

BORA: Bridging Offline Reinforcement Learning and Online Residual Adaptation for Real-World Dexterous VLA Models

Zhongxi Chen^{1*} Yifan Han^{2*} Yanming Shao³ Huanming Liu⁴
Congsheng Xu¹ Xiaoyu Chen¹ Yao Mu^{1†} Wenzhao Lian^{1†}

¹Shanghai Jiao Tong University (SJTU) ²CASIA

³Shanghai AI Laboratory ⁴USTC

* Co-first authors † Corresponding authors

Abstract: Vision-Language-Action (VLA) models have emerged as a promising paradigm for grounding visual-language understanding into real-world robotic manipulation. However, dexterous manipulation remains challenging for VLA policies due to high-dimensional hand control and compounding execution errors, which makes real-world RL post-training essential for bridging the gap between visually grounded action generation and physically reliable dexterous execution. However, high-dimensional dexterous exploration often triggers temporal inconsistency, sample inefficiency and hardware risks in the real world. To address these challenges, we propose BORA, an offline-to-online RL post-training framework designed for real-world dexterous VLA models. In the offline phase, BORA constructs a critic that takes both the VLM’s cognition tokens and action chunks as inputs. This design enables action-conditioned value guidance, allowing the critic to evaluate dexterous hand motions beyond visual context alone. During the subsequent online phase, BORA freezes the VLA base and introduces a lightweight, Human-in-the-Loop (HiL) chunk-wise residual adaptation mechanism to mitigate real-world execution errors and further correct the offline-learned intents within the actual physical environment. By inheriting the offline critic and employing intervention-driven rewards, BORA effectively corrects execution discrepancies and adapts to real-world physical variances while preserving the pretrained policy as a stable prior. Extensive evaluations across five complex real-world dexterous tasks demonstrate that BORA significantly outperforms pure imitation learning and traditional decoupled RL baselines, achieving a 33% absolute increase in average success rate under standard settings and up to a 43% improvement in unseen object generalization.

Keywords: Dexterous Manipulation, Post-Training, VLA Models

1 Introduction

While Vision-Language-Action (VLA) models have emerged as a powerful paradigm for generalizable robot control, they face severe challenges in real-world dexterous manipulation due to the high degrees of freedom (DoFs) involved. Furthermore, dexterous tasks exhibit inherent action diversity, where multiple distinct hand poses and configurations can successfully execute the same task. In such highly noisy and multimodal continuous action spaces, Imitation Learning (IL) struggles to extract generalized physical interaction intents. Consequently, an efficient Reinforcement Learning (RL) post-training framework that distills intent comprehension from offline data, coupled with online fine-tuning revision, is critical to pushing the boundaries of real-world dexterous control.



Figure 1: **BORA: Bridging Offline Reinforcement Learning and Online Residual Adaptation for Real-World Dexterous VLA Models.** We propose an offline-to-online RL post-training framework for dexterous VLAs, bridging semantic intents and physical dynamics to significantly elevate real-world deployment reliability and task success rates.

Nevertheless, existing VLA post-training schemes [1, 2] encounter two fundamental challenges when applied to dexterous manipulation. The first is credit assignment failure in action generation: mainstream generative action architectures (e.g., Diffusion Models or Flow Matching) typically rely on denoising chains spanning tens or even hundreds of steps. Considering that offline dexterous data often contains redundant or unintentional micro-actions [3], and that the actions are highly diverse, the resulting RL gradients inherently possess significant noise. Backpropagating these gradients through temporal computation graphs spanning thousands of steps leads to severe noise accumulation, rendering the model incapable of effectively extracting high-level intents to guide the underlying action manifold [4]. The second challenge is visual occlusion at the perceptual level. Dexterous manipulation frequently involves severe occlusions; during real-robot RL, traditional decoupled critics are prone to overfitting to background visual artifacts. Rather than evaluating the true physical contact consequences of the actions, these critics provide erroneous guidance to the dexterous VLA models.

Beyond these offline training challenges, real-world deployment inevitably introduces execution errors arising from complex friction and contact dynamics, resulting in state deviations and action-outcome mismatches relative to the offline data distribution. Consequently, the offline-learned intents must be continuously adapted to remain effective under such distribution shifts. While online fine-tuning can mitigate these issues by reinforcing intent guidance, directly updating all parameters of a VLA model is impractical. This is primarily because fragile dexterous hardware limits large-scale data collection, and the resulting online intervention data is often noisy, mixing optimal and suboptimal trajectories. Under traditional offline-to-online RL [5], early-stage distribution shifts and noisy gradients can easily trigger catastrophic feature drift in the pre-trained VLM [6]. Consequently, we need a new framework that prevents catastrophic forgetting during full-parameter updates, regularizes the Q-network within the offline manifold during online estimation, and retains sufficient exploration capability.

To address these challenges, we propose **BORA (Bridging Offline RL and Online Residual Adaptation)**, an RL post-training and adaptation framework custom-designed for dexterous VLAs spanning the offline-to-online spectrum. In the offline phase, BORA deploys a Consistency Policy [7, 8] as the action expert to generate continuous action chunks in just 1–3 steps, truncating the computation graph for efficient gradient backpropagation. Concurrently, to mitigate the impact of visual occlusions and prevent the critic from overfitting to background artifacts, we design a critic that explicitly fuses the continuous action chunks with the VLM’s cognition tokens. This ensures that the value estimation is fundamentally grounded in actual physical interactions rather than spurious visual features. During the subsequent online phase, BORA freezes the offline-trained VLA base to prevent catastrophic feature drift, introducing a lightweight, Human-in-the-Loop (HiL) chunk-wise residual adaptation mechanism. By directly inheriting the offline critic to ensure stable value estimation, this mechanism guides the residual actor to safely extract corrective priors from human inter-

vention data, robustly compensating for real-world execution deviations and intention mismatches. In summary, our main contributions are threefold:

- **Action-Conditioned Critic for Dexterous Manipulation:** we design a critic architecture that fuses continuous action chunks with the VLM’s cognition tokens. This enables precise, action-conditioned value guidance evaluated on physical execution consequences rather than visual context alone.
- **Lightweight Residual Online Adaptation:** We introduce an HiL chunk-wise residual RL mechanism for the real-world deployment phase. By freezing the VLA base, inheriting the offline critic, and leveraging intervention-driven rewards, we achieve safe and sample-efficient online adaptation. This design effectively corrects execution errors while preventing catastrophic forgetting of the pre-trained representation.
- **The BORA Unified Framework:** We present a comprehensive offline-to-online RL post-training framework tailored for dexterous VLAs. By utilizing a consistency policy to resolve generative credit assignment and employing progressive optimization to bridge offline intent learning with online physical execution, BORA significantly enhances real-world deployment robustness, achieving a 33% absolute increase in average success rate and up to a 43% improvement in unseen object generalization.

2 Related Work

Vision-Language-Action Models for Dexterous Manipulation Vision-Language-Action (VLA) models have emerged as a powerful paradigm for robot manipulation, typically adapting visual-language representations from pretrained VLMs to robot control through action heads. While earlier VLA models use autoregressive tokenized actions [9, 10, 11], recent action heads have evolved to diffusion- or flow-based continuous policies [12, 13], and further toward consistency-style few-step generation for efficient closed-loop control [14, 15]. This progress has also motivated dexterous VLA or vision-language-grasp models, as dexterous hands offer substantially greater capability than parallel grippers for fine-grained and contact-rich manipulation [16, 17, 18]. Recent works such as Being-H0 [19] and VITRA [12] study pretraining from human data, while Being-H0.5 [16] explores cross-embodiment transfer across dexterous hands. However, existing dexterous VLA models still suffer from limited real-world success rates, largely due to the high-dimensional hand-arm action space, complex contact dynamics, and noisy dexterous manipulation data, which make offline post-training unstable and inefficient.

Post-Training and Adaptation of Vision-Language-Action Policies Post-training has become an essential step for adapting pretrained VLA policies to downstream robots and task domains. Unlike VLM post-training, where preference optimization can often be conducted on static data, robotic post-training must optimize closed-loop physical interaction under sparse rewards, distribution shifts, and limited real-world samples. Existing methods can be broadly divided into offline and online paradigms. Offline RL and imitation-based fine-tuning can reuse collected robot data and are thus scalable, but they often bring limited gains due to suboptimal demonstrations, action multimodality, and the mismatch between offline trajectories and deployment dynamics [20]. Online RL directly optimizes task success in the target environment and has shown strong potential for robot manipulation, yet it remains costly and unstable, especially for high-DoFs dexterous hands where exploration is sample-inefficient and potentially unsafe [21, 22, 23]. Recent offline-to-online RL methods attempt to combine the scalability of offline data with the adaptivity of online interaction [24, 25]. However, most existing post-training and residual adaptation methods are developed for robot arms or parallel grippers, where the action space is relatively low-dimensional and contact uncertainty is easier to handle. For dexterous VLA policies, high-dimensional hand-arm actions, severe hand-object occlusions, and noisy correction data substantially destabilize value estimation and policy updates. As a result, directly applying conventional offline or online RL post-training methods often leads to inefficient training and limited success-rate improvement in real-world dexterous manipulation.

