

1 This document contains supplementary materials for our main paper. We provide further technical
2 details, additional experimental results, and more qualitative examples to complement the findings
3 presented in the main text. We hope this supplementary information will help readers better understand
4 our approach and results.

5 The remainder of this supplementary material is organized as follows. In Appendix A, we provide
6 the hardware specifications used in our experiments. In Appendix B, we list the hyperparameters
7 employed. Section Appendix C presents the detailed algorithms for training and inference. In Ap-
8 pendix D, we describe the ablation studies conducted. Section Appendix E outlines the details of our
9 real-world experiments. In Appendix F, we offer further discussion on VLA models and frequency
10 domain analysis. Finally, in Appendix G, we discuss the limitations of our approach.

11 A Computational Resources

12 To ensure reproducibility, we provide detailed information on the computational resources used in
13 our experiments. For all simulation environment experiments including training, inference, and time
14 benchmarking tests, we used NVIDIA RTX 2080Ti GPUs. Our model has 63M parameters, with DP3
15 at 255M, consuming approximately 4.5GB of memory during operation. For real-world environment
16 experiments, we employed NVIDIA RTX 4090 GPUs for training, inference, and time benchmarking
17 tests.

18 B Hyperparameters

19 In Table 1, we present the hyperparameters used in our experiments. For the baseline methods DP
20 and DP3, we use their default hyperparameters. For the Adroit, DexArt, and MetaWorld benchmarks,
21 our models are trained for 3,000 epochs. For the Robomimic and Push-T tasks, we use 1,000 training
22 epochs, and for the RoboTwin benchmark, our models are trained for 500 epochs.

Table 1: Hyperparameters used for various benchmark.

Hyperparameter	Value
Horizon (T_h)	16 (8 for RoboTwin)
Action step (T_a)	8 (6 for RoboTwin)
Observation step (T_o)	2 (3 for RoboTwin)
point_feature_dim	64
state_mlp_size	64
Batchsize	128
Num_iter	4
Num_training_steps(Diffusion training)	100
Num_sampling_steps(Diffusion sampling)	ddim10
Diffloss_d	3
Diffloss_w	1024
encoder_embed_dim	512
decoder_embed_dim	512
encoder_depth	4
decoder_depth	4
encoder_num_heads	8
decoder_num_heads	8
Optimizer	AdamW
Betas (β_1, β_2)	[0.95, 0.999]
Learning Rate	1.0e-4
Weight Decay	1.0e-6
Learning Rate Scheduler	Cosine

23 C Training and Inference Details

24 The training process for our FreqPolicy is outlined in Algorithm 1. At each epoch, we first encode
 25 the input observations using an observation encoder. The ground truth action sequence is then
 26 transformed into the frequency domain via the Discrete Cosine Transform (DCT), and a frequency
 27 index is randomly sampled. Conditional reconstruction is performed by applying the inverse DCT
 28 up to the sampled frequency level, enabling the model to focus on different frequency components
 29 during training. An adaptive mask ratio is determined based on the frequency index, and a mask
 30 is sampled accordingly. The masked observation and conditional reconstruction are then encoded
 31 and subsequently decoded by the FreqPolicy encoder and decoder, respectively. The diffusion
 32 model is trained to predict noise added to the actions at randomly sampled diffusion steps, using a
 33 standard mean squared error loss. Model parameters are updated by minimizing this loss throughout
 34 the training epochs. This procedure effectively leverages masked autoregressive modeling and
 35 diffusion-based generation, enabling the policy to learn robust representations across the frequency
 36 domain.

Algorithm 1 FreqPolicy Training

Require: Number of training epochs K , Observation Encoder \mathcal{E}_{obs} , FreqPolicy Encoder \mathcal{E} , FreqPolicy Decoder \mathcal{D} , diffusion model ϵ_θ , Observation \mathcal{O} , Ground truth action \mathbf{x} , Horizon T , diffusion steps T_{diff} , initial mask ratio m .

```

1: for  $e = 1$  to  $K$  do
2:    $z_{obs} \leftarrow \mathcal{E}_{obs}(\mathcal{O})$  ▷ Encode observations
3:    $\{X_0, X_1, \dots, X_{T-1}\} \leftarrow \text{DCT}(\mathbf{x}), k \sim \mathcal{U}(0, T)$  ▷ Apply DCT, sample index
4:    $y^k \leftarrow \begin{cases} \text{IDCT}(\{X_0, X_1, \dots, X_{k-1}\}) & \text{if } k > 0 \\ \mathbf{0} & \text{if } k = 0 \end{cases}$  ▷ Extract k-level reconstruction
5:    $\text{mask\_ratio} \leftarrow m \cdot (1 - k/T)$  ▷ Adaptive mask ratio
6:    $\text{mask} \sim \text{TruncNorm}(\text{mask\_ratio})$ 
7:    $z_{mask} \leftarrow \mathcal{E}(z_{obs}, y^k, k, \text{mask})$  ▷ Encode with mask
8:    $z^k \leftarrow \mathcal{D}(z_{obs}, z_{mask}, k, \text{mask})$  ▷ Decode with mask
9:    $t \sim \mathcal{U}(1, T_{diff}), \epsilon \sim \mathcal{N}(0, \mathbf{I})$ 
10:   $\mathbf{x}_t \leftarrow \sqrt{\bar{\alpha}_t} \mathbf{x} + \sqrt{1 - \bar{\alpha}_t} \epsilon$ 
11:   $\mathcal{L} \leftarrow \mathbb{E}_{\epsilon, t} [\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t, k, z^k)\|^2]$  ▷ Diffusion loss
12:  Update model parameters by minimizing  $\mathcal{L}$ 
13: end for
14: return Trained models  $\mathcal{E}_{obs}, \mathcal{E}, \mathcal{D}$ , and  $\epsilon_\theta$ 

```

37 The inference process for FreqPolicy is detailed in Algorithm 2. Given an input observation, we
 38 first encode it using the observation encoder. The action tokens are initialized to zeros and are fully
 39 masked at the beginning. For each iteration, the model progressively predicts actions at increasing
 40 frequency levels, guided by the current frequency index.

41 At each step k , the partially reconstructed tokens and the current mask are passed through the encoder
 42 and decoder to obtain the continuous latent code z^k . The diffusion sampler then generates an updated
 43 action prediction conditioned on z^k . If it is not the final step, the predicted actions are transformed
 44 into the frequency domain using the Discrete Cosine Transform (DCT), and the reconstruction is
 45 refined up to the next frequency level via inverse DCT. The masking ratio is adaptively reduced at
 46 each iteration, gradually revealing more of the reconstructed action sequence. This process continues
 47 until the entire action sequence is fully reconstructed. By progressively incorporating information
 48 across different frequency bands, our inference procedure enables the policy to generate high-fidelity
 49 predictions.

50 D Ablation

51 To verify the effectiveness of our proposed method and the contribution of each component, we
 52 conducted a series of ablation experiments.

