Diffusion-Based Imaginative Coordination for Bimanual Manipulation - Supplementary Material -

Huilin Xu[1*] Jian Ding[2] Jiakun Xu[3*] Ruixiang Wang[4*] Jun Chen[2] Jinjie Mai[2]
Yanwei Fu[1] Bernard Ghanem[2] Feng Xu[1] Mohamed Elhoseiny[2†]

[1]Fudan University [2]King Abdullah University of Science and Technology
[3]ETH Zurich [4]The Chinese University of Hong Kong, Shenzhen

The purpose of this supplementary material is to enhance the clarity and understanding of our proposed method by providing comprehensive details, additional experiments, in-depth analyses, and visualizations.

1. Differences between Our Model and GR-1

Our model differs fundamentally from GR-1 in both network architecture and training paradigm. Particularly, GR-1 focuses on multi-task generalization via large-scale video pretraining with an autoregressive GPT-style model. In contrast, we target spatio-temporal coordination using a diffusion-based model that jointly predicts future latents and actions without a pretraining stage. Our method also constructs the causal relationship between visual outcomes and action by action-conditioned attention mechanism. Besides, we predict future video latents rather than raw pixels used in GR-1 and show the advantage of latent prediction.

2. Additional Implementation Details

2.1. Implementation Details

We use DDIM [13] as the noise scheduler with a square cosine schedule [11], employing 100 diffusion steps during training and 10 steps during inference. Our model predicts the clean sample instead of epsilon. We utilize a pretrained Cosmos tokenizer [3] to extract visual tokens for future frames, using DV $4 \times 8 \times 8$ version with a compression ratio of 256. For ALOHA benchmark, we train for 20, 000 steps with a batch size of 32, while for RoboTwin benchmark, we train for 300 epochs with a batch size of 128. For real-world experiments, we train for 50 epochs with a batch size of 32. Across all experiments, we adopt a cosine learning rate scheduler with 500 linear warm-up steps to stabilize training.

Model	ALOHA	RoboTwin	Real-World
Image resolution	480×640	240×320	480×640
Backbone	Pretrained ResNet18	Pretrained ResNet18	Pretrained ResNet18
# encoder layer	4	4	4
# dncoder layer	7	7	7
Chunk size	100	20	40
# Predicted frames	40	20	40
Scheduler	DDIM	DDIM	DDIM
Prediction type	Sample	Sample	Sample
Diffusion steps	100	100	100
Diffusion steps (eval)	10	10	10
Noise scheduler	Squared_cosine	Squared_cosine	Squared_cosine
Tokenizer	Cosmos DV 4×8×8	Cosmos DV 4×8×8	Cosmos DV 4×8×8
Patch size	5	5	5
Training	ALOHA	RoboTwin	Real-world
Epochs	50	300	50
Batch size	32	256	32
Train/Validation ratio	4:1	9:1	49:1
Learning rate	5e-4	1e-4	5e-5
Lr scheduler	Cosine warmup	Cosine warmup	Cosine warmup
Optimizer	AdamW	AdamW	AdamW
Prediction weight	0.2	0.2	0.2
Image augmentation	RandomShift(15,20)	RandomShift(6,8)	RandomShift(15,20)

Table 1. **Hyperparameters of our method.**

2.2. Structure Details

We provide our architecture and hyperparameter setting details in three evaluate environments, as shown in Table 1. For normalization, we independently scale the minimum and maximum values of each action dimension and each video token dimension to the range $[-1, 1]$. Normalizing actions and tokens to $[-1, 1]$ is essential for DDPM and DDIM predictions, as these models clip their outputs to $[-1, 1]$ to ensure stability [6].

2.3. Baseline Implementations

To ensure fair comparison, we report baseline results on both the ALOHA and RoboTwin benchmarks based on either official publications or our own reproductions. For the ALOHA benchmark, we report ACT results directly from the original paper [16], and reproduce Diffusion Policy [2] using its publicly available code and default training settings. For the RoboTwin benchmark, we report the results

*Work was done during internship in KAUST.
†Corresponding author: mohamed.elhoseiny@kaust.edu.sa

	ALOHA	RoboTwin	Real-World
Chunk size	100	20	40
Batchsize	8	256	16
Learning rate	1e-5	1e-4	2e-5
Lr scheduler	Constant	Cosine warmup	Constant
Optimizer	AdamW	AdamW	AdamW