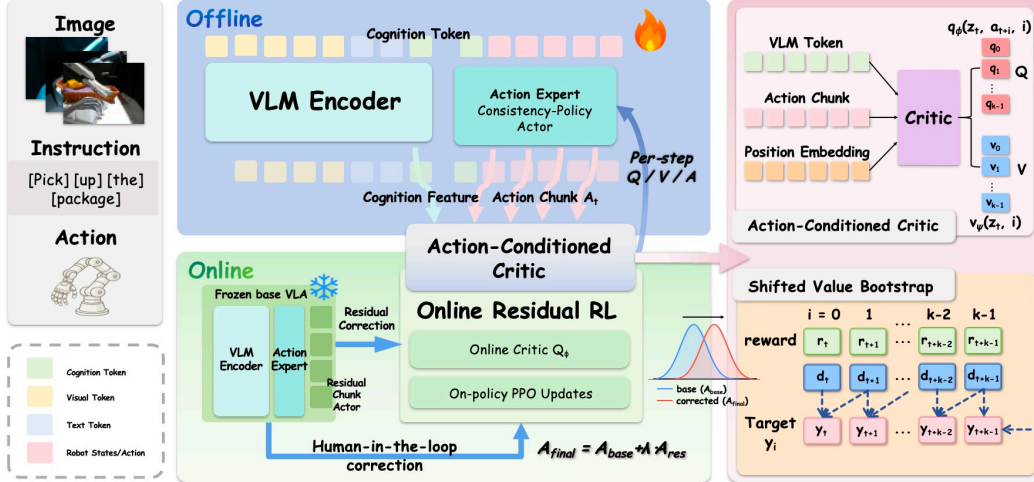


Figure 2: **Illustration of the BORA framework.** BORA bridges offline token-action reinforcement learning and online residual adaptation for real-world dexterous VLA policies. In the offline stage, the VLM encoder and action expert produce shared VLM cognition tokens and action chunks, jointly evaluated by an integrated critic Q_ϕ with semantic anchoring and IQL-based policy optimization. The right panel details the action-conditioned critic, which predicts per-step Q and V values from VLM tokens, action chunks, and position embeddings, with shifted value bootstrap for credit propagation within each chunk. In the online stage, the offline-trained base VLA is frozen, while a lightweight residual chunk actor π_{res} is trained with inherited critic feedback, sparse task rewards, and human-in-the-loop intervention signals. The final action chunk is obtained by residual composition, $A_{\text{final}} = A_{\text{base}} + \lambda A_{\text{res}}$ [26], enabling low-cost physical adaptation and improved dexterous execution.

3 Method

In this work, we propose **BORA**, a comprehensive offline-to-online reinforcement learning framework tailored for Vision-Language-Action (VLA) models in dexterous manipulation. The core philosophy of BORA is to establish a two-stage adaptation pipeline: it first extracts foundational manipulation skills via offline RL, and subsequently compensates for real-world execution errors through Human-in-the-Loop (HiL) online residual RL adaptation on the physical robot.

3.1 Offline RL with Action-Conditioned Critic and Consistency Policy

Existing generative action architectures typically rely on iterative denoising procedures, which inadvertently exacerbate credit assignment failures in continuous control [27]. This bottleneck is particularly detrimental in dexterous manipulation: given the exceptionally high DoFs, the action manifold is inherently multimodal and riddled with redundant, task-irrelevant micro-actions. Back-propagating RL signals through hundreds of temporal denoising steps over such a noisy space inevitably leads to severe noise accumulation. Consequently, the foundational VLM fails to receive effective gradients, disrupting the alignment between high-level task representations and low-level action execution.

To address this, we parameterize the actor as a consistency policy, enabling high-fidelity action chunk generation within just 1–3 denoising steps [7, 8]. By truncating the computation graph, this formulation ensures that informative gradients flow efficiently back into the VLM, thereby facilitating stable offline optimization. Furthermore, to avoid relying on privileged information for reward design [28], we adopt a minimalist sparse reward formulation consisting of a terminal success reward and a step-wise time penalty. However, evaluating high-dimensional action chunks under such sparse rewards—especially when compounded by severe visual occlusions—renders decoupled critics highly susceptible to overfitting to spurious background artifacts rather than actual physical interactions.

Action-Conditioned Critic. To counteract this visual overfitting and handle the high dimensionality of chunk-level evaluation, we condition our critic on VLM semantic tokens $z_t = \Phi_{\text{VLM}}(s_t)$ and decompose the chunk-level evaluation into atomic, step-wise scores. Specifically, our critic outputs k -dimensional value vectors defined as $Q_\phi(s_t, A_t) = [q_\phi(z_t, a_{t+i}, i)]_{i=0}^{k-1} \in \mathbb{R}^k$ and $V_\psi(s_t) = [v_\psi(z_t, i)]_{i=0}^{k-1} \in \mathbb{R}^k$, where the subscript i denotes the relative position encoded by a learnable positional embedding. To enforce temporal consistency across this vectorized horizon, credit is propagated backward within the chunk via a shifted value bootstrap:

$$y_{t,i} = \begin{cases} r_{t+i} + \gamma(1 - d_{t+i}) v_\psi(z_t, i + 1), & i < k - 1 \\ r_{t+k-1} + \gamma(1 - d_{t+k-1}) v_\psi(z_{t+1}, 0), & i = k - 1 \end{cases} \quad (1)$$

The critic parameters ϕ and ψ are optimized via a masked Bellman residual together with an IQL-style expectile value objective, which extracts conservative value targets from sub-optimal offline data (see Appendix A for details).

Conservative Policy Improvement. Instead of using isolated per-step residuals, we smooth the optimization signal using an intra-chunk Action-level GAE Recursion: $\hat{A}_{t,i} = \delta_{t,i} + \gamma\lambda(1 - d_{t+i})\hat{A}_{t,i+1}$ with $\hat{A}_{t,k} = 0$, where $\delta_{t,i} = q_\phi(z_t, a_{t+i}, i) - v_\psi(z_t, i)$ represents the instantaneous advantage. The consistency-based policy π_θ defines a Gaussian action distribution at each atomic step. Simulating policy updates with the importance sampling ratio $r_{t,i}(\theta) = \frac{\pi_\theta(a_{t+i}|s_t)}{\pi_{\theta_{\text{old}}}(a_{t+i}|s_t)}$, we optimize the actor using a validity-masked clipped PPO surrogate objective:

$$\mathcal{L}_{\text{PPO}}(\theta) = -\mathbb{E}_{(s_t, A_t) \sim \mathcal{D}} \left[\frac{1}{|M_t|} \sum_{i=0}^{k-1} m_{t,i} \min \left(r_{t,i} \hat{A}_{t,i}, \text{clip}(r_{t,i}, 1 - \epsilon, 1 + \epsilon) \hat{A}_{t,i} \right) \right] \quad (2)$$

where $m_{t,i}$ is a validity mask and $|M_t| = \sum_i m_{t,i}$ is the effective chunk length. To ensure policy conservatism and prevent out-of-distribution deviation, the full objective combines PPO with a Behavior Cloning (BC) regularization on the normalized action chunk: $\mathcal{L}_{\text{actor}} = \lambda_{\text{ppo}} \mathcal{L}_{\text{PPO}} + \lambda_{\text{bc}} \mathcal{L}_{\text{BC}}$.

3.2 Bridging the Deployment Gap: HiL Residual Chunk Adaptation

While offline RL equips the VLA model with robust behavioral priors, direct physical deployment inevitably exposes the policy to compounding execution drift induced by complex contact dynamics. Simultaneously, full-parameter online fine-tuning is computationally prohibitive and highly susceptible to catastrophic feature drift in pre-trained vision-language representations and value collapse [6].

To achieve rapid, sample-efficient real-world adaptation while preserving the foundation model’s priors, BORA freezes the offline-trained VLA base and introduces a lightweight Residual Chunk Actor π_{res} (parameterized by an MLP). Operating natively in the continuous temporal domain, π_{res} generates compensations at the action chunk level. Given the proprioceptive state s_{prop} , the base action chunk A_{base} , and the VLM tokens z_{VLM} , the final deployed composite chunk is formulated as $A_{\text{final}} = A_{\text{base}} + \lambda_{\text{res}} \cdot \pi_{\text{res}}(s_{\text{prop}}, A_{\text{base}}, z_{\text{VLM}})$, where λ_{res} is a scaling coefficient that restricts the intervention magnitude to guarantee the smoothness of the action manifold.

