H^{3}DP: Triply-Hierarchical Diffusion Policy for Visuomotor Learning

Yiyang Lu[1*], Yufeng Tian[4*], Zhecheng Yuan[1,2,3*], Xianbang Wang[1],
Pu Hua[1,2,3], Zhengrong Xue[1,2,3], Huazhe Xu[1,2,3]
[1] Tsinghua University IIIS, [2] Shanghai Qi Zhi Institute, [3] Shanghai AI Lab, [4] HIT
luyy24@mails.tsinghua.edu.cn, huazhe_xu@mail.tsinghua.edu.cn

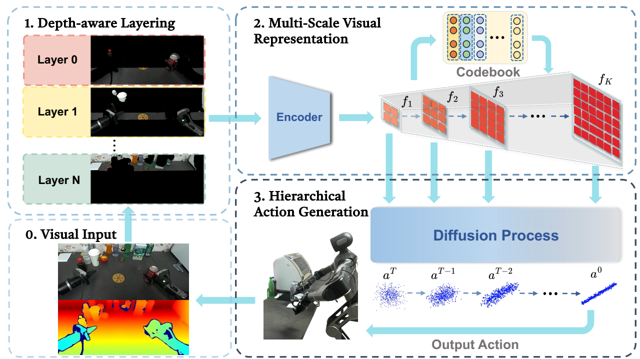

Figure 1. **Overview of H^{3}DP**.

Abstract

*We introduce **Triply-Hierarchical Diffusion Policy (H^{3}DP)**, a novel visuomotor learning framework that explicitly incorporates hierarchical structures to strengthen the integration between visual features and action generation. H^{3}DP contains **3** levels of hierarchy: (1) depth-aware input layering; (2) multi-scale visual representations; and (3) a hierarchically conditioned diffusion process. Extensive experiments demonstrate that H^{3}DP yields a $+\mathbf{27.5}\%$ average relative improvement over baselines across **44** simulation tasks and achieves superior performance in **4** challenging bimanual real-world manipulation tasks.*[1]

1. Introduction

Visuomotor policy learning is a prevailing paradigm in robotic manipulation [2, 3, 22, 23, 25]. Existing approaches have increasingly adopted powerful generative methods [5, 9, 12, 17, 20] to model the action generation process, but often overlook establishing a tight correspondence between perception and action. In this paper, we present **H^{3}DP**, a novel visuomotor policy learning framework grounded in three levels of hierarchy.

At the input level, H^{3}DP moves beyond prior 2D approaches [23, 26] by introducing a **depth-aware layering** strategy that partitions RGB-D input into distinct layers based on depth cues. For visual representation, to address limitations of flattening image features [7, 10, 16], H^{3}DP employs **multi-scale visual representation**, where different scales capture features at varying granularity levels. In action generation, H^{3}DP incorporates **hierarchical action generation**, leveraging the diffusion process's tendency to progressively reconstruct features from low to high-frequency components [4, 15, 19].

We validate H^{3}DP through extensive experiments on **44** simulation tasks across **5** diverse benchmarks, where it surpasses state-of-the-art methods by a relative average margin of $+\mathbf{27.5}\%$. Furthermore, real-world evaluations on bimanual robotic systems in cluttered, high-disturbance, long-horizon tasks show H^{3}DP achieves a $+\mathbf{32.3}\%$ performance improvement over Diffusion Policy.

2. Method

We employ three hierarchical structures to enhance the policy's understanding of visual input and predict more accurate action distributions. A detailed discussion of each part will be provided in the following sections.

2.1. Depth-aware Layering

To fully exploit the geometric structure inherent in depth maps, we introduce a depth-aware layering mechanism. Pixels with depth d are assigned to layer m using linear-increasing discretization [24] $m = \lfloor -0.5 + 0.5\sqrt{1 + 4(N+1)(N+2)\frac{d-d_{\min}}{d_{\max}-d_{\min}+\epsilon}} \rfloor$, which promotes the robot to focus more on its workspace. By explicitly encoding objects distributed across different depth planes, this structured representation retains all visual detail while strategically utilizing depth to impose a meaningful foreground-background separation, thereby enabling the policy to selectively attend to different regions of the image.