53 The results in Table 2 present the ablation study on the prediction horizon, it shows that the perfor-
 54 mance is not sensitive to this parameter, demonstrating the robustness of our method. When $T_h = 8$

Algorithm 2 FreqPolicy Inference

Require: Observation Encoder \mathcal{E}_{obs} , FreqPolicy Encoder \mathcal{E} , FreqPolicy Decoder \mathcal{D} , DiffusionSampler \mathcal{F} , diffusion steps T_{diff} , Observation \mathcal{O} , Horizon T , Number of iterations N_{iter} , Frequency indices $\{i_0, i_1, \dots, i_{N_{iter}-1}\}$.

```
1: Initialize:
2:   mask  $\leftarrow \mathbf{1}_{B \times T}$  ▷ Full masking initially
3:   tokens  $\leftarrow \mathbf{0}_{B \times T \times D_{action}}$  ▷ Zero initialization
4:    $z_{obs} \leftarrow \mathcal{E}_{obs}(\mathcal{O})$  ▷ Encode observations
5: for step = 0 to  $N_{iter} - 1$  do
6:    $k \leftarrow i_{step}$  ▷ Current frequency index
7:    $z_{mask} \leftarrow \mathcal{E}(z_{obs}, \text{tokens}, k, \text{mask})$ 
8:    $z^k \leftarrow \mathcal{D}(z_{obs}, z_{mask}, k, \text{mask})$ 
9:    $\hat{x} \leftarrow \mathcal{F}(z^k, k, T_{diff})$  ▷ Generate prediction via diffusion
10:  if step <  $N_{iter} - 1$  then
11:     $\{X_0, X_1, \dots, X_{T-1}\} \leftarrow \text{DCT}(\hat{x})$  ▷ Transform to frequency domain
12:    next_k  $\leftarrow i_{step+1}$  ▷ Next frequency level
13:    tokens  $\leftarrow \text{IDCT}(\{X_0, X_1, \dots, X_{next\_k-1}\})$ 
14:  else
15:    tokens  $\leftarrow \hat{x}$ 
16:  end if
17:  mask_ratio  $\leftarrow \cos\left(\frac{\pi}{2} \cdot \frac{step+1}{N_{iter}}\right)$ 
18:  mask  $\leftarrow \text{GenerateMask}(\text{mask\_ratio})$ 
19: end for
20: return tokens ▷ Final action sequence
```

55 or $T_h = 16$, the model achieves slightly better performance with an average score over 58 points,
56 while both smaller and larger horizons yield comparable results. This robustness across a range of
57 temporal scales highlights the flexibility of our approach. Notably, only at extremely long horizons
58 ($T_h = 64$) do we observe a noticeable decline in performance.

59 Table 3 presents the results of the ablation study on masking strategies. Our frequency policy
60 mask significantly improves performance across all tasks, increasing the average score from 40 to
61 58—an improvement of approximately 45%. These results clearly demonstrate the importance of the
62 frequency-based masking strategy for prediction tasks.

Table 2: **Ablation study on prediction horizon.** Analysis of Horizon (T_h), Action step (T_a), and Observation step (T_o).

T_h	T_o	T_a	Hammer	Door	Pen	Pick Out of Hole	Soccer	Stick Pull	Average
4	2	1	65±6	76±3	55±5	37±4	23±4	60±2	53±17
4	2	2	62±4	71±4	53±6	31±3	38±3	64±2	53±14
8	2	4	100±0	68±2	52±4	25±2	38±4	62±0	58±24
8	2	6	100±0	72±4	50±5	29±3	31±2	64±3	58±25
16	2	8	100±0	65±5	59±5	30±2	32±4	62±0	58±23
16	2	12	100±0	59±4	51±3	35±2	35±4	55±5	56±21
32	2	16	98±2	58±5	38±2	34±3	19±2	52±4	50±25
64	2	32	80±4	35±4	42±2	35±5	32±3	38±4	44±17

Table 3: **Ablation study on mask.** This experiment analyzes the model performance with and without mask.

	Hammer	Door	Pen	Pick Out of Hole	Soccer	Stick Pull	Average
W/o mask	99±1	35±4	31±3	15±0	27±2	34±2	40±27
Freqpolicy	100±0	65±5	59±5	30±2	32±4	62±0	58±23

Table 4: **Main results on 48 simulation tasks.** Results for each task are provided in this table.

Alg \ Task	Meta-World [6] (Easy)									
	Button Press	Button Press Wall	Coffee Button	Dial Turn	Door Close	Reach Wall	Door Open	Door Unlock	Drawer Close	Drawer Open
DP3	100±0	99±1	100±0	66±1	100±0	68±3	99±1	100±0	100±0	100±0
Diffusion Policy	99±1	97±3	99±1	63±10	100±0	59±7	98±3	98±3	100±0	93±3
DP3*	100±0	100±0	100±0	58±5	100±0	47±5	100±0	100±0	100±0	100±0
Mamba Policy*	100±0	100±0	100±0	56±4	100±0	50±3	100±0	100±0	100±0	100±0
ours	100±0	100±0	100±0	72±4	100±0	71±4	100±0	100±0	100±0	100±0

Alg \ Task	Meta-World (Easy)									
	Faucet Open	Handle Press	Lever Pull	Plate Slide	Plate Slide Back	Plate Slide Back Side	Plate Slide Side	Reach	Window Close	Window Open
DP3	100±0	100±0	79±8	100±1	99±0	100±0	100±0	24±1	100±0	100±0
Diffusion Policy	100±0	81±4	49±5	83±4	99±0	100±0	100±0	18±2	100±0	100±0
DP3*	100±0	86±5	84±2	100±0	100±0	100±0	100±0	22±4	100±0	100±0
Mamba Policy*	100±0	83±5	74±6	100±0	100±0	100±0	100±0	17±3	100±0	100±0
ours	100±0	90±3	84±4	100±0	100±0	100±0	100±0	30±2	100±0	100±0

Alg \ Task	Meta-World (Medium)									
	Hammer	Peg Insert Side	Push Wall	Soccer	Sweep	Sweep Into	Basketball	Bin Picking	Box Close	Coffee Pull
DP3	76±4	69±7	49±8	18±3	96±3	15±5	98±2	34±30	42±3	87±3
Diffusion Policy	15±6	34±7	20±3	14±4	18±8	10±4	85±6	15±4	30±5	34±7
DP3*	80±5	62±5	87±5	22±3	100±0	17±2	100±0	35±10	50±6	95±0
Mamba Policy*	90±5	63±4	92±2	28±3	95±5	15±3	100±0	26±3	48±10	96±2
ours	96±1	51±4	97±3	32±4	85±4	19±5	83±8	31±2	56±4	100±0

Alg \ Task	Meta-World (Hard)					Meta-World (Very Hard)				
	Assembly	Hand Insert	Pick Out of Hole	Pick Place	Push	Shelf Place	Disassemble	Stick Pull	Stick Push	Pick Place Wall
DP3	99±1	14±4	14±9	12±4	51±3	17±10	69±4	27±8	97±4	35±8
Diffusion Policy	15±1	9±2	0±0	0±0	30±3	11±3	43±7	11±2	63±3	5±1
DP3*	95±3	11±2	12±7	42±6	54±4	31±4	87±3	57±3	83±3	82±6
Mamba Policy*	96±2	15±3	25±5	36±10	60±4	38±4	90±2	55±2	82±5	80±5
ours	97±2	17±3	30±2	37±4	63±2	27±3	92±6	62±0	85±5	85±3

Alg \ Task	Adroit [5]			DexArt [1]			Average (41+3+4)	
	Hammer	Door	Pen	Laptop	Faucet	Toilet	Bucket	
DP3	100±0	62±4	43±6	83±1	63±2	82±4	46±2	71.36
Diffusion Policy	45±5	37±2	13±2	69±4	23±8	58±2	46±1	53.25
DP3*	100±0	53±2	50±5	83±3	33±2	70±6	24±4	72.91
Mamba Policy*	100±0	59±3	55±2	79±3	35±6	65±5	23±2	73.71
ours	100±0	65±5	59±5	85±4	30±3	77±3	25±3	75.54

63 E Real-world Experiment Details

64 In real-world experiments, we employ an RGB-D camera (Kinect) to capture environmental point
65 clouds at a rate of 30 Hz. The FoundationPose model serves as our object pose estimation module,
66 which processes the point clouds from the RGB-D camera to predict the 6D object pose, achieving a
67 93.22% ADD AUC score on the YCBInEOAT dataset at 30 FPS. By sampling 4,096 points from the
68 object’s mesh and applying the estimated pose from FoundationPose, we obtain cleaned object point
69 clouds for the past 5 frames ($T_0 = 5$).

