CUBic: Coordinated Unified Bimanual Perception and Control Framework

Supplementary Material

1. Overview

We begin by providing a granular and comprehensive description of the diverse manipulation tasks defined within the RoboTwin environment, including the specific success criteria and state-action spaces. To ensure full reproducibility of our results, we also elaborate on the detailed hyperparameter configurations and training protocols utilized in our experiments. Subsequently, we conduct a systematic series of ablation studies to quantitatively isolate and validate the contribution of each individual module in our proposed method, offering empirical evidence for our architectural design choices. Finally, we present qualitative visualizations of the task execution trajectories. These visual demonstrations not only illustrate the superior performance of our approach in practical scenarios but also highlight its robustness in handling complex coordination compared to baseline methods.

2. Task Description

We first provide a detailed description of the testing tasks, which are strategically categorized to evaluate distinct manipulation capabilities. Specifically, 'Pick Apple Messy' is designed to rigorously benchmark the model's aptitude for precise object localization and fine-grained grasping in cluttered environments. To assess synchronous coordination, we employ the 'Dual Bottles Pick' and 'Dual Shoes Place' tasks, which challenge the model's ability to perform collaborative precise grasping. Furthermore, the 'Blocks Stack (Easy)', 'Block Handover', and 'Put Apple Cabinet' tasks are utilized to evaluate advanced dual-arm collaborative manipulation, requiring complex temporal and spatial reasoning.

- **Pick Apple Messy**: Apples and four random items are places randomly on the table. The robotic arm picks up the apple and lifts it.
- **Dual Bottles Pick (Easy)**: A red bottle is placed randomly on the left side, and a green bottle is placed randomly on the right side of the table. Both bottles are standing upright. The left and right arms are used simultaneously to lift the two bottles to a designated location.
- **Blocks Stack (Easy)**: Red and black cubes are placed randomly on the table. The robotic arms stack the cubes in order, placing the red cubes first, followed by the black cubes, in the designated target location.
- **Dual Bottles Pick (Hard)**: A red bottle is placed randomly on the left side, and a green bottle is placed randomly on the right side of the table. The bottles are lying on their sides. Both left and right arms are used simultaneously to lift the two bottles to a designated location.
- **Block Handover**: A long block is placed on the left side of the table. The left arm grasps the upper side of the block and then hands it over to the right arm, which places the block on the blue mat on the right side of the table.
- **Dual Shoes Place**: One shoe is placed randomly on the left and right sides of the table. The shoes are the same pair with random designs that do not repeat in the training and testing sets. Both left and right arms are used to place them in the blue area with the shoe heads facing the left side of the table.
- **Put Apple Cabinet**: Initially, an apple is placed randomly. The robotic arm uses the left arm to open the cabinet and the right arm to pick up the apple and place it inside.

3. Hyperparameters

We hereby report the precise hyperparameter specifications used to train and evaluate RoboTwin in table 1. To accelerate the inference process while preserving generation quality, we employ the Denoising Diffusion Implicit Models (DDIM) sampler, which effectively reduces the number of required sampling steps. Regarding the quantization architecture, we enhance the reconstruction performance by integrating four hierarchical layers of residual quantization within the Residual VQ module. To address the instability often inherent in discrete representation learning, we opt for an Exponential Moving Average (EMA) mechanism to optimize the codebook vectors rather than direct gradient descent. Additionally, we configure a global EMA scheme for the entire network, maintaining a running average of the model parameters to smooth out fluctuations during optimization. Consequently, the EMA replica weights are retained and utilized for the final evaluation to ensure superior performance.

4. Ablation Study

We present comprehensive ablation results to validate the effectiveness of our proposed method. Specifically, we investigate the impact of the two-stage training strategy, the dual-codebook perceptual coordination mechanism, the number of latent tokens and codebook size. The results demonstrating the contribution of each component are detailed below.

4.1. Impact of two-stage training

To evaluate the effectiveness of our two-stage training paradigm integrating perception collaboration and control

Hyperparameter	Value
Observation Horizon (T_o)	1
Action Horizon (T_a)	8
Prediction Action Horizon (T_p)	8
Optimizer	AdamW
Betas (β_1, β_2)	[0.95, 0.999]
Learning Rate	1.0e-5
Weight Decay	1.0e-6
Pre-training epochs	900
Post-training epochs	900
Learning Rate Scheduler	Cosine
Training Timesteps (T)	100
Inference Timesteps	10
Prediction Type	ϵ-prediction
Image Resolution	240×320
RVQ Layer Number (K)	4
Noise Scheduler	DDIM

