s). We use EMA smoothing over diffusion-policy actions for stable deployment.

Baselines. We compare against: **WBC-policy** (the designed WBC without any learning); **BC** with a Diffusion Policy [5] trained on the WBC data; **RL (TD3)** [23], direct off-policy RL from scratch with the same reward function and an MLP Gaussian policy; and three offline-RL baselines that share our IQL critic but differ in policy extraction: **IQL+DDPG_BC** [19] (DDPG+BC extraction), **IDQL** [20] (diffusion policy + highest-Q sample at test time), and **RISE** [21] (IDQL with spectral-norm regularization). All learned baselines use the same Q-chunking transformer critic and transformer Diffusion Policy; only the RL baseline uses an MLP Gaussian policy.

Real-world setup. We transfer simulation-trained policies directly to a real TIAGo++ holonomic mobile manipulator on RealDrawerOpenOneCloseAnother and RealCupboardOpenAndPlace (Figure 4), without any teleoperated or real-world training data. Since TIAGo does not accept direct joint-velocity commands, predicted joint velocities are converted to positions via Euler integration and sent through the robot's position controller. Because Diffusion Policy inference ($\sim$80–100 ms per chunk of H=16) cannot sustain a 40 Hz control rate, we decouple inference from control asynchronously: a dedicated inference thread is triggered

(a) WBC: local optimum (b) WHOLE-MoMa: success

Fig. 4: Real-world RealDrawerOpenOneCloseAnother qualitative comparison: the WBC gets stuck in a local optimum and fails to fully open/close the drawers (left); WHOLE-MoMa performs the simultaneous close-and-open successfully (right).

early to predict the next action chunk while the control loop still consumes the last actions of the previous chunk; when a new chunk is ready, control switches over to the matching index. EMA smoothing across chunk boundaries prevents discontinuities. Articulated-object states are obtained via motion-capture markers, and safety handles are designed to snap off under excessive force, protecting the robot and the objects.

IV. Results

Simulation (Table I). On the easier **Door** task the WBC already reaches 86%. Offline-RL methods are tightly clustered, with WHOLE-MoMa achieving near-perfect 98%. The small gap reflects that the WBC data is already close to optimal here, so even BC (78%) captures most of the behavior. The **Drawer** task, requiring bimanual base-arm coordination, is harder: TD3 drops sharply to 44%, reflecting the difficulty of learning bimanual coordination from scratch; BC matches IDQL/RISE at 70–72%; and WHOLE-MoMa reaches 80%. The remaining 20% failure rate indicates that some WBC configurations lead to states where even reward-weighted stitching cannot recover optimal bimanual timing. On the hardest **Cupboard** task, TD3 fails entirely (0%) and IQL+DDPG_BC collapses at 6%, confirming that unstable policy extraction via critic gradients is particularly problematic with diffusion policies. IDQL/RISE reach 64%, and WHOLE-MoMa achieves 78% full success and 100% partial (grasping) success: the policy always grasps successfully, but the simultaneous articulation-and-placement phase remains challenging, where precise coordination between the two arms and the base is most critical. Across tasks, IDQL/RISE additionally suffer from sampling noise between consecutive actions – even N_a=256 candidates does not fully resolve this. WHOLE-MoMa's AWR extraction avoids this by directly reweighting the training data, yielding smoother, more consistent policies.

Real world (Table II). On **RealDrawerOpenOneCloseAnother** (Figure 4), the WBC gets stuck in local optima and fails to fully open or close the drawers (13/25), BC applies excessive lateral force and breaks the safety handle (15/25), while WHOLE-MoMa succeeds in 20/25 (80%) with near-perfect grasping (24/25) and high articulation success (20/24 given grasping). Notably, WHOLE-MoMa's real-world performance matches its simulation result, demonstrating effective state-based sim-to-real transfer for this task. On **RealCupboardOpenAndPlace**, the gap to simulation is larger: the WBC achieves only 4/25 (16%) as its myopic optimization leads to configurations where the handle snaps off under force; BC improves to 8/25; and WHOLE-MoMa reaches 17/25 (68%) with the highest grasping (22/25) and articulation (17/22) success. Remaining failures concentrate in the simultaneous articulation-and-placement phase and also reflect a sim-to-real dynamics gap (object compliance, snap-off safety handles) that could be mitigated by additional domain randomization of masses, frictions, and articulation-joint compliance during sim data generation.

TABLE II: Real-world evaluation over 25 trials per task. Articulation success is counted only when grasping succeeded.

Metric	WBC	BC (Diff. Pol.)	WHOLE-MoMa
RealDrawerOpenOneCloseAnother			
Success	13/25	15/25	**20/25**
Grasp	22/25	22/25	24/25
Articulate	13/22	15/22	20/24
Time (s)	24.5	34.4	31.1
RealCupboardOpenAndPlace			
Success	4/25	8/25	**17/25**
Grasp	17/25	19/25	22/25
Articulate	4/17	8/19	17/22
Time (s)	45.5	76.1	70.5

TABLE III: Simulation ablations comparing WHOLE-MoMa with architectural and design variants. Success rates over 50 episodes with 95% confidence intervals. Time to success averaged over successful trials.