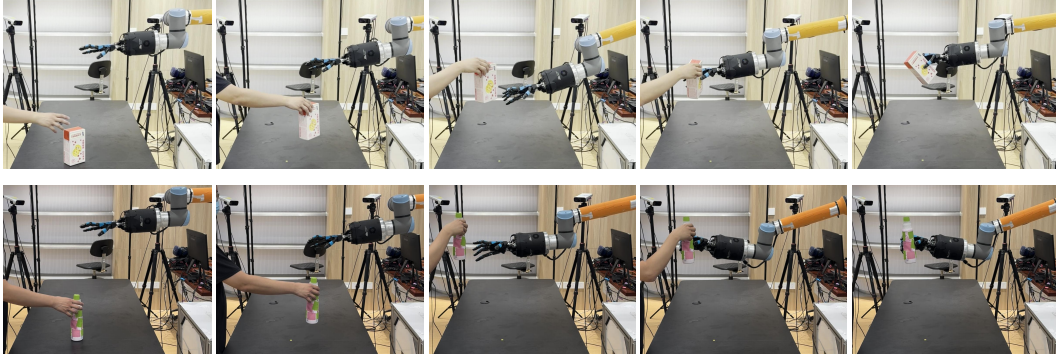


Figure 1: **Additional Visualization of Real-World Experiments on 2 Robotic Handover Tasks.**

70 Our algorithm takes these cleaned object point clouds along with the shadow hand poses from the
71 previous 5 frames as input conditions, and predicts the hand’s poses for the next 3 frames ($T_a = 3$).
72 These predicted poses are executed directly on the physical robot, enabling a fully end-to-end action
73 generation system. In terms of system performance, the perception module operates at approximately
74 30 FPS, while the action prediction module achieves over 70 FPS with a single iteration setting. The
75 complete end-to-end system maintains a comprehensive operating rate of approximately 25 FPS,

76 meeting real-time interaction requirements. Figure 1 shows the results of two additional sets of
77 handover tasks.

78 F Discussion

79 F.1 Further Discussion on Vision-Language-Action (VLA) Models

80 In the main sections of this paper, we have detailed and validated the superior performance of
81 FreqPolicy in learning specific robotic manipulation tasks. Its innovative frequency-domain autore-
82 gressive mechanism and the use of continuous tokens have demonstrated significant advantages in
83 both precision and efficiency for single-task learning. However, a natural and promising extension
84 is to investigate the adaptability of FreqPolicy in more complex and generalized multitask learning
85 scenarios, especially when task instructions are given in natural language. Vision-Language-Action
86 (VLA) models provide a powerful framework for achieving such language-driven multitask robot
87 control.

88 Therefore, this section aims to conduct a preliminary exploration of FreqPolicy’s potential when
89 applied to VLA models. Given that policy learning in multitask environments is still a multifaceted
90 and challenging research problem, we do not aim to provide a solution here. Instead, our goal is to
91 assess FreqPolicy on a multitask benchmark and to provide a forward-looking discussion on whether
92 the core ideas of FreqPolicy, such as frequency-domain decomposition and hierarchical learning, can
93 benefit VLA models. We also outline possible directions for future research. We believe that this
94 discussion can help to provide a more comprehensive understanding of FreqPolicy’s potential and its
95 future development.

96 **Multitask Benchmark.** To preliminarily assess FreqPolicy’s performance in a multitask setting,
97 we selected RoboCasa [4] as the benchmark platform. RoboCasa comprises a suite of tasks defined
98 within a simulated kitchen environment, representing complex interaction scenarios that robots might
99 encounter in the real world. In this exploratory experiment, we focused on 24 "atomic" tasks, which
100 cover fundamental sensorimotor skills such as pick-and-place, opening and closing doors, pressing
101 buttons, and turning faucets. Placing our method in such a simulated environment, which possesses a
102 certain level of difficulty and a rich variety of tasks, helps us to preliminarily understand its potential
103 and areas for exploration in multitask learning.

104 **Baseline.** For effective comparison, we first selected the Diffusion Policy and the autoregressive
105 method BC-Transformer [3] implemented in RoboCasa as direct baselines for our FreqPolicy. Con-
106 sidering the characteristics of VLA models, we further introduced GR00T-N1 [2] as a reference.
107 GR00T-N1 is a VLA model with a tightly coupled dual-system, where its vision-language module
108 is responsible for understanding the environment and instructions, and the subsequent Diffusion
109 Transformer module employs Flow Matching technology to generate smooth action sequences in
110 real-time. Selecting these baselines helps us to more clearly position the relative performance of
111 FreqPolicy within existing VLA frameworks.

112 **Implementation Details.** In this initial exploration, our FreqPolicy model is directly integrated on
113 top of the Diffusion Policy framework within RoboCasa. To ensure a fair comparison, the main
114 parameters and observation inputs used by FreqPolicy are kept consistent with those of Diffusion
115 Policy. This simplified integration aims to quickly validate the basic adaptability of FreqPolicy’s core
116 mechanisms in a multitask scenario, rather than to perform deep customization and optimization.

117 **Results and Discussion.** The experimental results, as shown in Table 5, offer an initial insight into
118 FreqPolicy’s performance on the RoboCasa multitask benchmark. The data indicate that, compared
119 to Diffusion Policy and BC-Transformer, our method demonstrates a certain advantage in overall
120 multitask success rates. This preliminarily suggests that FreqPolicy’s frequency-domain processing
121 mechanism, and its hierarchical modeling approach to action sequences, are not only effective for
122 single-task learning but may also bring positive impacts to scenarios requiring the simultaneous
123 handling of multiple tasks. However, when compared to the GR00T-N1 model, which is specifically
124 designed for VLA, FreqPolicy’s current performance still shows a gap in multi-task success rates.
125 We attribute this primarily to the fact that FreqPolicy, in its current design and similar to Diffusion
126 Policy, focuses more on action generation and optimization. It has not been specifically enhanced
127 for deep understanding of complex language instructions and multi-modal scene perception to
128 the same extent as GR00T-N1. One of the core strengths of VLA models lies in their powerful

Table 5: **Multitask results on RoboCasa.** Experimental results of BC-Transformer, Diffusion Policy and GR00T-N1 are from the GR00T-N1 paper.

	BC-Transformer	Diffusion Policy	GR00T-N1	FreqPolicy(Ours)
Success Rate	26.3%	25.6%	32.1%	27.4%

semantic understanding and scene perception capabilities, enabling them to generalize better to unseen instruction and environment combinations.

Nevertheless, these preliminary results also provide us with important insights: could FreqPolicy’s unique frequency-domain analysis and modeling approach serve as a beneficial supplement when integrated into more powerful VLA frameworks? For instance, FreqPolicy’s ability to capture the smoothness and structural information of action signals might assist VLA models in generating more stable and physically plausible action sequences. Exploring how to effectively combine the frequency-domain strengths of FreqPolicy with the semantic understanding capabilities of VLA models, with the aim of further enhancing overall multi-task learning performance, will be a highly valuable research direction for us in the future. This might involve designing new fusion mechanisms or tailoring FreqPolicy’s frequency decomposition strategies and autoregressive processes to the specific characteristics of VLA tasks. In summary, while FreqPolicy was not natively designed for VLA tasks, its core ideas demonstrate a potential worthy of further investigation in the broader field of multi-task robot learning.

F.2 Discussion on Frequency Domain

In the preceding discussions, we have preliminarily shown that the action signals in robotic manipulation tasks exhibit significant compressibility in the frequency domain, with most critical information concentrated in the lower frequency bands. This section aims to provide a more in-depth discussion and analysis of the frequency-domain characteristics of action signals, based on a broader set of tasks and more detailed visualizations (as shown in Figures 2 to 25). These supplementary figures provide action visualizations (a), frequency band energy heatmaps for each action dimension (b), and success rate curves for actions reconstructed with varying frequency ratios (c) for each task, thereby offering more robust support for our core arguments.