Table 2. **Hyperparameters of ACT.**

	ALOHA	RoboTwin	Real-World
Chunk size	100	8	40
N_obs_step	2	3	2
Batchsize	20	128	20
Learning rate	1e-4	1e-4	1e-4
Lr scheduler	Cosine warmup	Cosine warmup	Cosine warmup
Optimizer	AdamW	AdamW	AdamW

Table 3. **Hyperparameters of Diffusion Policy**.

of Diffusion Policy [2] and 3D Diffusion Policy [14] from the RoboTwin paper [10]. For other baselines, including ACT [16], GR-MG [8], and RDT-1B [9], we retrain them using their official implementations and configurations, and evaluate them under the same data and testing protocols as our method. The hyperparameters for ACT and Diffusion Policy are summarized in Table 2 and Table 3, respectively.

2.4. Cosmos Tokenizer

Cosmos Tokenizer is a core component of the Cosmos World Foundation Model Platform [3], designed to efficiently transform raw visual data (images and videos) into compact token representations. It supports both continuous and discrete tokenization, preserving spatio-temporal information while reducing computational costs. Designed with a causal architecture, it ensures that token computation depends only on past and current frames, making it well-suited for real-time and sequential tasks.

For an input video of shape $(1 + T, C, H, W)$, Cosmos Tokenizer compresses it based on the spatial compression factor s_{HW} and the temporal compression factor s_T, producing an output of shape $(1 + T/s_T, C, H/s_{HW}, W/s_{HW})$. The first temporal token represents the first input frame, while subsequent tokens capture temporal dependencies. Spatially, the feature map is downsampled by a factor of s_{HW}, resulting in a reduced resolution of $(H/s_{HW}, W/s_{HW})$.

Fig. 1 illustrates video reconstruction of the pretrained Cosmos Tokenizer on two simulated benchmarks. Since the Cosmos tokenizer is trained in various domains such as robotics, driving, egocentric, and web videos, it demonstrates strong generalization capabilities. This makes it a suitable choice as a plug-and-play module for compressing video while retaining information-rich features.

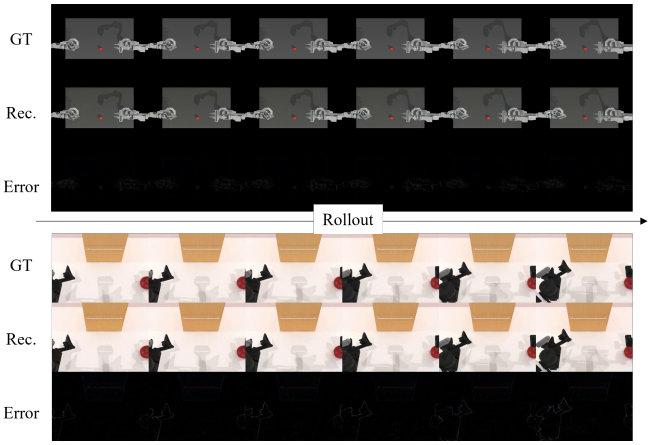


Figure 1. **Visualization of video reconstruction** using Cosmos Tokenizer on ALOHA [16] and RoboTwin [10] benchmark (17 frame sub-clip, DV4x8x8 version).

Method	Avg. SR ↑	Transfer Cube (Human)	Insertion (Human)
InterACT [7]	<u>63.0</u>	82	<u>44</u>
ARP [15]	59.4	**94**	24.8
Ours	**65.3**	<u>84</u>	**46.7**

Table 4. **Comparison with more transformer-based methods.**

3. Additional Experimental Results

3.1. Comparison with Transformer-Based Baselines

We compare our method against more transformer-based baselines, including InterACT [7] and ARP [15], as shown in Table 4. For a fair comparison, we report the original results of InterACT and ARP from their respective papers and retrain our model under the same experimental settings. All results are averaged over 3 random seeds with 50 evaluation episodes each. Our method achieves the highest average success rate (Avg. SR) of 65.3, consistently outperforming both InterACT and ARP.

3.2. Results in Multi-task setting

To evaluate the effectiveness of our method in more complex multi-task scenarios, we extend our single-task framework to a language-conditioned multi-task policy, inspired by prior work [1]. Specifically, we incorporate FiLM-based conditioning [12] to inject task descriptions into the policy. We construct a multi-task benchmark using three representative bimanual tasks from RoboTwin and compare our method against RDT-1B [9]. As shown in Table Tab. 6, our approach consistently outperforms RDT-1B, highlighting its strong ability in complex dual-arm manipulation.