Metric	Full model	U-Net Diff. Policy	MLP Q function	no Q-chunking
Door task				
Success %	**98** [89.5, 99.6]	86 [73.8, 93.0]	98 [89.5, 99.6]	90 [78.6, 95.7]
Time (s)	10.6	13.8	10.8	11.9
Drawer task				
Success %	**80** [67.0, 88.8]	42 [29.4, 55.8]	76 [62.6, 85.7]	60 [46.2, 72.4]
Time (s)	17.4	22.5	17.8	19.6
Cupboard task				
Success %	**78** [64.8, 87.2]	24 [14.3, 37.4]	72 [58.3, 82.5]	58 [44.2, 70.6]
Time (s)	18.7	24.4	18.9	20.7

Ablations (Table III). A transformer-based Diffusion Policy is critical on the hardest task: on Cupboard, the transformer reaches 78% vs. 24% for a U-Net, which struggles with the complex multi-stage coordination. A transformer critic offers a smaller gain over an MLP critic (78% vs. 72% on Cupboard); this is expected in IQL where the critic is only evaluated on in-distribution data, so even an MLP can fit it reasonably well. Q-chunking provides a clear benefit: removing it drops Cupboard success from 78% to 58%, confirming that chunk-level credit assignment and temporally consistent action-sequence prediction matter for these tasks. We found a state history of 5 and an action horizon of H=16 to work best. Overall, these results show that by treating a simple sub-optimal WBC as a structural prior and combining it with chunked offline RL, we obtain whole-body policies that outperform both pure-optimization and pure-imitation paradigms, and transfer directly to the real robot without any teleoperated or real-world training data.

REFERENCES

[1] O. Brock, J. Park, and M. Toussaint, "Mobility and manipulation," in *Springer Handbook of Robotics, 2nd Edition*, 2016.

[2] S. Yenamandra, A. Ramachandran, K. Yadav, A. S. Wang, M. Khanna, T. Gervet, T.-Y. Yang, V. Jain, A. Clegg, J. M. Turner *et al.*, "Homerobot: Open-vocabulary mobile manipulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 1975–2011.

[3] S. Jauhri, J. Peters, and G. Chalvatzaki, "Robot learning of mobile manipulation with reachability behavior priors," *IEEE Robotics and Automation Letters*, 2022.

[4] S. Uppal, A. Agarwal, H. Xiong, K. Shaw, and D. Pathak, "Spin: Simultaneous perception, interaction and navigation," *CVPR*, 2024.

[5] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, 2025.

[6] P. Arm, M. Mittal, H. Kolvenbach, and M. Hutter, "Pedipulate: Enabling manipulation skills using a quadruped robot's leg," in *IEEE Conference on Robotics and Automation (ICRA 2024)*, 2024.

[7] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," in *Conference on Robot Learning (CoRL)*, 2024.

[8] X. Xu, J. Park, H. Zhang, E. Cousineau, A. Bhat, J. Barreiros, D. Wang, and S. Song, "Hommi: Learning whole-body mobile manipulation from human demonstrations," 2026.

[9] D. Honerkamp, T. Welschehold, and A. Valada, "Learning kinematic feasibility for mobile manipulation through deep reinforcement learning," *IEEE Robotics and Automation Letters (RA-L)*, 2021.

[10] R. Yang, Y. Kim, A. Kembhavi, X. Wang, and K. Ehsani, "Harmonic mobile manipulation," *arXiv preprint arXiv:2312.06639*, 2023.

[11] S. B. Moyen, R. Krohn, S. Lueth, K. Pompetzki, J. Peters, V. Prasad, and G. Chalvatzaki, "The role of embodiment in intuitive whole-body teleoperation for mobile manipulation," in *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2025.

[12] L. Sentis and O. Khatib, "Synthesis of whole-body behaviors through hierarchical control of behavioral primitives," *International Journal of Humanoid Robotics*, 2005.

[13] A. Escande, N. Mansard, and P.-B. Wieber, "Hierarchical quadratic programming: Fast online humanoid-robot motion generation," *The International Journal of Robotics Research*, 2014.

[14] M. Mittal, D. Hoeller, F. Farshidian, M. Hutter, and A. Garg, "Articulated object interaction in unknown scenes with whole-body mobile manipulation," in *IEEE/RSJ international conference on intelligent robots and systems (IROS)*, 2022.

[15] J. Pankert and M. Hutter, "Perceptive model predictive control for continuous mobile manipulation," *IEEE Robotics and Automation Letters*, 2020.

[16] A. D. Prete, N. Mansard, O. E. Ramos, O. Stasse, and F. Nori, "Implementing torque control with high-ratio gear boxes and without joint-torque sensors," in *Int. Journal of Humanoid Robotics*, 2016.

[17] I. Kostrikov, A. Nair, and S. Levine, "Offline reinforcement learning with implicit q-learning," in *International Conference on Learning Representations*, 2022.

[18] D. Tian, O. Celik, and G. Neumann, "Chunking the critic: A transformer-based soft actor-critic with n-step returns," *arXiv preprint arXiv:2503.03660*, 2025.

[19] S. Park, K. Frans, S. Levine, and A. Kumar, "Is value learning really the main bottleneck in offline rl?" *Advances in Neural Information Processing Systems*, 2024.

[20] P. Hansen-Estruch, I. Kostrikov, M. Janner, J. G. Kuba, and S. Levine, "Idql: Implicit q-learning as an actor-critic method with diffusion policies," *arXiv preprint arXiv:2304.10573*, 2023.

[21] K. Huang, R. Scalise, C. Winston, A. Agrawal, Y. Zhang, R. Baijal, M. Grotz, B. Boots, B. Burchfiel, M. Itkina, P. Shah, and A. Gupta, "Using non-expert data to robustify imitation learning via offline reinforcement learning," *Under Review*, 2025.

[22] H. Geng, H. Xu, C. Zhao, C. Xu, L. Yi, S. Huang, and H. Wang, "Gapartnet: Cross-category domain-generalizable object perception and manipulation via generalizable and actionable parts," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[23] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *International Conference on Machine Learning*, 2018.