High-Dimensional Tasks. In high-dimensional action spaces (22 dimensions) within Dexart tasks (Figures 2-5, Dexart Bucket, Faucet, Laptop and Toilet), we observe consistent trends. The success rate curves (c) for these tasks generally show that even using only 30%-70% of the low-frequency components is often sufficient to reconstruct action sequences capable of task completion, strongly supporting the core hypothesis that high-frequency components contribute relatively little to the macroscopic success of these complex tasks. Concurrently, energy heatmaps (b) clearly demonstrate that different action dimensions exhibit varying dependencies on frequency components; some dimensions (like large-range arm movements) have energy highly concentrated in very low-frequency bands, while others (like fine finger postures) might retain significant energy in relatively higher bands. For instance, in Dexart Faucet (Figure 3b), energy distribution in the 10-30% or even higher frequency bands for some dimensions might correspond to fine adjustments for turning a faucet. Although overall trends are similar, the minimum frequency ratio for high success rates varies slightly across Dexart tasks, with Dexart Laptop (Figure 4c), for example, reaching a success plateau around a 0.4-0.6 frequency ratio, suggesting subtle differences in action signal fidelity requirements for various complex manipulations.

Low-Dimensional Tasks. Compared to high-dimensional Dexart tasks, low-dimensional Meta-World tasks (Figures 6-25, typically 4-dimensional action spaces) exhibit a more pronounced low-frequency dominance. In most Meta-World tasks, success rate curves (c) indicate that only 10%-40% of the low-frequency ratio is sufficient for near-perfect task success, with tasks like Meta-World Coffee-Pull (Figure 8c) and Meta-World Disassemble (Figure 10c) requiring only about 20% low-frequency signal. This suggests higher compressibility in action signals for these simpler robotic tasks, corroborated by their energy heatmaps (b) where most action dimensions show energy concentrated in the lowest 0-10% band, consistent with their typically smoother, direct motion trajectories. Minor exceptions, such as Meta-World Shelf-Place (Figure 23c) or Meta-World Push-Wall (Figure 21c), might need slightly more frequency components due to potentially higher precision demands at the end-effector.

177 **Discussion.** Synthesizing these analyses, we discover the universality and variability of frequency
178 compression across task types: action signals are generally compressible in both high-dimensional
179 complex and low-dimensional structured tasks, though the required frequency bandwidth varies with
180 task dimensionality, complexity, and operational specifics. The heterogeneity of action dimensions
181 highlighted by heatmap analysis suggests that future work on dimension-adaptive frequency pro-
182 cessing, rather than global uniform cutoffs, could be a promising optimization, such as dynamically
183 allocating frequency components based on each dimension’s energy to balance performance and
184 efficiency. A deeper understanding of task action frequency characteristics can also guide policy
185 learning algorithm design, for example, by using low-frequency biases or stronger regularization for
186 low-frequency dominated tasks to accelerate learning and enhance generalization.

187 In conclusion, the additional frequency domain analysis in this section provides more comprehensive
188 empirical support for FreqPolicy’s core mechanisms, points to valuable directions for optimizing
189 frequency-based robot learning, deepens our understanding of robotic action nature, and underscores
190 the potential of the frequency-domain perspective in building efficient, robust robotic agents.

191 **G Limitations**

192 Since all of our experiments used condition inputs consistent with DP3 or DP, we have not yet explored
193 the potential impact of altering condition input methods on model performance. Additionally, it
194 should be noted that our method still has room for improvement in 2D tasks, and performance tends
195 to decrease when frequency domain partitioning becomes too fine-grained.

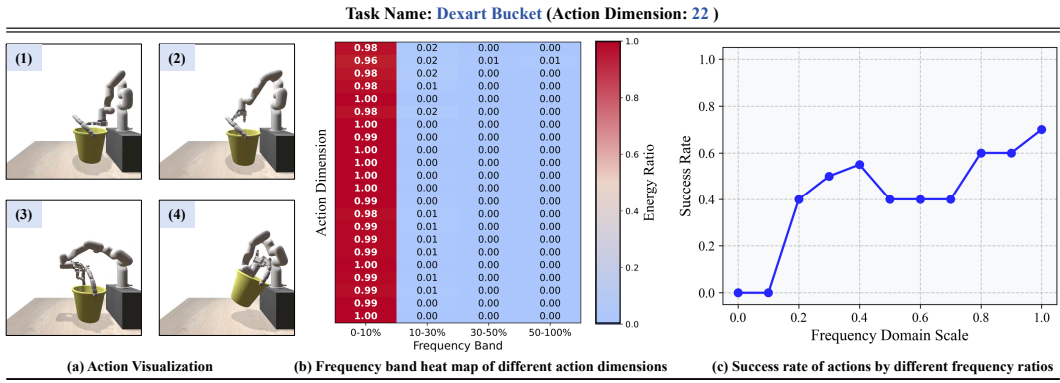


Figure 2: Frequency Domain Analysis of Dexart Bucket.

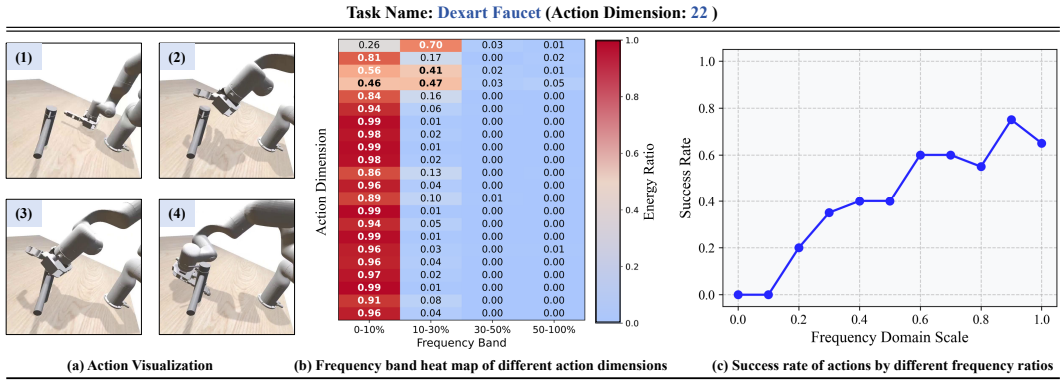


Figure 3: Frequency Domain Analysis of Dexart Faucet.

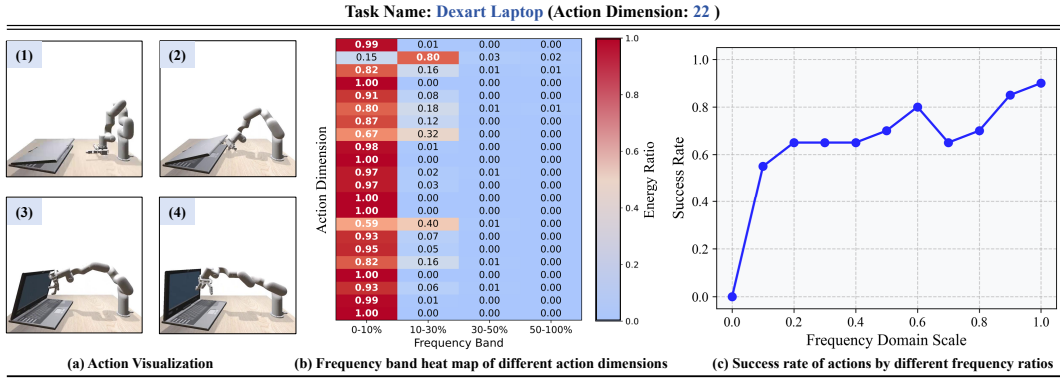


Figure 4: Frequency Domain Analysis of Dexart Laptop.