3.3. Detailed Result on Data Efficiency Setting

We evaluate methods under the data efficiency setting, shown in Table 5. The performance of all methods de-

Method	Avg. Success ↑	Block Hammer Beat	Block Handover	Blocks Stack (Easy)	Blocks Stack (Hard)	Bottle Adjust	Container Place	Diverse Bottles Pick	Dual Bottles Pick (Easy)
3D Diffusion Policy [14]	**30.4**	55.7	**89**	-	-	**64.7**	**52.7**	**11.3**	**40.3**
Diffusion Policy [2]	1.5	0.0	0.0	0.0	0.0	6.3	1.7	0.7	1.7
ACT [16]	15.4	45.3	65.67	3.67	0.0	38.33	9.67	0.7	27.7
Ours	27.3	**60.33**	88.33	**4.33**	**0.33**	38.33	40.33	1.33	32.33

Method	Dual Bottles Pick (Hard)	Dual Shoes Place	Empty Cup Place	Mug Hanging (Easy)	Mug Hanging (Hard)	Pick Apple Messy	Put Apple Cabinet	Shoe Place	Coord. Avg. ↑
3D Diffusion Policy [14]	31.7	**4.0**	**33.7**	**7.3**	**4**	4	50.0	**38**	25.1
Diffusion Policy [2]	8.0	0.0	0.0	0.0	12.0	5.3	0.0	0.0	0.0
ACT [16]	17.0	2.7	3.0	0.0	0.0	7.0	14.33	12	13.9
Ours	**34.33**	3.67	16.67	0.67	1.0	**8.67**	**76.0**	29.33	**28.4**

Table 5. **Evaluation on data efficiency setting.** We report the mean of success rates averaged over 3 random seeds. Best score in **bold**, second-best underlined. Coord. Avg. denotes the averaged success rate of tasks in the coordinated subset.

Method	Avg. ↑ Success (%) ↑	Diverse Bottle Pick	Blocks Stack (Easy)	Put Apple Cabinet
RDT-1B [9]	54.0	14	**66**	82
Ours	**62.0** (+8.0)	**24**	62	**100**

Table 6. **Performance comparison under multi-task setting.** We extend our methods to a multi-task policy and beats RDT-1B [9].

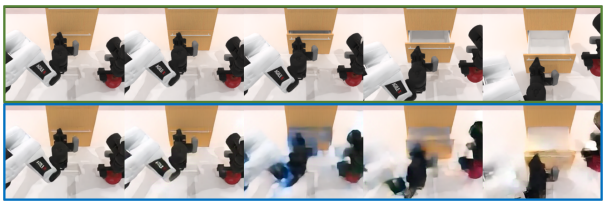


Figure 2. Visualization of video prediction results.

grades compared to the default setting due to the limited training data. Among the baselines, 3D Diffusion Policy achieves the highest overall success rate, benefiting from its 3D point cloud representation, which has been shown to exhibit strong sample efficiency. Our method achieves the best result among all 2D policy models. In *Seq-coordinate* tasks, our method even outperforms 3D DP, demonstrating its effectiveness in capturing sequential dependencies on the data efficiency setting.

3.4. Visualization of Video Prediction

We visualize predicted frames in Fig. 2. While not photorealistic, our method generates semantically meaningful predictions (e.g., drawer opening) that reflect task-relevant dynamics. We respectfully clarify that our model is not designed to generate photorealistic frames, but rather to capture key task dynamics (e.g., drawer opening) for robotic action prediction. As prior work has shown that accurate modeling of task-relevant latent dynamics is more important than pixel-level reconstruction for effective control, e.g. TD-MPC2 [4], MPI [5]. We do not explicitly optimize for video quality, as our objective is not visual fidelity but task-relevant dynamics

3.5. Impact of Video Token Type

We also study the impact of various video tokens from Cosmos-Tokenizer[3]. As shown in Tab. 7, two different video tokens, discrete tokens and continuous tokens are implemented based on our model and the best results are reported for both tokens with same training seed. Our empirical analysis demonstrates that discrete tokens outperform continuous tokens in action prediction. Notably, this finding contrasts with results from video prediction studies, indicating that the underlying factors merit further investigation in future research.