During online adaptation, BORA couples **Critic Inheritance** with an **Intervention-Driven RLPD** pipeline to stabilize optimization dynamics. The offline-trained critic Q_ϕ initializes the online value function, preserving a pre-shaped value landscape. To enforce monotonic policy improvement, the residual actor optimizes a conservative value guidance objective (Appendix B.3) that penalizes residual actions underperforming the frozen VLA base. Concurrently, human corrections are integrated via an RLPD protocol that maintains a 1:1 sampling ratio between the static offline dataset and an online buffer dynamically enriched by human interventions [25]. This joint optimization over mixed data streams regularizes the model against non-stationary online trajectories, enabling precise calibration of the residual action advantage.

Integrating with the stabilization method, we then design a Human-in-the-Loop (HiL) intervention system following DexHiL [2]. Specifically, BORA allows human operators to take control to recover

the task whenever the model encounters failures (*e.g.*, incomplete closing of a finger or wrist misplacement). To mitigate kinematic discontinuities during the transition from autonomous execution to human takeover, a linear interpolation mechanism for action smoothing is employed, ensuring the temporal consistency of the hand poses.

Finally, to supply the core optimization signals for the Intervention-Driven RLPD pipeline, we engineer an asymmetric Intervention-Driven Reward function. If the policy drifts OOD and triggers an intervention, an instant penalty r_{int} is imposed. Conversely, upon completion of a human corrective action, a positive recovery reward r_{rec} is granted. This asymmetric reward mechanism directly guides the RLPD value updates, compelling the residual policy to aggressively penalize high-risk states while efficiently learning from high-quality recovery trajectories, ultimately closing the physical adaptation loop with minimal real-world interactions.

4 Experiments

In this section, we design a series of experiments to investigate our proposed BORA by asking the following 3 questions:

- RQ1:** Can BORA improve offline RL post-training for dexterous VLA policies by providing action-conditioned value guidance under high-dimensional action generation and visual occlusion?
- RQ2:** How can the proposed online adaptation bridge the real-world deployment gap in a sample-efficient way?
- RQ3:** How does the inherited action-conditioned critic support residual policy learning during online fine-tuning?

4.1 Experimental Setup

As detailed in Appendix C, We deployed BORA on a real-world dexterous manipulation platform composed of a Franka arm equipped with a 12-DoF dexterous hand. We evaluate the models across the following five tasks: Pick the plush toy, Pick and Place, Open the box, Pull the tissue, Press the button. For each task, we conduct 20 trials under the standard configuration and another 20 trials involving novel, unseen objects to assess BORA’s generalization capabilities.

To systematically ablate our framework, we compare BORA against four baseline settings. Built upon the pre-trained VITRA VLM, these baselines all adopt the VLA architecture:

- **VITRA (Fine-tuned Base) [12]:** A baseline fine-tuned on top of the vanilla pre-trained VITRA weights, constructed via a multimodal VLM backbone paired with a Diffusion Action Expert.
- **CP Base (consistency policy):** An Imitation Learning baseline utilizing the consistency policy as an action expert without any reinforcement learning fine-tuning.
- **Decoupled-Critic Baseline [29]:** An offline RL baseline with a separately trained critic that does not jointly condition on the VLM cognition tokens and generated action chunks.
- **BORA-Offline (Ours):** Our offline RL framework with an action-conditioned critic that evaluates VLM cognition tokens together with continuous action chunks.

Then, our setting, **BORA-Full (Ours)**, is the complete pipeline, featuring BORA-Offline followed by HiL online residual chunk adaptation through 2 rounds of real-world physical reinforcement learning. The online training process incorporating these components is formally summarized in Algorithm 1.

4.2 Main Results and Analysis

We report the main quantitative results in Tables 1 and 2, and provide a visual summary in Fig. 3. The tables present exact success counts over 20 real-world trials for each task and setting, while the bar plots summarize the relative trends across baselines. Fig. 5 in the Appendix further presents representative real-robot results across the five evaluation tasks.

Table 1: Real-World Dexterous Manipulation Success Rates under the **Standard** Configuration.

| Task | Pure Imitation Learning (IL) | | Reinforcement Learning (RL) Fine-tuning | | |
|--------------------|------------------------------|-------------|---|--------------|-------------------------|
| | VITRA (Diffusion) | CP Base | CP + Decoupled Critic | BORA-Offline | BORA-Full (Ours) |
| Pick the plush toy | 14/20 (70%) | 12/20 (60%) | 10/20 (50%) | 17/20 (85%) | 20/20 (100%) |
| Pick and Place | 9/20 (45%) | 10/20 (50%) | 8/20 (40%) | 14/20 (70%) | 18/20 (90%) |
| Open the box | 12/20 (60%) | 11/20 (55%) | 5/20 (25%) | 11/20 (55%) | 15/20 (75%) |
| Pull the tissue | 9/20 (45%) | 7/20 (35%) | 10/20 (50%) | 12/20 (60%) | 16/20 (80%) |
| Press the button | 10/20 (50%) | 13/20 (65%) | 12/20 (60%) | 13/20 (65%) | 17/20 (85%) |

Table 2: Real-World Dexterous Manipulation Success Rates under the **Object-Unseen** Setting.

| Task | Pure Imitation Learning (IL) | | Reinforcement Learning (RL) Fine-tuning | | |
|--------------------|------------------------------|------------|---|--------------|-------------------------|
| | VITRA (Diffusion) | CP Base | CP + Decoupled Critic | BORA-Offline | BORA-Full (Ours) |
| Pick the plush toy | 9/20 (45%) | 7/20 (35%) | 11/20 (55%) | 15/20 (75%) | 17/20 (85%) |
| Pick and Place | 6/20 (30%) | 6/20 (30%) | 8/20 (40%) | 9/20 (45%) | 14/20 (70%) |
| Open the box | 2/20 (10%) | 3/20 (15%) | 1/20 (5%) | 7/20 (35%) | 10/20 (50%) |
| Pull the tissue | 7/20 (35%) | 3/20 (15%) | 10/20 (50%) | 10/20 (50%) | 14/20 (70%) |
| Press the button | 9/20 (45%) | 8/20 (40%) | 11/20 (55%) | 11/20 (55%) | 15/20 (75%) |

Offline Intent Alignment with Action-Conditioned Value Guidance (RQ1): As shown in Tables 1 and 2, **CP Base** reaches 53.0% average success under the standard setting but drops to 27.0% under object-unseen evaluation, indicating limited generalization in high-dimensional and multi-modal dexterous action spaces. By routing informative policy gradients back into the foundational VLM within minimal steps, **BORA-Offline** extracts high-fidelity physical intents, improving the averages to 67.0% and 52.0%, respectively, yielding a 14-point gain in the standard setting and a 25-point gain under object-unseen evaluation. This indicates that, although standard diffusion architectures hinder direct offline RL fine-tuning due to severe gradient vanishing across lengthy denoising iterations, our truncated consistency policy formulation effectively addresses this bottleneck.

Crucially, this architectural synergy successfully immunizes the model against the visual occlusion. Traditional critic-decoupled architectures (**Decoupled-Critic baseline**) overfit to raw pixel values and spurious background features, propagating erroneous gradients back to the VLM backbone. In the standard setting, the Decoupled-Critic baseline falls below CP Base on average. Under object-unseen evaluation, although it improves over CP Base on some tasks, it collapses on Open-the-Box, indicating unstable value guidance under severe occlusion. Conversely, by conditioning our Integrated Critic directly on the VLM’s semantic tokens (z_t), **BORA-Offline** binds value updates strictly to the multimodal latent space.

This is qualitatively supported by the t-SNE visualization in Appendix Fig. 8, where BORA-Offline stays closer to the SFT/BC manifold after offline RL, suggesting better preservation of the policy’s action-representation structure. Appendix Fig. 9 and Fig. 10 further show that BORA assigns stronger saliency to the dexterous hand and task-relevant contact regions, while the decoupled critic exhibits more scattered responses over the table, background, and other non-contact regions.