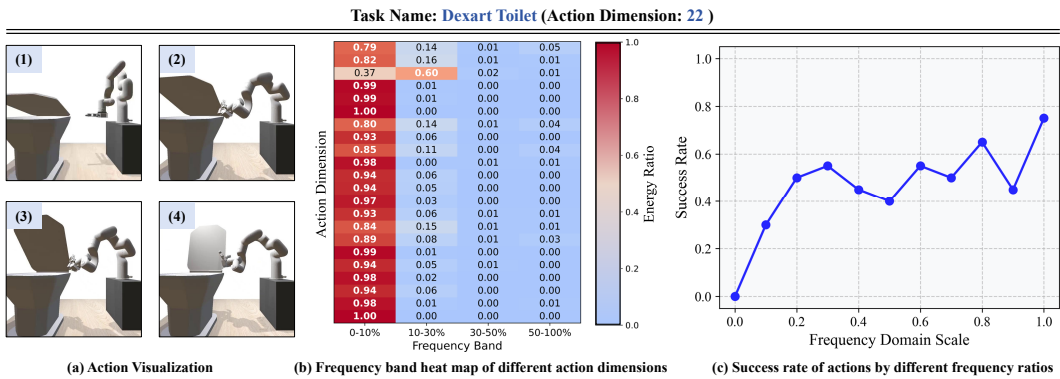


Figure 5: Frequency Domain Analysis of Dexart Toilet.

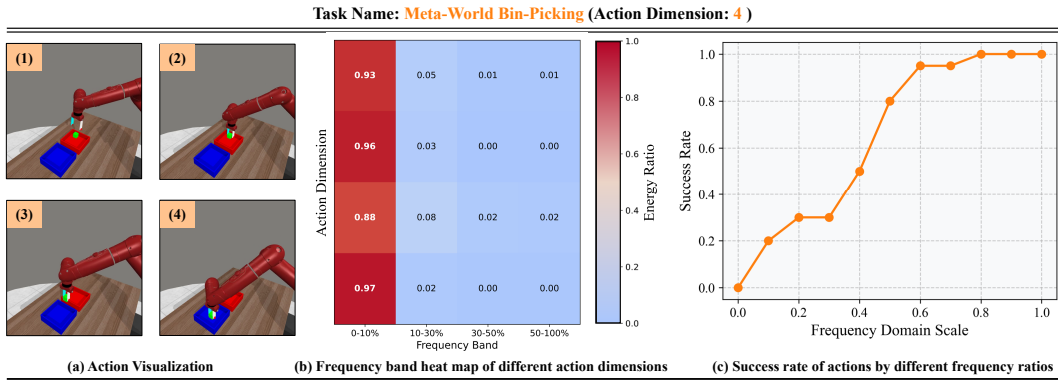


Figure 6: Frequency Domain Analysis of Meta-World Bin-Picking.

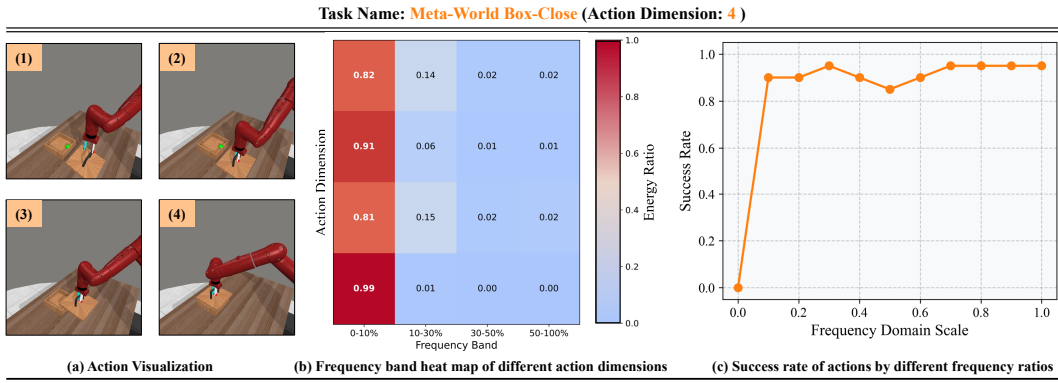


Figure 7: Frequency Domain Analysis of Meta-World Box-Close.

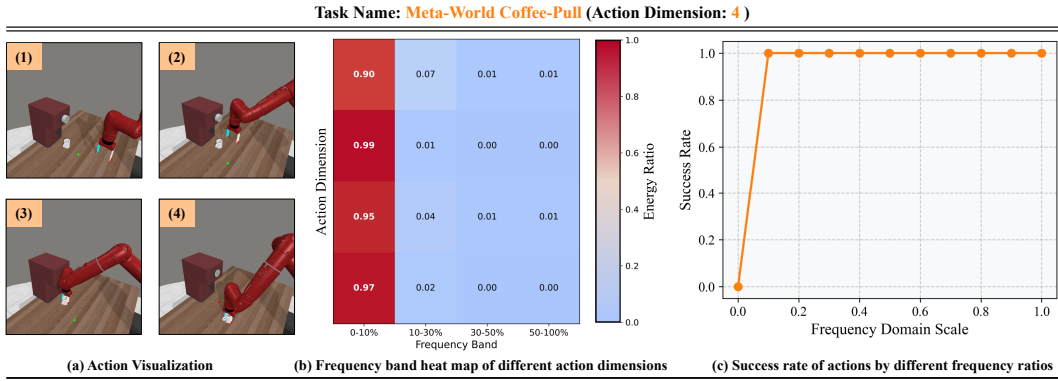


Figure 8: Frequency Domain Analysis of Meta-World Coffee-Pull.

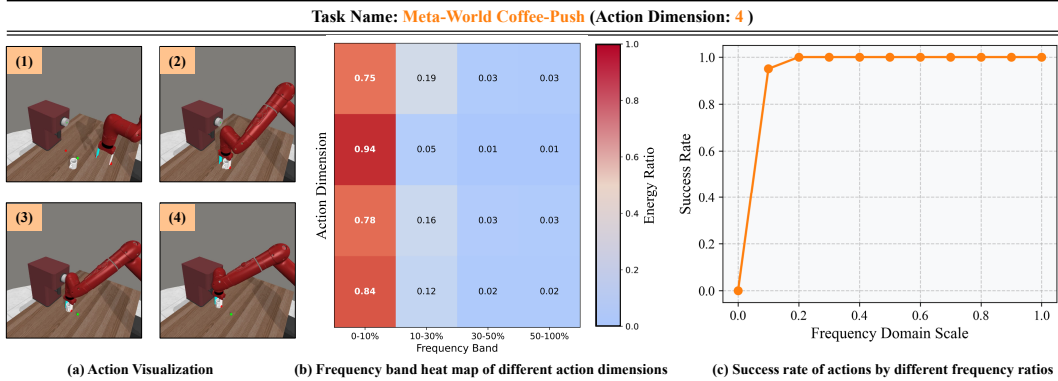


Figure 9: Frequency Domain Analysis of Meta-World Coffee-Push.