Video Token	Avg. Success ↑	Transfer Cube Scripted	Transfer Cube Human	Insertion Scripted	Insertion Human
DV	73.7	98	78	79	40
CV	70.7	95	77	86	25

Table 7. **Ablation of video token.** DV denotes discrete video token and CV denotes continuous video token.

4. Additional Task Details

4.1. Task Description

For details of tasks in the ALOHA [16] and RoboTwin [10] simulation benchmarks, please refer to their original papers. Table 8 provides a comprehensive overview of the real-world benchmark robot tasks, illustrated in Fig. 3.

4.2. Task Category

We categorize the bimanual tasks in RoboTiwn benchmark into three types:

- **Dominant-select** – The executing arm is chosen based on the object's position. Tasks include:

Task	#Steps	Task Description
Water Wipe	500	Lift a bottle and wipe the exposed area with a cloth.
Coffee Stir	350	Pick up a cup and a pen, then stir inside the cup.
Cup Stack	400	Grasp both cups, place the right one first, then stack the left.
Can Handover	450	The left arm hands a can to the right arm, which places it.

Table 8. **Real-World task descriptions**

  - *Block Hammer Beat, Empty Cup Place, Pick Apple Messy, Shoe Place, Bottles Ajust, Container Place*
- **Sync-bimanual** – Both arms operate independently but simultaneously. Tasks include:
  - *Diverse Bottle Pick, Dual Bottles Pick (Easy), Dual Bottles Pick (Hard), Dual Shoes Place*
- **Seq-coordinate** – Tasks require sequential coordination with temporal dependencies. Tasks include:
  - *Block Handover, Blocks Stack (Easy), Blocks Stack (Hard), Mug Hanging (Easy), Mug Hanging (Hard), Put Apple Cabinet*

References

[1] Homanga Bharadhwaj, Jay Vakil, Mohit Sharma, Abhinav Gupta, Shubham Tulsiani, and Vikash Kumar. Roboagent: Generalization and efficiency in robot manipulation via semantic augmentations and action chunking. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4788–4795. IEEE, 2024. 2

[2] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023. 1, 2, 3

[3] NVIDIA et. al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 1, 2, 3

[4] Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024. 3

[5] Zeng Jia, Bu Qingwen, Wang Bangjun, Xia Wenke, Chen Li, Dong Hao, Song Haoming, Wang Dong, Hu Di, Luo Ping, Cui Heming, Zhao Bin, Li Xuelong, Qiao Yu, and Li Hongyang. Learning manipulation by predicting interaction. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024. 3

[6] Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. *arXiv preprint arXiv:2402.10885*, 2024. 1

[7] Andrew Choong-Won Lee, Ian Chuang, Ling-Yuan Chen, and Iman Soltani. Interact: Inter-dependency aware action chunking with hierarchical attention transformers for bimanual manipulation. In *8th Annual Conference on Robot Learning*, 2024. 2

[8] Peiyan Li, Hongtao Wu, Yan Huang, Chilam Cheang, Liang Wang, and Tao Kong. Gr-mg: Leveraging partially-annotated data via multi-modal goal-conditioned policy. *IEEE Robotics and Automation Letters*, 2025. 2

[9] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024. 2, 3

[10] Yao Mu, Tianxing Chen, Shijia Peng, Zanxin Chen, Zeyu Gao, Yude Zou, Lunkai Lin, Zhiqiang Xie, and Ping Luo. Robotwin: Dual-arm robot benchmark with generative digital twins (early version). *arXiv preprint arXiv:2409.02920*, 2024. 2, 3

[11] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 1

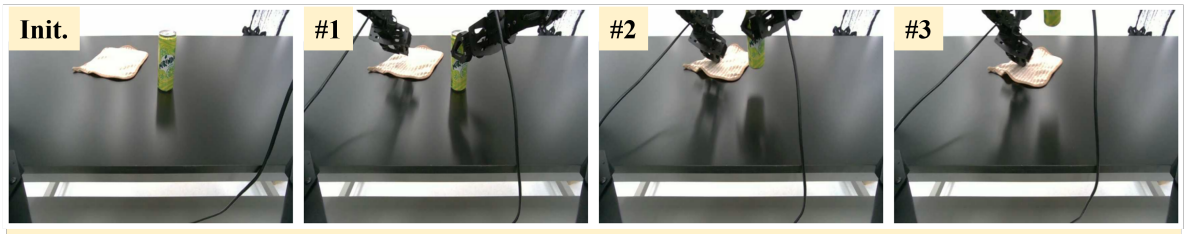
[12] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 2