Table 1. **Hyperparameters used for RoboTwin.**

Task	DP3	DP	Ours (end-to-end)	Ours (2-stage)
Pick Apple	9.7	29.3	**45.0**	40.0
Bottles (Easy)	55.3	**85.7**	72.3	84.3
Stack (Easy)	-	8.0	10.0	**16.0**
Bottles (Hard)	58.0	**59.3**	54.1	58.0
Block Handover	77.3	76.0	56.7	**85.7**
Shoes Place	**12.0**	3.0	7.9	10.0
Apple Cabinet	66.3	8.0	48.7	**68.7**
Average (%)	39.8	38.5	42.1	**51.8**

Table 2. **Impact of two-stage training.** We report and compare the success rates across seven tasks. The best score is shown in **bold**, and the second-best is underlined.

collaboration, we compared the success rates of the end-to-end trained model across multiple tasks. As shown in the table 2, the end-to-end trained model has already acquired capabilities 42.1% comparable to those of DP 38.5% and DP3 39.8%, particularly surpassing baseline performance in more challenging tasks such as "Pick Apple Messy" and "Blocks Stack Easy". Furthermore, our control collaborative post-training yields a 9.7% improvement in accuracy relative to end-to-end trained model alone, demonstrating the substantial efficacy of our two-stage training approach.

4.2. Impact of shared-mapping for dual codebooks

We extensively evaluated the impact of the dual-codebook shared mapping mechanism on performance across seven tasks. As shown in Table 3, our method achieves a 11.7% improvement in accuracy compared to independent codebook encoding. Notably, we observe significant gains of 12.0% and 17.0% in the "Blocks Stack Easy" and "Put Apple Cabinet" tasks, respectively. These results underscore the importance of dual-codebook perceptual coordination for bimanual manipulation tasks.

Task	DP3	DP	Ours (Ind.)	Ours (Share)
Pick Apple	9.7	29.3	28.0	**40.0**
Bottles (Easy)	55.3	**85.7**	72.0	84.3
Stack (Easy)	-	8.0	4.0	**16.0**
Bottles (Hard)	58.0	**59.3**	30.0	58.0
Block Handover	77.0	76.0	**90.0**	85.7
Shoes Place	**12.0**	3.0	5.0	10.0
Apple Cabinet	66.3	8.0	51.7	**68.7**
Average (%)	39.8	38.5	40.1	**51.8**

Table 3. **Impact of shared mapping for dual codebooks.** We report the success rates of our method with two independent codebooks (Ind.). The best score is shown in **bold**, and the second-best is underlined.

Task	DP3	DP	Ours (w/o latent)	Ours (8, 512)	Ours (4, 256)
Pick Apple	9.7	29.3	0.0	38.0	**40.0**
Bottles (Easy)	55.3	**85.7**	0.0	58.0	84.3
Stack (Easy)	-	8.0	0.0	10.0	**16.0**
Bottles (Hard)	58.0	**59.3**	0.0	56.0	58.0
Block Handover	77.3	76.0	0.0	70.0	**85.7**
Shoes Place	**12.0**	3.0	0.0	6.0	10.0
Apple Cabinet	66.3	8.0	0.0	43.7	**68.7**
Average (%)	39.8	38.5	0.0	40.2	**51.8**

Table 4. **Impact of latent tokens and codebook size.** We report the success rates of our method with different numbers of latent tokens and codebook sizes. The best score is shown in **bold**, and the second-best is underlined.

4.3. Impact of latent tokens and codebook size

We provide a detailed performance comparison between the absence of latent tokens and varying configurations of latent token counts with their corresponding codebook sizes. As shown in the table 4, the model fails to function without latent tokens serving as a bridge. Furthermore, performance degrades when the number of latent tokens becomes excessive. These results underscore the critical importance of our specific parameter selection for optimal model performance.