Sample-Efficient Online Physical Adaptation (RQ2): Despite the intent awareness established offline, real-world deployment introduces covariate shifts, execution errors, and intricate contact friction, causing a performance drop on object-unseen tasks. To address this offline-to-online gap, the online residual chunk adaptation in **BORA-Full** provides a sample-efficient solution. Specifically, BORA-Full exhibits rapid performance convergence within the initial two online reinforcement learning rounds, beyond which further improvement becomes minimal. In practice, the human operator only needs to intervene 1–2 times per task, which demands a minimal time investment of approximately 20% of the online trajectory execution. Benefiting from this residual training and the human-intervention value guidance mechanism, BORA-Full improves the overall average success rate to 86.0% in the Standard setting, and elevates it from 52.0% to 70.0% under the Object-Unseen configuration. This efficient adaptation achieves the best performance among the evaluated base-

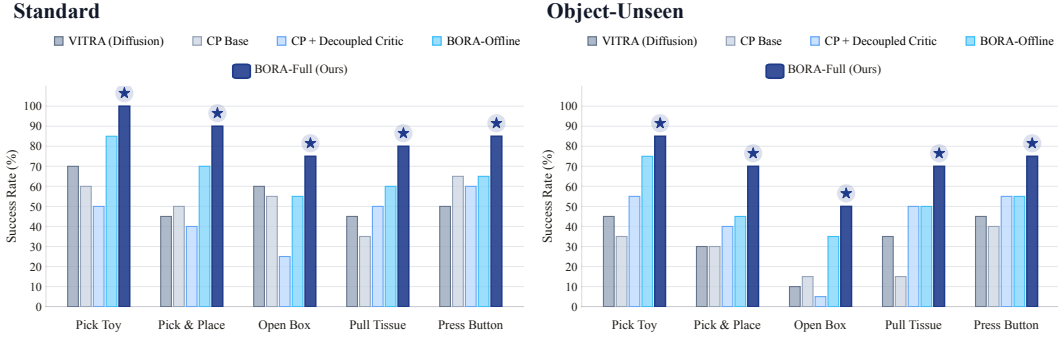


Figure 3: Visual summary of real-world dexterous manipulation results. Success rates are reported across five tasks under the standard and object-unseen settings. BORA-Offline improves over imitation and decoupled-critic baselines, while BORA-Full further improves performance through online residual adaptation.

lines on dexterous execution tasks, while preventing representation drift in the frozen pre-trained VLM foundation.

Mechanistic Analysis of Critic Inheritance (RQ3): To understand the underlying mechanism behind the rapid online adaptation, we analyze the behaviors of the inherited value function, with additional value-profile visualization provided in Appendix Fig. 11. The empirical value profiles reveal that the offline-trained action-conditioned critic, when combined with the online asymmetric intervention-driven reward, provides stable and responsive value estimation. Throughout autonomous execution, the inherited critic exhibits clear discriminative capability: it maintains high value confidence along successful trajectories, while its Q-value estimation remains consistently low along failed trajectories. Importantly, this discriminative capability is preserved even in visually similar states, such as when the multi-fingered hand grasps an object under severe occlusion. Although the raw camera pixels appear nearly identical to those in the decoupled baseline, the action-conditioned critic inherited from the offline stage allows the critic to evaluate state-action pairs at a structural semantic level. Consequently, the inherited critic mitigates superficial visual ambiguities, penalizes high-risk execution states, and guides the lightweight residual policy π_{res} to incorporate human corrective priors with few physical interactions.

5 Limitations

While BORA demonstrates compelling performance, it presents two main limitations. First, the framework relies on visuo-proprioceptive inputs and lacks dense tactile feedback. Integrating high-fidelity tactile arrays into VLM tokens could further tighten the perception-action loop under severe visual occlusion. Second, our physical evaluation is constrained to a single arm-hand topology. Verifying BORA’s cross-embodiment generalization across diverse multi-fingered hands with varying kinematics and degrees of freedom remains an important direction for future scaling.

6 Conclusion

In this paper, we presented BORA, an offline-to-online RL post-training framework custom-designed for real-world dexterous VLA models. By seamlessly bridging offline alignment with efficient online fine-tuning, BORA addresses key challenges in dexterous VLA post-training, including critic overfitting to visual artifacts, real-world execution discrepancies, and catastrophic feature drift. Extensive evaluations across five complex real-world dexterous tasks demonstrate that BORA significantly outperforms imitation learning and decoupled RL baselines, achieving a 33% absolute increase in average success rate and up to 43% improvement in unseen object generalization. Moving forward, we aim to extend this framework toward high-precision dexterous skills and investigate its scalability to structurally complex, long-horizon operational tasks.

References

- [1] T.-Y. Xiang, A.-Q. Jin, X.-H. Zhou, M.-J. Gui, X.-L. Xie, S.-Q. Liu, S.-Y. Wang, S.-B. Duan, F.-C. Xie, W.-K. Wang, et al. Parallels between vla model post-training and human motor learning: Progress, challenges, and trends. *arXiv preprint arXiv:2506.20966*, 2025.
- [2] Y. Han, Z. Chen, Y. Zhao, C. Xu, Y. Shao, Y. Peng, Y. Mu, and W. Lian. Dexhil: A human-in-the-loop framework for vision-language-action model post-training in dexterous manipulation. *arXiv preprint arXiv:2603.09121*, 2026.
- [3] A. Mandlekar, D. Xu, J. Wong, S. Nasiriany, C. Wang, R. Kulkarni, L. Fei-Fei, S. Savarese, Y. Zhu, and R. Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. *arXiv preprint arXiv:2108.03298*, 2021.
- [4] D. Zhang, J. Sun, C. Hu, X. Wu, Z. Yuan, R. Zhou, F. Shen, and Q. Zhou. Pure vision language action (vla) models: A comprehensive survey. *arXiv preprint arXiv:2509.19012*, 2025.
- [5] P. Intelligence, A. Amin, R. Aniceto, A. Balakrishna, K. Black, K. Conley, G. Connors, J. Darpinian, K. Dhabalia, J. DiCarlo, et al. $\pi_{0.6}^*$: A vla that learns from experience. *arXiv preprint arXiv:2511.14759*, 2025.
- [6] Z. Zhou, A. Peng, Q. Li, S. Levine, and A. Kumar. Efficient online reinforcement learning fine-tuning need not retain offline data. In *International Conference on Learning Representations*, volume 2025, pages 32343–32368, 2025.
- [7] A. Prasad, K. Lin, J. Wu, L. Zhou, and J. Bohg. Consistency policy: Accelerated visuomotor policies via consistency distillation. *arXiv preprint arXiv:2405.07503*, 2024.
- [8] G. Lu, Z. Gao, T. Chen, W. Dai, Z. Wang, W. Ding, and Y. Tang. Manicm: Real-time 3d diffusion policy via consistency model for robotic manipulation. *arXiv preprint arXiv:2406.01586*, 2024.
- [9] 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*, pages 2165–2183. PMLR, 2023.
- [10] O. Mees, D. Ghosh, K. Pertsch, K. Black, H. R. Walke, S. Dasari, J. Hejna, T. Kreiman, C. Xu, J. Luo, et al. Octo: An open-source generalist robot policy. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.
- [11] 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.
- [12] Q. Li, Y. Deng, Y. Liang, L. Luo, L. Zhou, C. Yao, L. Zeng, Z. Feng, H. Liang, S. Xu, et al. Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. *arXiv preprint arXiv:2510.21571*, 2025.
- [13] M. Chen, Y. Wang, Z. Li, H. Bharadhwaj, Y. Chen, C. Qin, Z. Kou, Y. Tian, E. Whitmire, R. Sodhi, et al. Flowing from reasoning to motion: Learning 3d hand trajectory prediction from egocentric human interaction videos. *arXiv preprint arXiv:2512.16907*, 2025.
- [14] 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.
- [15] Physical 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.