[13] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 1

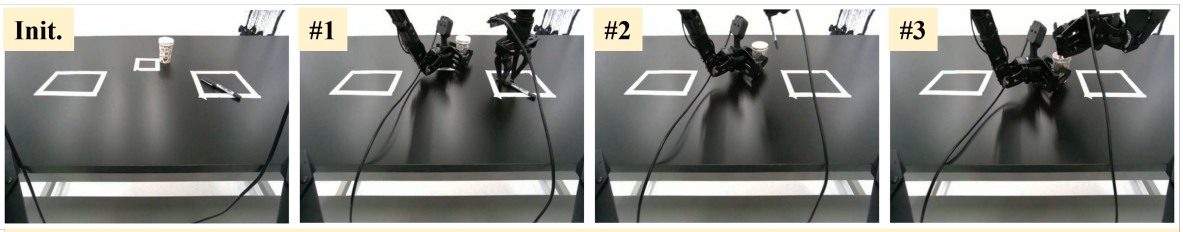
[14] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. In *Proceedings of Robotics: Science and Systems (RSS)*, 2024. 2, 3

[15] Xinyu Zhang, Yuhan Liu, Haonan Chang, Liam Schramm, and Abdeslam Boularias. Autoregressive action sequence learning for robotic manipulation. *arXiv preprint arXiv:2410.03132*, 2024. 2

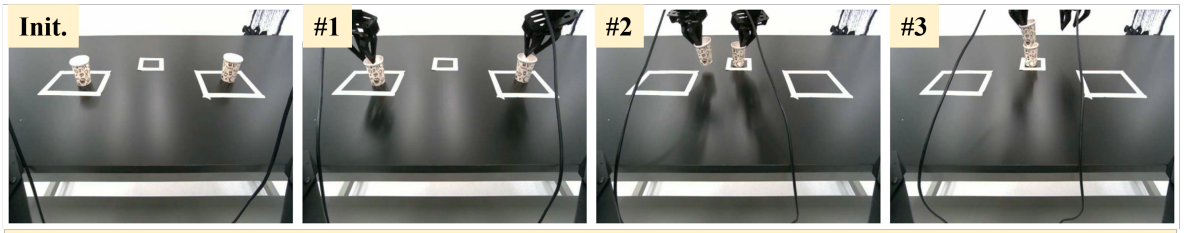
[16] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023. 1, 2, 3



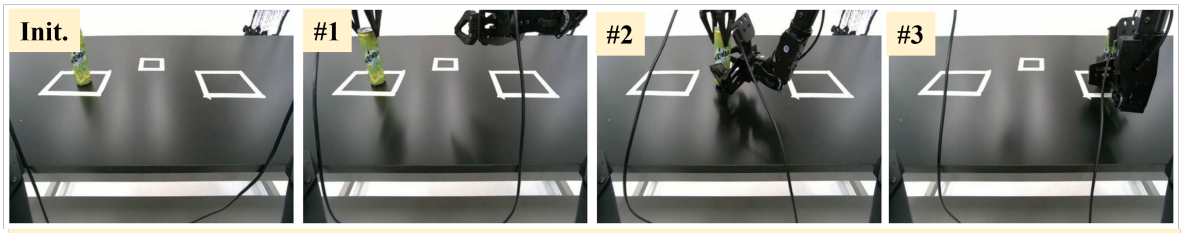
Water Wipe: Initially a bottle and a cloth are randomly placed on the table. Each robot arm grasps the nearest object (#1). The arm holding the bottle lifts it (#2), while the other arm wipes the exposed area with the cloth (#3). Each demo has 500 steps or 10 s.



Coffee Stir: Initially a cup and a pen are randomly placed on the table. Each robot arm grasps the corresponding object (#1). The right arm adjusts the pen to an upright position (#2) and places it inside the cup, performing a stirring motion (#3). Each demo has 350 steps or 7 s.



Cup Stack: Initially two cups are randomly placed on the table. The robot arms simultaneously grasp both cups (#1). The right arm places its cup in the designated area (#2), and the left arm stacks its cup on top (#3). Each demo has 400 steps or 8 s.



Can Handover: Initially a can is randomly placed on the table. The left arm grasps the upper side of the can (#1), then hands it over to the right arm (#2), which places it in the designated target position (#3). Each demo has 450 steps or 9 s.

Figure 3. **Task definition of real-world experiments.**