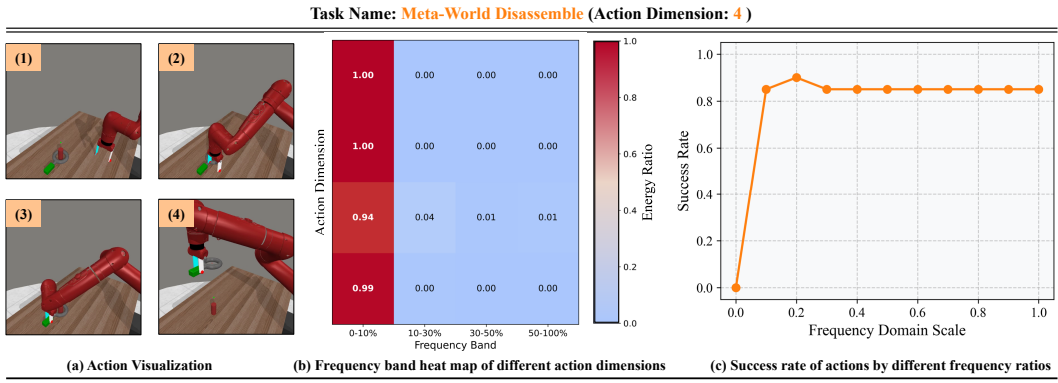


Figure 10: Frequency Domain Analysis of Meta-World Disassemble.

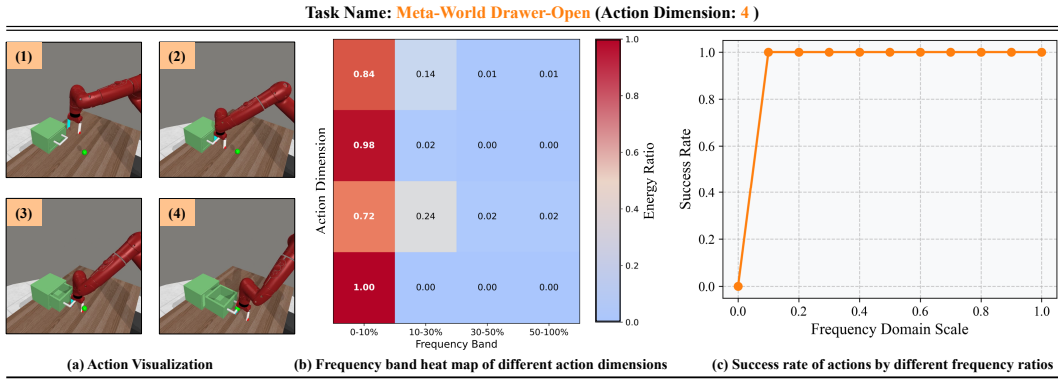


Figure 11: Frequency Domain Analysis of Meta-World Drawer-Open.

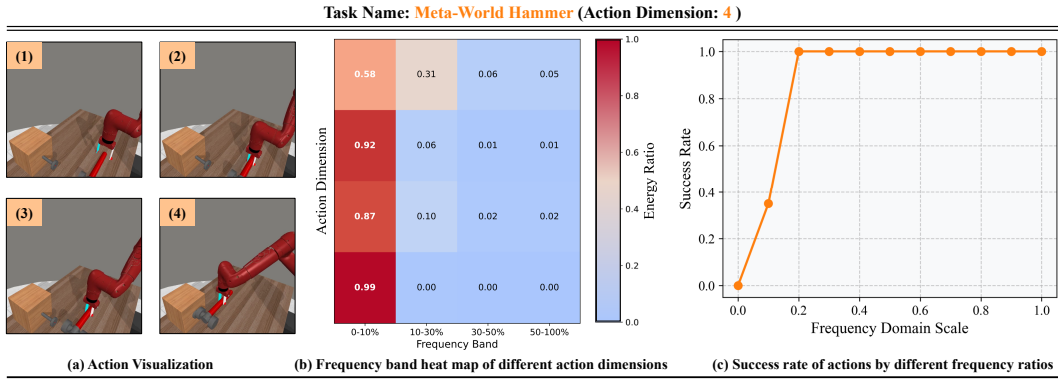


Figure 12: Frequency Domain Analysis of Meta-World Hammer.

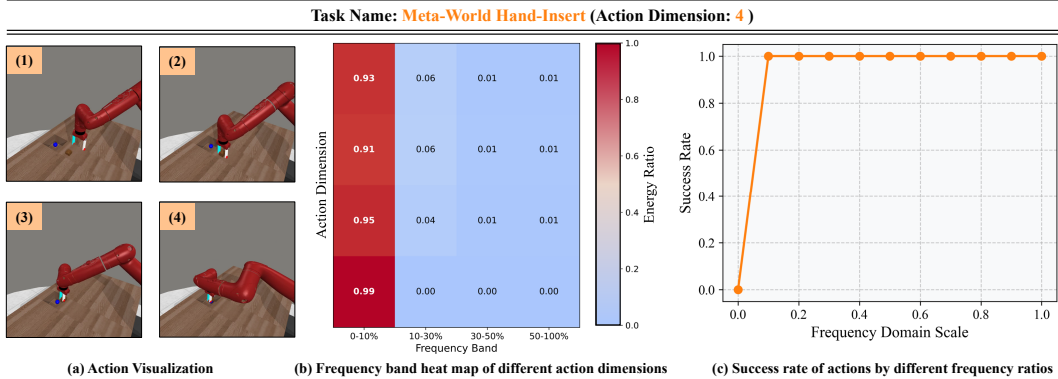


Figure 13: Frequency Domain Analysis of Meta-World Hand-Insert.

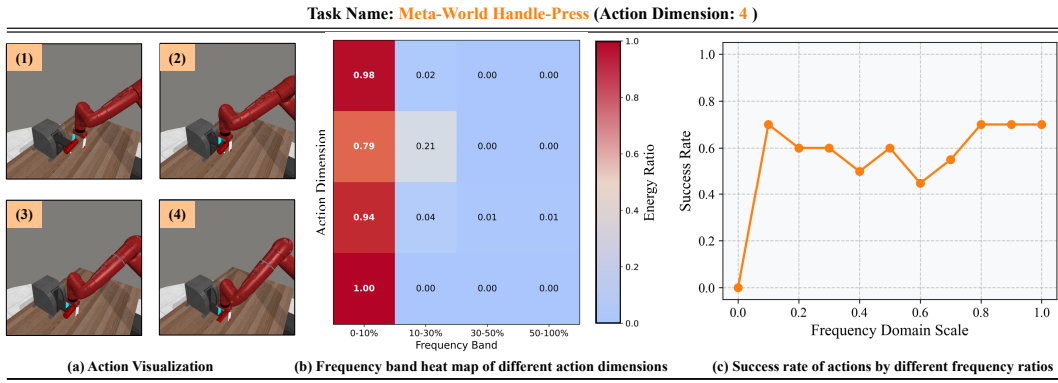


Figure 14: Frequency Domain Analysis of Meta-World Handle-Press.

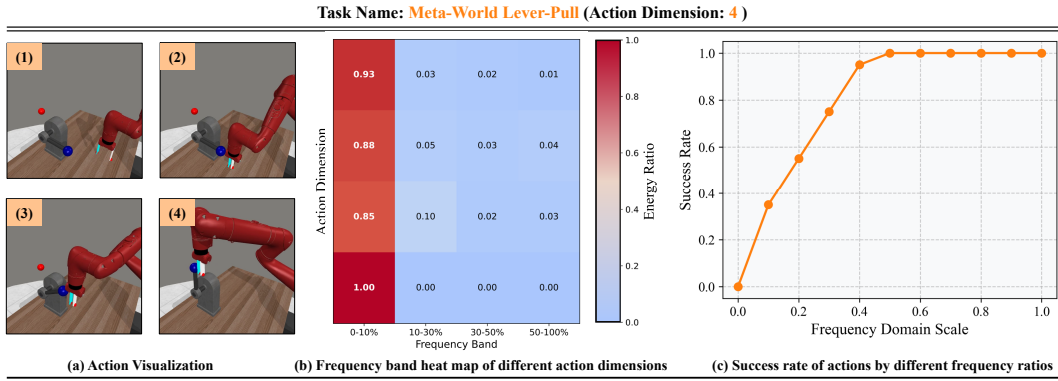


Figure 15: Frequency Domain Analysis of Meta-World Lever-Pull.