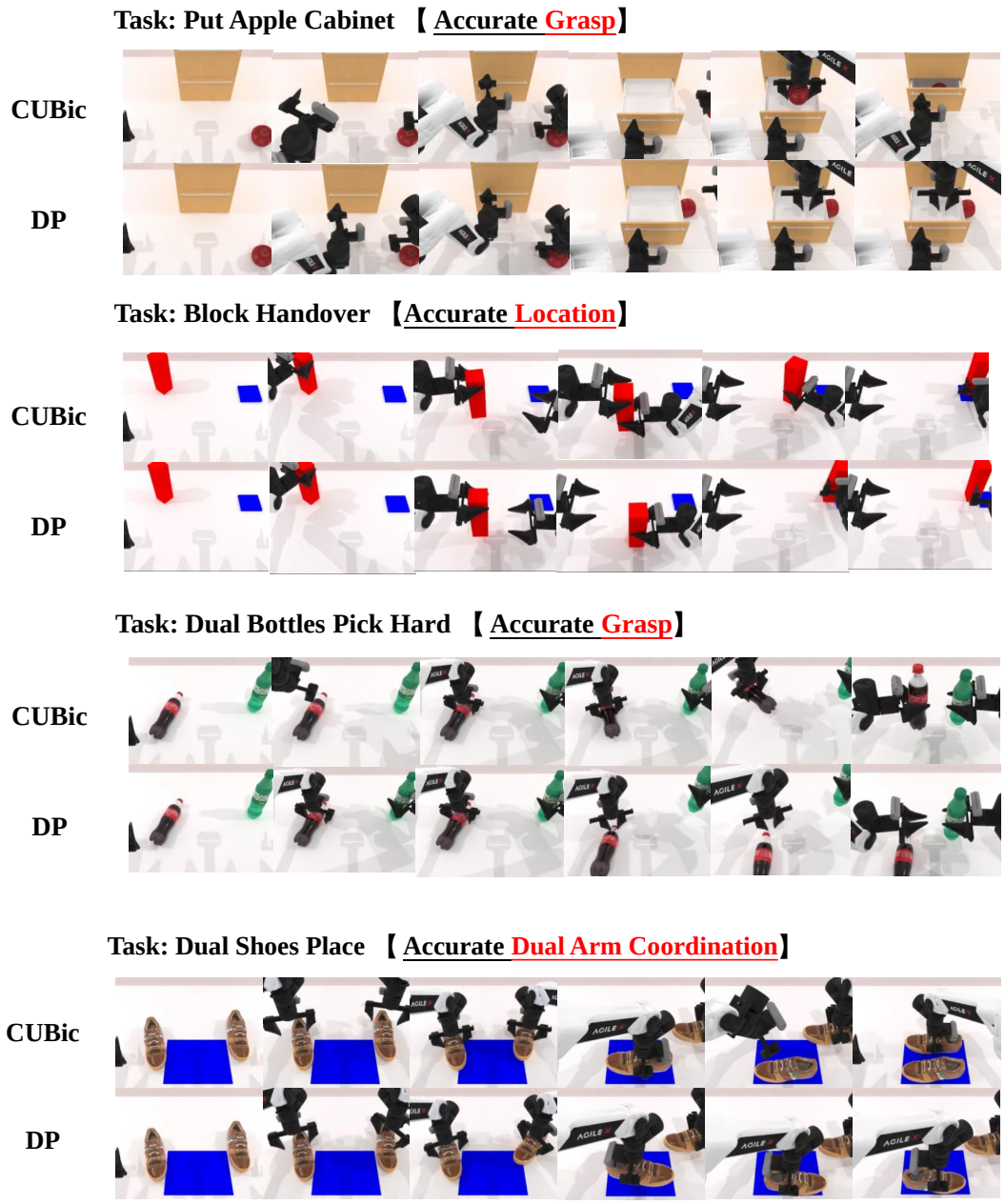

Figure 1. **Visualization in RoboTwin.** As illustrated, compared with Diffusion Policy (DP), CUBic demonstrates superior performance in target localization and precise object grasping, while maintaining strong coordination in bimanual manipulation scenarios.

5. Visualization

We visualize the execution processes of CUBic and DP across four task scenarios in the RoboTwin in Figure 1. As observed, in bimanual manipulation tasks such as "Put Apple Cabinet," "Block Handover," and "Dual Shoes Place," our method achieves substantially higher precision than Diffusion Policy (DP) in several key stages of execution, including apple grasping, target localization, and dual-arm cooperative placement. In particular, our policy is able to identify the target object more reliably in cluttered scenes, approach it with more stable and accurate end-effector motions, and maintain better coordination between the two arms during handover and placement. By contrast, DP more frequently exhibits imprecise grasp poses, less consistent spatial alignment, and weaker inter-arm synchronization, which often leads to unstable manipulation outcomes or task failure.

These qualitative results further suggest that the performance gains of our method are not limited to final success rates, but are also reflected in the entire manipulation process. The improved behavior indicates that our framework learns a more coherent coupling between visual understanding and action generation, enabling the robot to react more effectively to scene variations while preserving smooth and coordinated control. Overall, these results demonstrate that our framework exhibits strong synergetic capabilities across both perception and control, which is particularly beneficial for complex bimanual manipulation tasks.