- [16] H. Luo, Y. Wang, W. Zhang, S. Zheng, Z. Xi, C. Xu, H. Xu, H. Yuan, C. Zhang, Y. Wang, et al. Being-h0. 5: Scaling human-centric robot learning for cross-embodiment generalization. *arXiv preprint arXiv:2601.12993*, 2026.
- [17] Y. Zhong, X. Huang, R. Li, C. Zhang, Z. Chen, T. Guan, F. Zeng, K. N. Lui, Y. Ye, Y. Liang, et al. Dexgraspvla: A vision-language-action framework towards general dexterous grasping. *arXiv preprint arXiv:2502.20900*, 2025.
- [18] J. He, D. Li, X. Yu, Z. Qi, W. Zhang, J. Chen, Z. Zhang, Z. Zhang, L. Yi, and H. Wang. Dexvlg: Dexterous vision-language-grasp model at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14248–14258, 2025.
- [19] H. Luo, Y. Feng, W. Zhang, S. Zheng, Y. Wang, H. Yuan, J. Liu, C. Xu, Q. Jin, and Z. Lu. Being-h0: vision-language-action pretraining from large-scale human videos. *arXiv preprint arXiv:2507.15597*, 2025.
- [20] D. Huang, Z. Fang, T. Zhang, Y. Li, L. Zhao, and C. Xia. Co-rft: Efficient fine-tuning of vision-language-action models through chunked offline reinforcement learning. *arXiv preprint arXiv:2508.02219*, 2025.
- [21] Y. Chen, S. Tian, S. Liu, Y. Zhou, H. Li, and D. Zhao. Conrft: A reinforced fine-tuning method for vla models via consistency policy. *arXiv preprint arXiv:2502.05450*, 2025.
- [22] C. Xu, J. T. Springenberg, M. Equi, A. Amin, A. Esmail, S. Levine, and L. Ke. Rl token: Bootstrapping online rl with vision-language-action models. *arXiv preprint arXiv:2604.23073*, 2026.
- [23] K. Chen, Z. Liu, T. Zhang, Z. Guo, S. Xu, H. Lin, H. Zang, Q. Zhang, Z. Yu, G. Fan, et al. π rl: Online rl fine-tuning for flow-based vision-language-action models. *arXiv preprint arXiv:2510.25889*, 2025.
- [24] M. Nakamoto, S. Zhai, A. Singh, M. Sobol Mark, Y. Ma, C. Finn, A. Kumar, and S. Levine. Cal-ql: Calibrated offline rl pre-training for efficient online fine-tuning. *Advances in Neural Information Processing Systems*, 36:62244–62269, 2023.
- [25] P. J. Ball, L. Smith, I. Kostrikov, and S. Levine. Efficient online reinforcement learning with offline data. In *International Conference on Machine Learning*, pages 1577–1594. PMLR, 2023.
- [26] E. Su, T. Westenbroek, A. Nagabandi, and A. Gupta. Rfs: Reinforcement learning with residual flow steering for dexterous manipulation. In *The Fourteenth International Conference on Learning Representations*, 2026.
- [27] H. Ma, T. Chen, K. Wang, N. Li, and B. Dai. Efficient online reinforcement learning for diffusion policy. *arXiv preprint arXiv:2502.00361*, 2025.
- [28] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, and S. Levine. Learning Complex Dexterous Manipulation with Deep Reinforcement Learning and Demonstrations. In *Proceedings of Robotics: Science and Systems (RSS)*, 2018.
- [29] J. Luo, Z. Hu, C. Xu, Y. L. Tan, J. Berg, A. Sharma, S. Schaal, C. Finn, A. Gupta, and S. 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.

Appendix

A Detailed Critic Formulation

Following the chunk-wise Bellman backup defined in the main text, the Q-function parameters ϕ and value-function parameters ψ are updated in the same offline training loop, but with decoupled objectives. For notation brevity, we write $q_{t,i} \triangleq q_\phi(z_t, a_{t+i}, i)$ and $v_{t,i} \triangleq v_\psi(z_t, i)$. The critic objectives are

$$\mathcal{L}_Q(\phi) = \mathbb{E}_{(s_t, A_t) \sim \mathcal{D}} \left[\frac{1}{|M_t|} \sum_{i=0}^{k-1} m_{t,i} (q_{t,i} - y_{t,i})^2 \right], \quad (3)$$

$$\mathcal{L}_V(\psi) = \mathbb{E}_{(s_t, A_t) \sim \mathcal{D}} \left[\frac{1}{|M_t|} \sum_{i=0}^{k-1} m_{t,i} |\tau - \mathbf{1}[q_{t,i} - v_{t,i} < 0]| (q_{t,i} - v_{t,i})^2 \right], \quad (4)$$

where $m_{t,i} \in \{0, 1\}$ is a validity mask that keeps all atomic steps up to and including the first terminal step within the chunk. The parameter $\tau \in (0.5, 1.0)$ is the expectile asymmetry coefficient, following the value-learning design of IQL. In implementation, the Q estimate used in \mathcal{L}_V is treated as a stop-gradient target.

B Implementation Details and Training Safeguards

To stabilize offline post-training of the VLM-based actor, we use a decoupled actor-critic optimization scheme. The critic is trained with the Bellman and expectile objectives above, while the actor is optimized separately using a PPO-clipped likelihood-ratio objective on demonstration actions together with behavior cloning regularization. In the current implementation, no direct pathwise critic gradient $\nabla_a Q_\phi$ is backpropagated into the actor.

B.1 Advantage Gating Mechanism

To reduce harmful policy updates under noisy offline value estimates, we optionally gate the actor update using the mean validity-masked chunk advantage:

$$\bar{A}_t = \frac{1}{|M_t|} \sum_{i=0}^{k-1} m_{t,i} \hat{A}_{t,i}. \quad (5)$$

In practice, this gating score may be computed from either raw or normalized advantages. If $\bar{A}_t \leq \alpha$, the actor update for the current batch is skipped; the critic update is still executed.

B.2 Two-Stage Post-Training Pipeline

The offline post-training procedure is divided into two stages:

1. **Stage 1: BC Warm-up.** The actor is optimized only with the BC objective, while the critic is trained simultaneously from offline transitions. PPO guidance is disabled during this phase.
2. **Stage 2: BC+PPO Guidance.** After the warm-up stage, the actor is optimized with the combined objective $\mathcal{L}_{\text{actor}} = \lambda_{\text{ppo}} \mathcal{L}_{\text{PPO}} + \lambda_{\text{bc}} \mathcal{L}_{\text{BC}}$, using critic-derived action-level advantages as weighting signals. The critic remains decoupled from the actor in the backward pass.

B.3 Online Residual Alignment via Conservative Value Guidance

In the online adaptation phase, to guarantee that the lightweight residual actor π_{res} strictly improves execution quality over the frozen VLA base a_{t+i}^{base} , we introduce the conservative improvement hinge

loss:

$$L_{\text{improve}} = \mathbb{E} \left[\frac{1}{\sum_i m_{t,i}} \sum_i m_{t,i} \max \left(0, Q_\phi(z_t, a_{t+i}^{\text{base}}) + \delta - Q_\phi(z_t, \hat{a}_{t+i}) \right) \right]. \quad (6)$$

This formulation can be theoretically justified as a relaxation of a constrained optimization problem. Ideally, we seek a residual policy that maximizes the expected return while satisfying a local policy improvement bound:

$$\max_{\pi_{\text{res}}} \mathbb{E} [Q_\phi(z_t, \hat{a}_{t+i})] \quad \text{s.t.} \quad Q_\phi(z_t, \hat{a}_{t+i}) \geq Q_\phi(z_t, a_{t+i}^{\text{base}}) + \delta \quad (7)$$

By constructing the Lagrangian function with local multipliers $\alpha_{t+i} \geq 0$, we have:

$$\mathcal{L}(\pi_{\text{res}}, \alpha) = Q_\phi(z_t, \hat{a}_{t+i}) - \alpha_{t+i} (Q_\phi(z_t, a_{t+i}^{\text{base}}) + \delta - Q_\phi(z_t, \hat{a}_{t+i})) \quad (8)$$

In practical deep reinforcement learning, to avoid gradient instability and maintain manifold smoothness, we parameterize α_{t+i} as a constant coefficient λ_{improve} and convert the constraint into a one-sided hinge penalty. When the composite action \hat{a}_{t+i} fails to outperform the base action by the margin δ , $\mathcal{L}_{\text{improve}}$ provides explicit pathwise repulsive gradients through the learned critic Q_ϕ , regularizing the residual adaptation within a safe optimization boundary.

B.4 Online Residual Adaptation Algorithm

The complete online training process incorporating history-aware base behavior and intervention-driven RLPD is detailed in Algorithm 1.

Algorithm 1 BORA Online Residual Adaptation (Multi-Round Iteration)

Require: Frozen offline VLA base π_{base} , Prior round residual actor $\pi_{\text{res}}^{\text{old}}$ (None if Round 1), Inherited critic Q_ϕ and value network V_ψ , Trainable residual actor π_{res} , Datasets $\mathcal{D}_{\text{offline}}$, $\mathcal{D}_{\text{online}}$, Normalization parameters $(\mu_s, \sigma_s, \mu_a, \sigma_a)$.