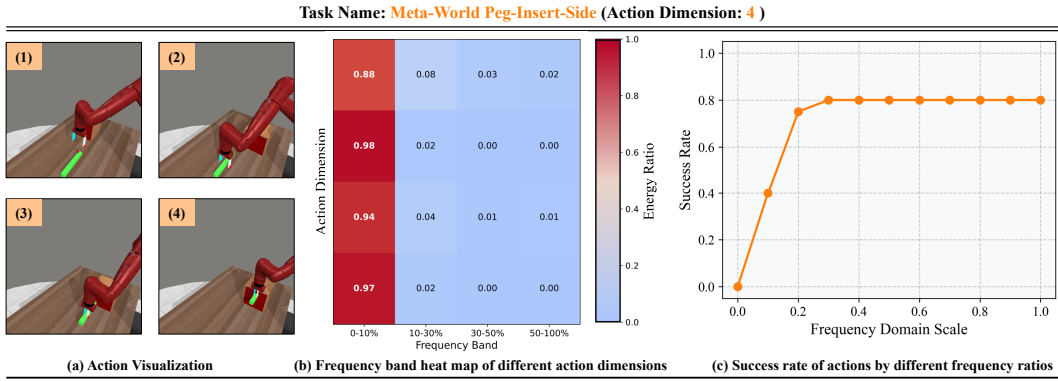


Figure 16: Frequency Domain Analysis of Meta-World Peg-Insert-Side.

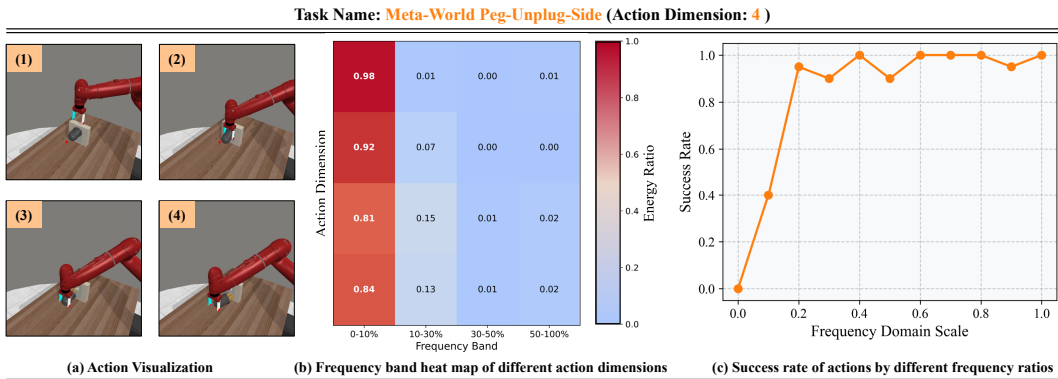


Figure 17: Frequency Domain Analysis of Meta-World Peg-Unplug-Side.

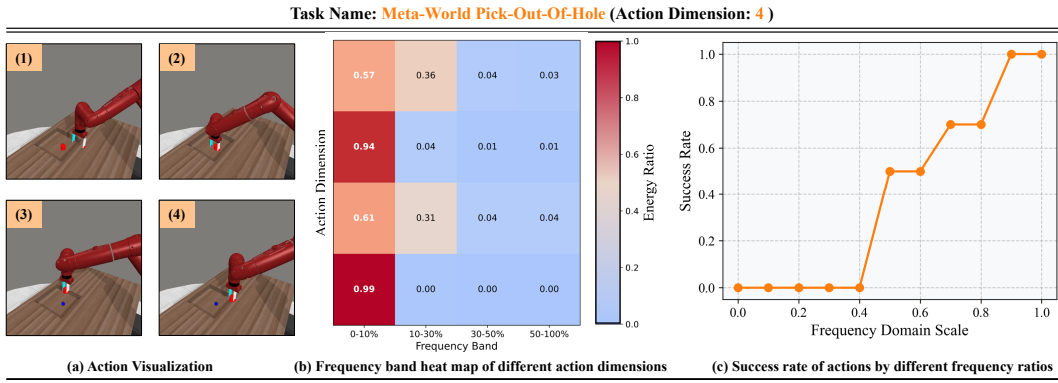


Figure 18: Frequency Domain Analysis of Meta-World Pick-Out-Of-Hole.

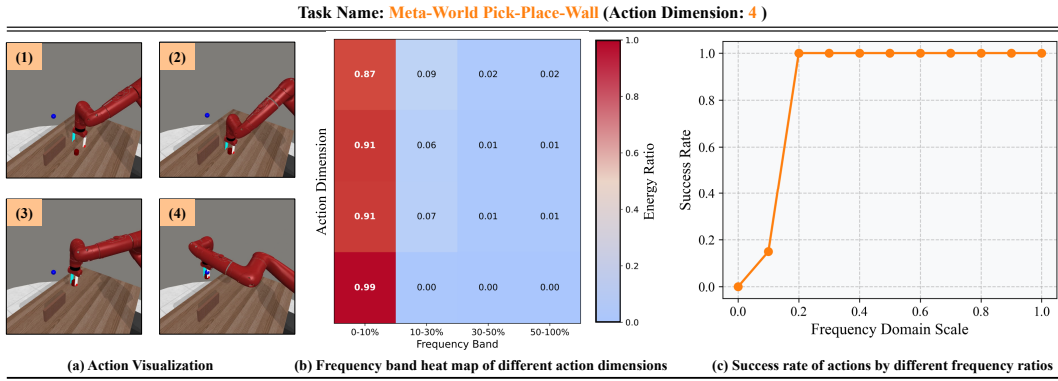


Figure 19: Frequency Domain Analysis of Meta-World Pick-Place-Wall.

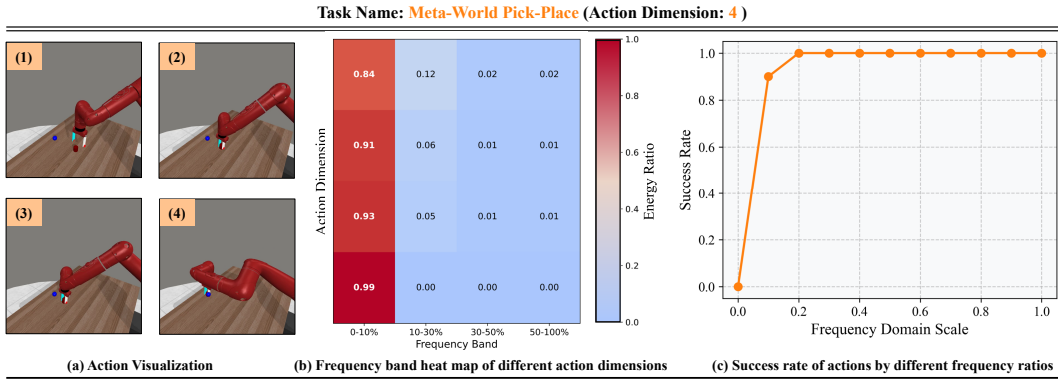


Figure 20: Frequency Domain Analysis of Meta-World Pick-Place.