Ensure: Optimized residual chunk actor π_{res} .

```

1: for episode = 1 to  $M$  do
2:   Receive initial environment state  $s_0$ , set  $t \leftarrow 0$ 
3:   while not terminal do
4:     Normalize state:  $s_t^{\text{norm}} \leftarrow (s_t - \mu_s) / (\sigma_s + \epsilon)$  and extract VLM tokens  $z_t \leftarrow \Phi_{\text{VLM}}(s_t^{\text{norm}})$ 
5:     Generate prior-policy actions:  $A^{\text{VLA}} \leftarrow \pi_{\text{base}}(z_t)$ 
6:     if  $\pi_{\text{res}}^{\text{old}}$  is None then ▷ Round 1: Use pure VLA as base
7:       Set current base action chunk:  $A_{\text{norm}}^{\text{base}} \leftarrow A^{\text{VLA}}$ 
8:     else ▷ Subsequent Rounds: VLA + prior residual as base
9:       Set current base action chunk:  $A_{\text{norm}}^{\text{base}} \leftarrow A^{\text{VLA}} + \lambda_{\text{old}} \pi_{\text{res}}^{\text{old}}(s_t^{\text{norm}}, A^{\text{VLA}}, z_t)$ 
10:    end if
11:    Compute new residual chunk:  $A_{\text{res}} \leftarrow \pi_{\text{res}}(s_t^{\text{norm}}, A_{\text{norm}}^{\text{base}}, z_t)$ 
12:    Linear schedule factor:  $\alpha_t \leftarrow \alpha_s + (\alpha_e - \alpha_s) \min(t/T_\alpha, 1.0)$ 
13:    Map composite chunk to physical space:  $A_{\text{final}} \leftarrow (A_{\text{norm}}^{\text{base}} + \alpha_t A_{\text{res}}) \cdot \sigma_a + \mu_a$ 
14:    if Human Intervention Triggered then
15:       $A_{\text{exec}} \leftarrow (1 - \beta) A_{\text{final}} + \beta A_{\text{human}}$ , set  $r_{\text{int}}$ ,  $\text{is\_int} \leftarrow \text{True}$ 
16:    else
17:      Set  $A_{\text{exec}} \leftarrow A_{\text{final}}$ ,  $r_{\text{int}} \leftarrow 0$ ,  $\text{is\_int} \leftarrow \text{False}$ 
18:    end if
19:    Step environment:  $s_{t+1}, r_{\text{env}}, \text{done}, \text{info} \leftarrow \text{step}(A_{\text{exec}})$ 
20:    Compute reward:  $r_t \leftarrow r_{\text{env}} + r_{\text{int}} + (r_{\text{rec}} \text{ if } [\text{is\_int and info.recovered}] \text{ else } 0)$ 
21:    Store transition  $(s_t, A_{\text{exec}}, r_t, s_{t+1}, \text{done})$  into  $\mathcal{D}_{\text{online}}$  and update  $s_t \leftarrow s_{t+1}, t \leftarrow t + 1$ 
22:    if  $|\mathcal{D}_{\text{online}}| \geq N_{\text{start}}$  then
23:      Sample mixed batch (1 : 1 ratio):  $\mathcal{B} \leftarrow \mathcal{B}_{\text{online}} \sim \mathcal{D}_{\text{online}} \cup \mathcal{B}_{\text{offline}} \sim \mathcal{D}_{\text{offline}}$ 
24:      Compute bootstrap targets  $y_{t+i}$  and update Critic  $Q_\phi$  via  $\mathcal{L}_Q$ 
25:      Decay imitation weight  $\lambda_{\text{BC}}(t)$  and update Actor  $\pi_{\text{res}}$  via  $\mathcal{L}_{\text{actor}}$ 
26:    end if
27:  end while
28: end for

```

C Experimental Details

C.1 Robot Hardware and Teleoperation Setup

Fig. 4 shows the robot hardware and teleoperation setup used for BORA evaluation. The robotic platform consists of a Franka robotic arm equipped with a DexHand021 dexterous hand, together with Intel RealSense D435 RGB-D cameras for visual observation. For human-in-the-loop online adaptation, we use a wearable teleoperation device to provide corrective demonstrations when the autonomous policy enters failure-prone states. The device captures human hand motions and maps them to dexterous hand commands, allowing the operator to recover the task while preserving the physical interaction context. These corrective trajectories are then incorporated into the online residual adaptation stage as intervention data.

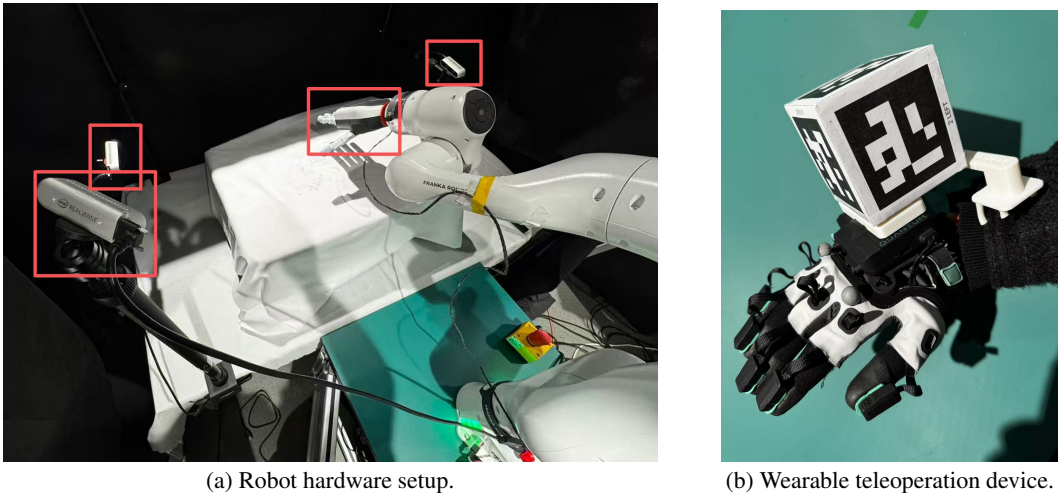


Figure 4: Robot hardware and teleoperation setup. (a) The real-world platform consists of a Franka robotic arm, a DexHand021 dexterous hand, and Intel RealSense D435 RGB-D cameras for multi-view visual observation. Red boxes indicate the main sensing and manipulation components used during evaluation. (b) The wearable teleoperation device captures human hand motions and provides corrective demonstrations for human-in-the-loop online residual adaptation.

C.2 Rollout Visualizations

Fig. 5 provides representative real-world rollout sequences across the five dexterous manipulation tasks. Each row shows temporally ordered frames from one successful execution, covering grasping, placing, pulling, opening, and pressing behaviors. These qualitative examples illustrate that the learned policy can produce coherent action chunks for contact-rich dexterous manipulation under severe hand-object occlusions.

C.3 Object and Task Configurations

Fig. 6 illustrates the object configurations used in our real-world experiments. The object-seen setting contains the object instances used during offline data collection, whereas the object-unseen setting introduces novel instances at evaluation time. Both settings share the same task semantics and robot platform, allowing us to isolate the effect of object-level distribution shift.

Fig. 7 further visualizes the task-level configuration variations used during evaluation. In addition to changing object identities, we also vary physical configurations such as the box opening angle, object orientation, object pose, and object position. These variations are designed to evaluate whether the policy remains robust under realistic layout changes, occlusion patterns, and contact-rich manipulation conditions.

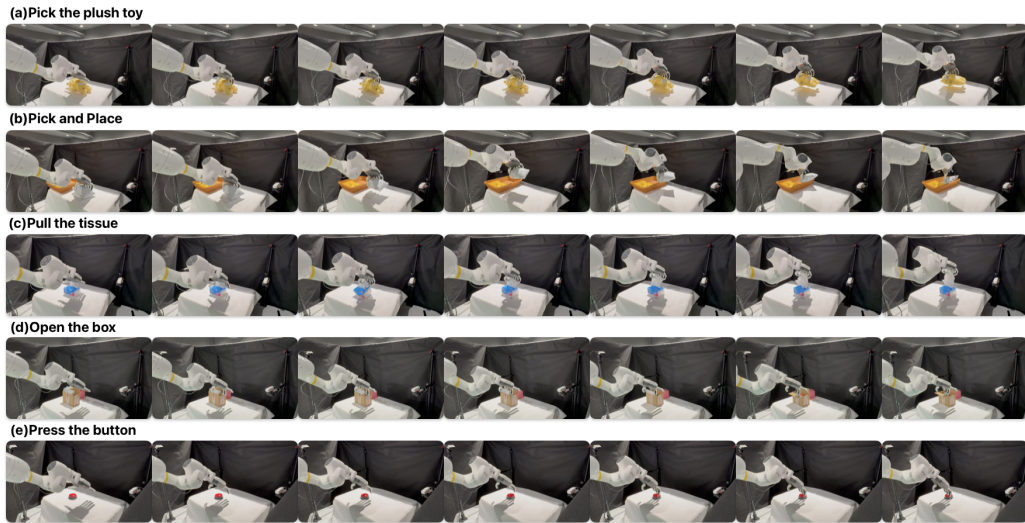


Figure 5: Representative real-world rollout visualizations. From top to bottom, the rows show successful executions of *Pick-the-Plush-Toy*, *Pick-and-Place*, *Pull-the-Tissue*, *Open-the-Box*, and *Press-the-Button*. Each row contains temporally ordered frames from a single rollout, demonstrating coherent dexterous execution across diverse contact-rich manipulation tasks.



(a) Object-seen offline data.



(b) Evaluation with seen and unseen objects.

Figure 6: **Seen and unseen object configurations.** The object-seen setting corresponds to the offline data distribution, while the object-unseen setting includes novel object instances for evaluating generalization under cluttered and occluded real-world dexterous manipulation.



Figure 7: **Task-level evaluation variations.** We visualize representative standard and object-unseen configurations, together with pose, orientation, position, and box opening-angle shifts used to evaluate real-world robustness.