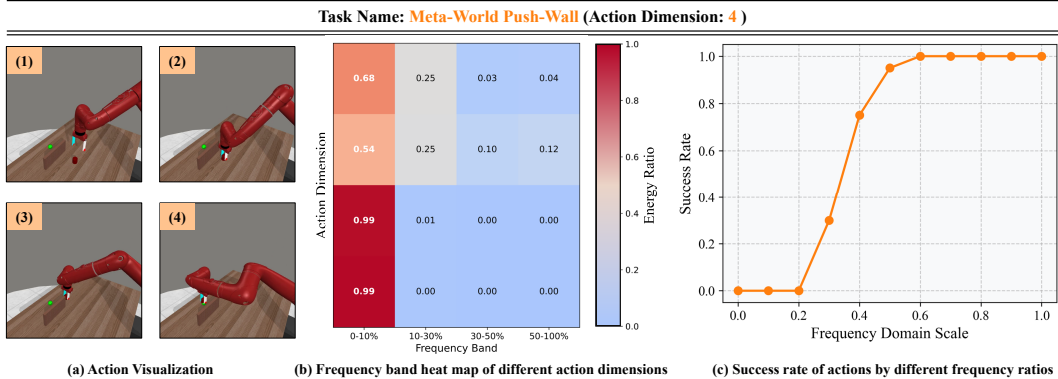


Figure 21: Frequency Domain Analysis of Meta-World Push-Wall.

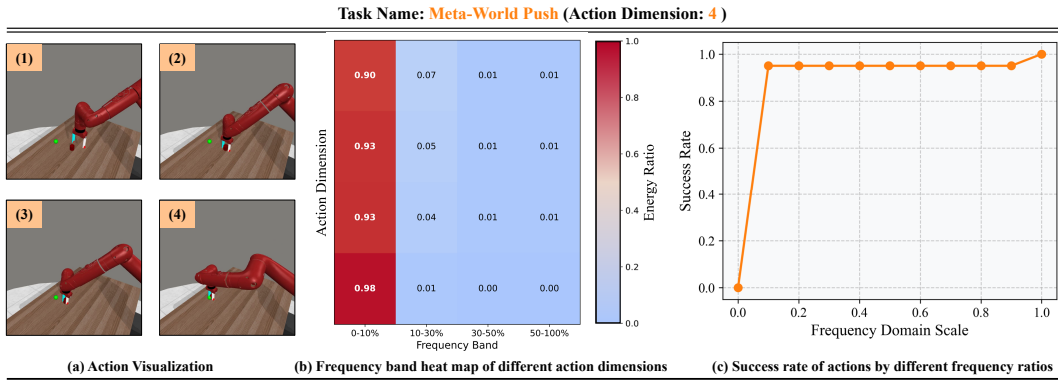


Figure 22: Frequency Domain Analysis of Meta-World Push.

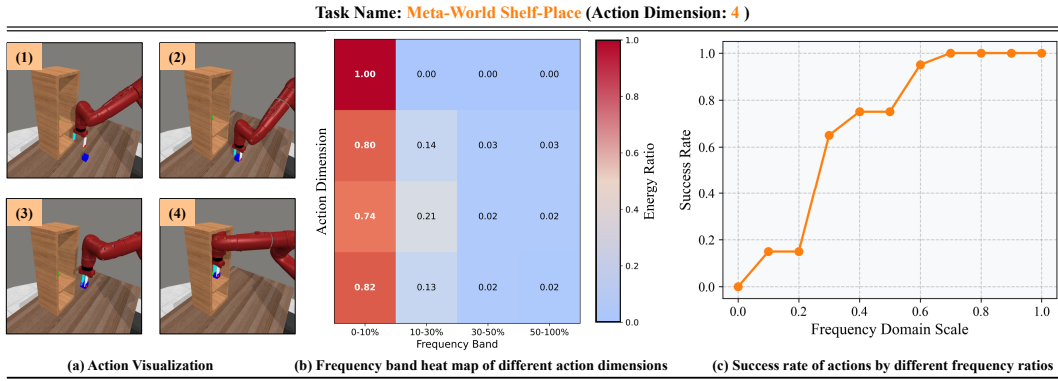


Figure 23: Frequency Domain Analysis of Meta-World Shelf-Place.

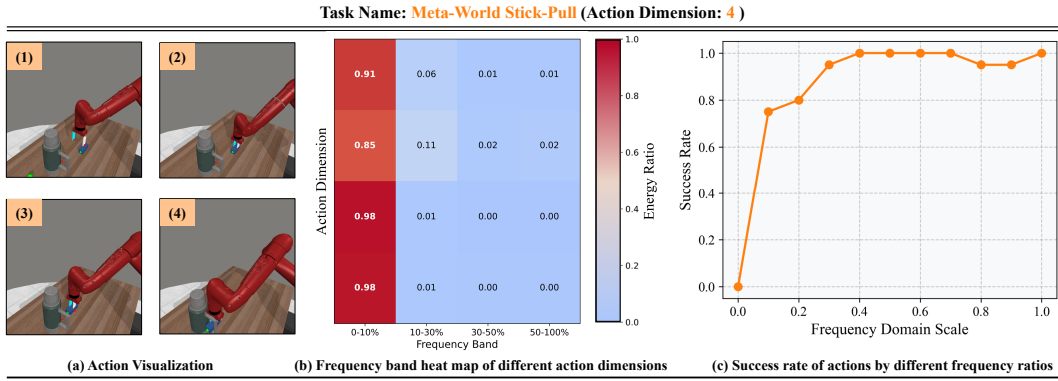


Figure 24: Frequency Domain Analysis of Meta-World Stick-Pull.

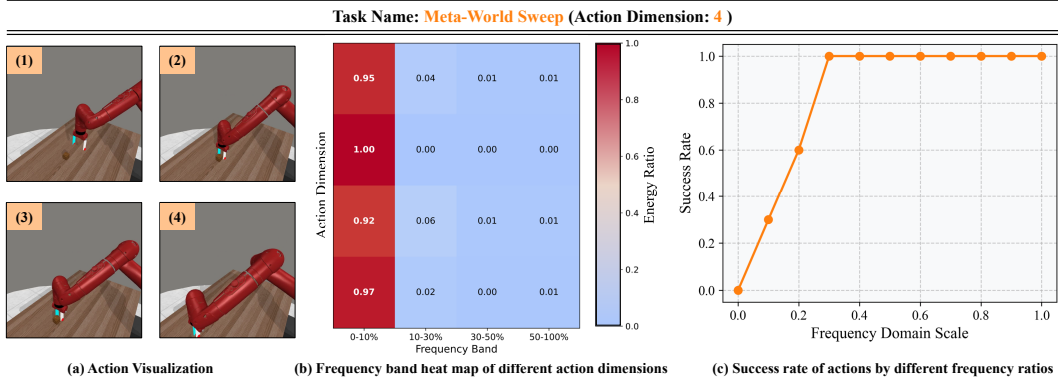


Figure 25: Frequency Domain Analysis of Meta-World Sweep.

References

- [1] Chen Bao, Helin Xu, Yuzhe Qin, and Xiaolong Wang. Dexart: Benchmarking generalizable dexterous manipulation with articulated objects. In *CVPR*, 2023a.
- [2] NVIDIA: Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi "Jim" Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, Joel Jang, Zhenyu Jiang, Jan Kautz, Kaushil Kundalia, Lawrence Lao, Zhiqi Li, Zongyu Lin, Kevin Lin, Guilin Liu, Edith Llontop, Loic Magne, Ajay Mandlekar, Avnish Narayan, Soroush Nasiriany, Scott Reed, You Liang Tan, Guanzhi Wang, Zu Wang, Jing Wang, Qi Wang, Jiannan Xiang, Yuqi Xie, Yinzhen Xu, Zhenjia Xu, Seonghyeon Ye, Zhiding Yu, Ao Zhang, Hao Zhang, Yizhou Zhao, Ruijie Zheng, and Yuke Zhu. Gr00t n1: An open foundation model for generalist humanoid robots, 2025.
- [3] Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation, 2021.
- [4] Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots, 2024.
- [5] Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. *arXiv*, 2017.
- [6] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *CoRL*, 2020.