D Supplementary Visualizations

D.1 Additional Representation Visualization

Fig. 8 provides the full t-SNE visualization of projected action representations used in the main analysis. The left two panels correspond to the External Critic variant on the *Pick-the-Plush-Toy* task, while the right two panels correspond to BORA-Offline on the *Pick-and-Place* task. For each setting, we visualize the relation between the SFT/BC manifold anchor and the representation obtained after the same number of offline RL training epochs.

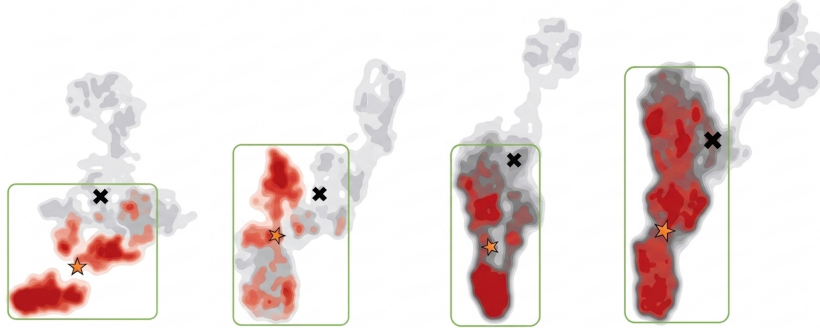


Figure 8: t-SNE visualization of representation drift under offline RL. The left two panels show External Critic on *Pick-the-Plush-Toy*, and the right two panels show BORA-Offline on *Pick-and-Place*. The orange star denotes the SFT/BC manifold anchor, the black cross denotes the RL-updated representation, and the green box marks the local BC support. Red density regions indicate critic-preferred high-value areas, while gray regions denote lower-value regions.

D.2 V-Critic Saliency Visualization

Fig. 9 and Fig. 10 visualize gradient-based saliency maps of the value critic on the *Open-the-Box* and *Pull-the-Tissue* tasks, respectively. The saliency is computed with respect to the visual patch features used by each critic, indicating which visual regions contribute most to the current value estimate. Across both tasks, BORA’s integrated token-action critic places comparatively stronger emphasis on the dexterous hand and task-relevant interaction regions, while the decoupled critic exhibits more scattered responses on object-irrelevant regions. This supports our claim that token-action conditioning encourages value estimation to rely more on task-relevant interaction cues under severe occlusion.

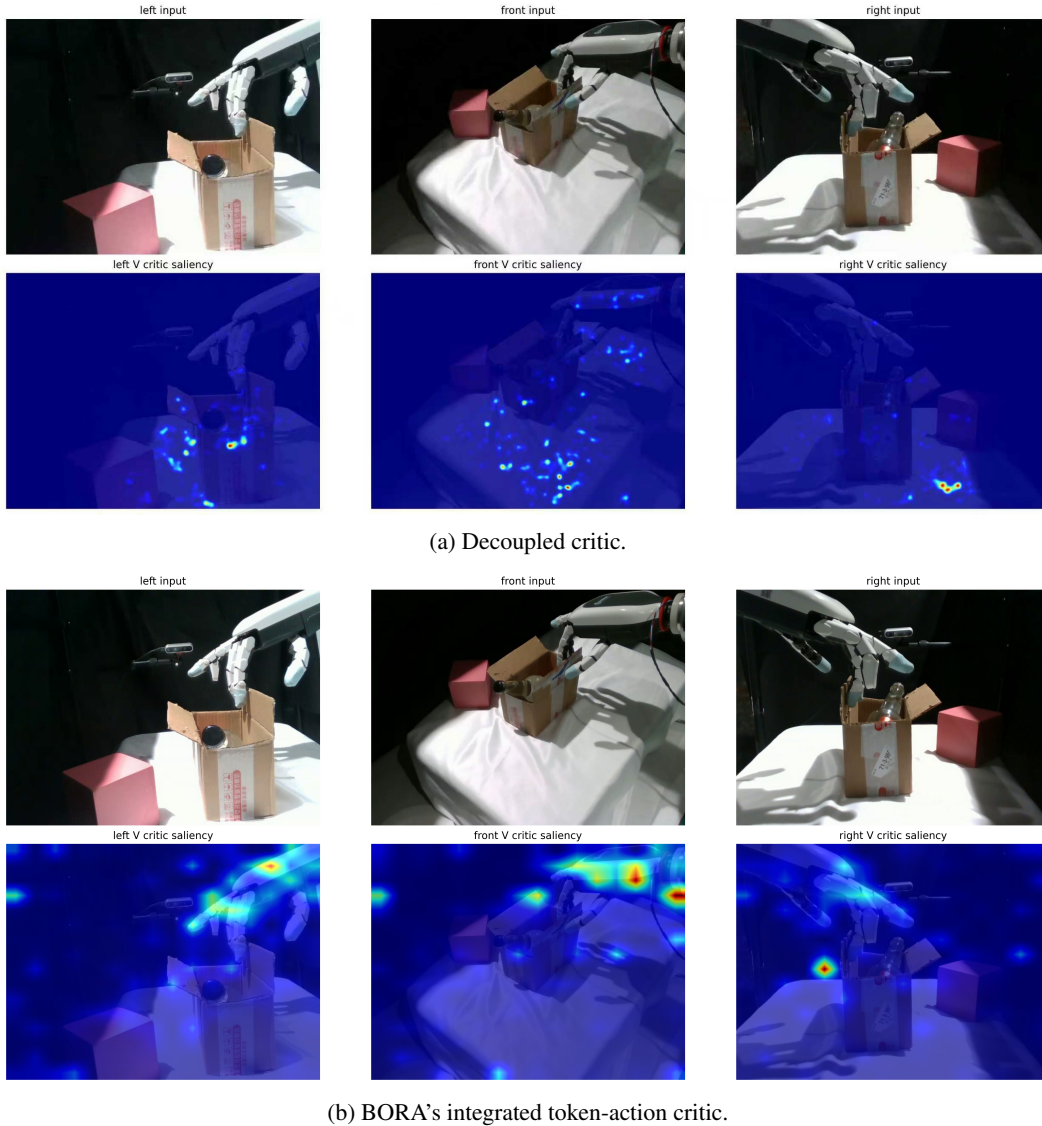
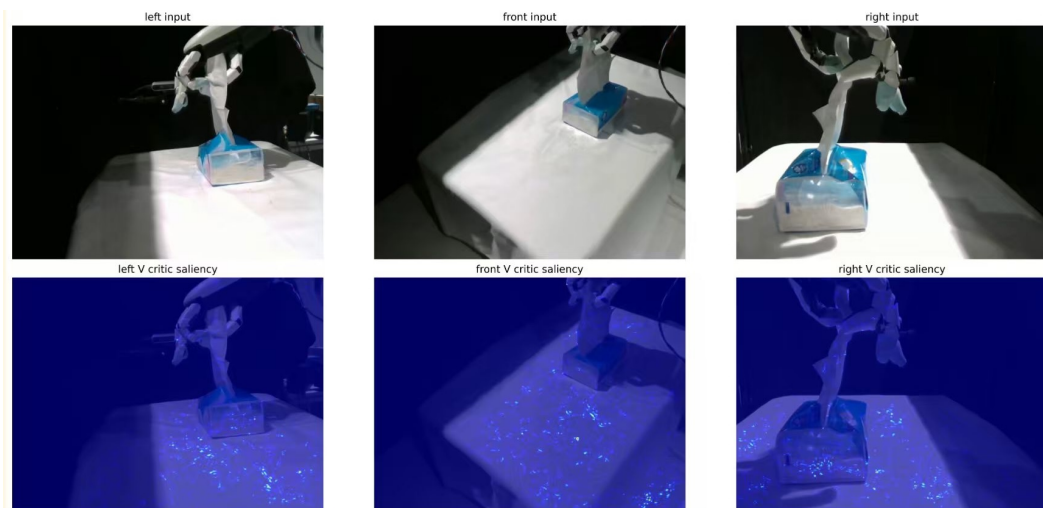
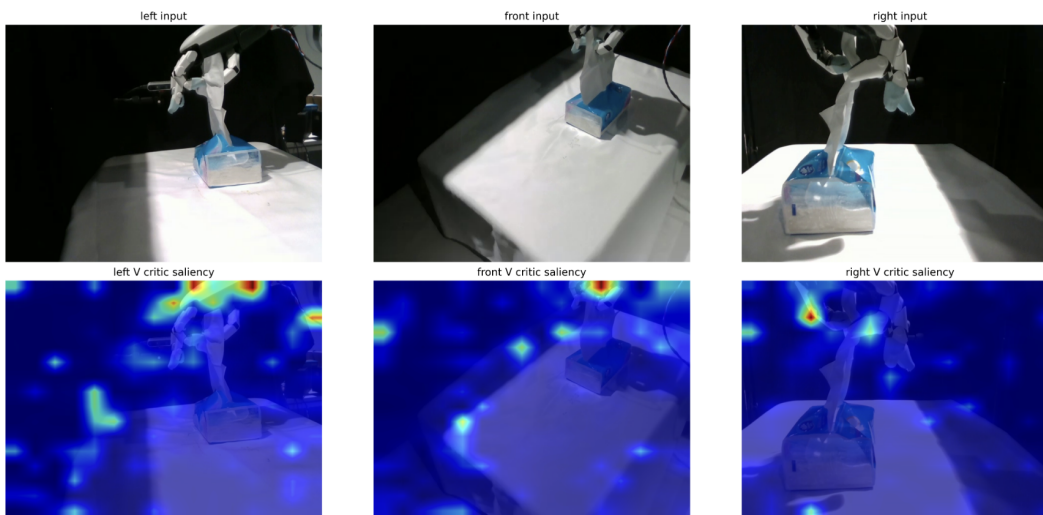


Figure 9: **V-critic saliency maps on the *Open-the-Box* task.** Each subfigure shows RGB observations from the left, front, and right camera views in the top row, with the corresponding gradient-based saliency maps in the bottom row. Warmer colors indicate larger influence on the current value estimate.



(a) Decoupled critic.



(b) BORA's integrated token-action critic.

Figure 10: **V-critic saliency maps on the *Pull-the-Tissue* task.** Each subfigure shows RGB observations from the left, front, and right camera views in the top row, with the corresponding gradient-based saliency maps in the bottom row. Warmer colors indicate larger influence on the current value estimate.

D.3 Value-Function Visualization

Fig. 11 visualizes the inherited critic's value profiles during online execution. The successful episode receives consistently high value estimates, whereas the failure episode remains low and further drops near the failure state. This indicates that the inherited critic provides discriminative value guidance for online residual adaptation.

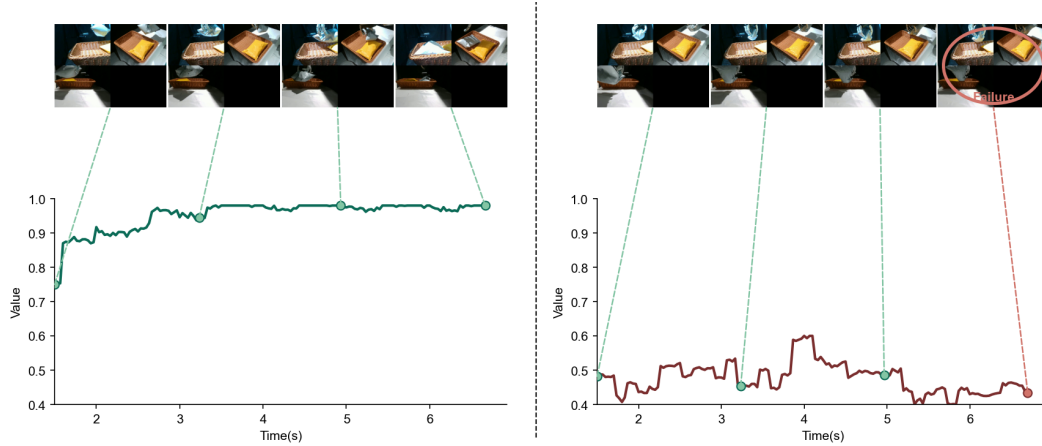


Figure 11: Value-function visualization of the inherited critic. The left panel shows a successful episode, where the critic maintains high value estimates throughout execution. The right panel shows a failure episode, where the value estimates remain lower and further drop near the failure state. Dashed lines connect representative frames to their corresponding value estimates.

E Hyperparameters and Implementation Details

The complete hyperparameter configurations for both the Phase 1 (Offline VLM pre-training) and Phase 2 (Online residual adaptation) are summarized in Table 3.

Table 3: Hyperparameter Configurations for Offline and Online Phases.

| Category | Hyperparameter | Value |
|--|---|---|
| Hardware Setup | Phase 1 (Offline) Cluster | 8 × NVIDIA H100 GPUs |
| | Phase 2 (Online) Workstation | 1 × NVIDIA RTX 4090 GPU |
| Phase 1: Offline VLM BC & PPO | Training Mode / Precision | Full Fine-tuning / BF16 |
| | Base Optimizer / Weight Decay | AdamW / 1×10^{-4} |
| | Total Offline Updates | 70,000 |
| | Batch Size per GPU | 8 |
| | Action Chunk Size (k) | 32 |
| | Actor VLM Learning Rate | 1×10^{-5} |
| | Actor Consistency Learning Rate | 5×10^{-5} |
| | PPO Optimization Start Step | 40,000 |
| | Behavior Cloning Coefficient (λ_{BC}) | 1.0 |
| | PPO Guidance Coefficient | 0.01 |
| Advantage Normalization Clip | 2.0 | |
| OPE Gate / Margin (δ_{OPE}) | Enabled / 0.0 | |
| Phase 2: Online Residual RLPD | Base Optimizer / Weight Decay | AdamW / 1×10^{-4} |
| | Max Online Updates | 15,000 |
| | Batch Size | 64 – 128 |
| | Action Chunk Size (k) | 32 |
| | Discount Factor (γ) | 0.99 |
| | Actor / Critic Learning Rate | 1×10^{-4} / 1×10^{-4} |
| | Offline-to-Online Batch Ratio | 1 : 1 |
| | Target Network Soft Update (τ_{target}) | 0.005 |
| | Initial / End BC Lambda | 2.0 / 1.0 (Decay over 10,000 steps) |
| | Critic Regularization (λ_Q) | 0.25 |
| | Residual Blend Factor (λ_t) | 0.2 → 0.75 (Warmup over 4,000 steps) |
| | Actor Anchor Lambda | 0.03 |
| | Conservative Improvement Margin (δ) | 0.0 |
| Conservative Improvement Lambda | 0.35 | |
| Actor Residual L_2 Regularization | 0.0002 | |

F Dataset Statistics and Task Metrics

Table 4 summarizes the task definitions, success criteria, and data statistics used in our real-world experiments. For each task, we report the number of offline trajectories used for offline token-action RL, together with the number of online intervention trajectories collected per adaptation round for BORA-Full. The success criteria are defined at the task level and are used consistently for both the standard and object-unseen evaluations.

Table 4: Task Definition, Success Criteria, and Dataset Statistics for Offline and Online Phases.

| Task | Offline Trajectories | Online Trajectories per Iteration (Ours) | Success Criteria |
|-------------------------|-----------------------------|---|---|
| Pick Plush Toy | 60 | 10 | Lift the plush toy completely off the tabletop surface and maintain a stable grasp. |
| Pick & Place | 100 | 10 | Successfully pick up the package, transfer it, and place it entirely inside the basket. |
| Open Box | 100 | 10 | Fully rotate and flip open one side of the box lid to a completely open state. |
| Pull Tissue | 60 | 10 | Securely pinch and extract a single whole sheet of tissue completely out of the box container. |
| Press Button | 60 | 10 | Actuate the button by applying stable downward force to fully depress it without finger slippage. |

G Failure Mode and Quantitative Analysis

Table 5 provides a qualitative summary of the typical failure modes observed across tasks, evaluation settings, and method families. The comparison shows how failure patterns change from pure imitation learning to offline RL baselines and BORA-Full.

Table 5: Qualitative Analysis of Typical Failure Modes Across Settings.

| Task | Setting | Pure IL Baselines | Offline RL Baselines | BORA-Full (Ours) |
|-------------------------|----------|---|---|--------------------------------|
| Pick Plush Toy | Standard | Minor rotation; unconfident grasp closure. | Minor object rotation during approach. | — |
| | Unseen | Severe loose grasp under shape variations. | Minor object rotation (same as Standard). | — |
| Pick & Place | Standard | Basket collision; placement hesitation. | Basket collision during transfer. | — |
| | Unseen | Grasp failure driven by novel shape/color. | Basket collision (same as Standard). | — |
| Open Box | Standard | Insufficient lifting; incomplete lid opening. | Missed lid contact; localized hesitation. | — |
| | Unseen | Insufficient lid opening (same as Standard). | Missed lid contact under layout shifts. | Collision with inner contents. |
| Pull Tissue | Standard | Insecure pinch, leading to tear/slip. | Off-center pinching; partial extraction. | Finger trembling; loose grip. |
| | Unseen | Missed pinch due to box dimension shifts. | Off-center pinching (same as Standard). | Finger trembling; loose grip. |
| Press Button | Standard | Insufficient downward pressing force. | Off-center contact, causing finger to slip. | — |
| | Unseen | Insufficient force (same as Standard). | Off-center contact and slipping. | — |