2.2. Multi-Scale Visual Representation

Existing methods typically extract features at a single spatial scale or compress them into a fixed-resolution representation, limiting the expressiveness of learned features [7, 10, 16]. To address this problem, we hierarchically

---

*Equal Contribution
[1]Project Page: https://lyy-iiis.github.io/h3dp.

Table 1. **Simulation task results.**

Method \ Tasks	MetaWorld (Medium 11)	MetaWorld (Hard 5)	MetaWorld (Hard++ 5)	ManiSkill (Deformable 4)	ManiSkill (Rigid 4)	Adroit (3)	DexArt (4)	RoboTwin (8)	**Average (44)**
H³DP	**98.3**	**87.8**	**95.8**	**59.3**	**65.3**	**87.3**	53.3	**57.4**	**75.6±18.6**
DP	78.2	52.6	58.0	22.3	27.5	79.0	44.3	22.8	48.1±23.1
DP (w/ depth)	77.7	57.2	71.2	44.5	40.8	76.0	42.0	12.6	52.8±22.2
DP3	89.1	52.6	88.4	26.5	33.5	84.0	**54.8**	45.9	59.3±24.9

partition the feature map into multiple scales, enabling the capture of both coarse global and detailed local information.

Interpolation and Quantization. After applying depth-aware layering to the input image I, each layer I_m is independently encoded into multi-scale feature maps $\{f_{m,k}|f_{m,k} \in \mathbb{R}^{h_k \times w_k \times C}\}_{k=1}^{K}$, where $\{(h_k, w_k)\}_{k=1}^{K}$ denotes the spatial resolutions across scales. Adopting the quantization design in VQ-VAE [14, 18], these feature maps $\{f_{m,k}\}_{k=1}^{K}$ are quantized into discrete vectors drawn from a learnable codebook $\mathcal{Z}_m \in \mathbb{R}^{V \times C}$. Specifically, each feature vector $f_{m,k}^{(i,j)}$ is mapped to its nearest neighbor in Euclidean distance: $f_{m,k}^{(i,j)} \leftarrow \arg\min_{z \in \mathcal{Z}_m} \|z - f_{m,k}^{(i,j)}\|_2$. By applying differentiable interpolation and lightweight convolution to the quantized features $f_{m,k}$, we then obtain the multi-scale visual representations $\{\hat{f}_{m,k}\}_{k=1}^{K}$ for each layer I_m.

2.3. Hierarchical Action Generation

To match the inherent inductive biases of denoising process [4, 15, 19], we leverage multi-scale visual representations to model action generation in a coarse-to-fine manner.

Inference. Our action generation module is a denoising diffusion model conditioned on multi-scale features $F = \{\hat{f}_k = \{\hat{f}_{m,k}\}_{m=0}^{N-1}\}_{k=1}^{K}$ and robot poses q. The denoising process unfolds over T steps partitioned into K stages $\cup_{k=1}^{K}(\tau_{k-1}, \tau_k]$. When $t \in (\tau_{k-1}, \tau_k]$, the denoising network $\epsilon_\theta^{(t)}$ conditioning on the corresponding feature map $\hat{f}_k$ and robot poses q, predicts the noise component $\epsilon^t = \epsilon_\theta^{(t)}(a^t|\hat{f}_k, q)$, then generates $a^{t-1} = \alpha_t a^t + \beta_t \epsilon^t + \sigma_t \tilde{\epsilon}^t$, gradually transforming the Gaussian noise a^T into the noise-free action a^0, where $\alpha_t, \beta_t, \sigma_t$ are fixed parameters, and $\tilde{\epsilon}^t \sim \mathcal{N}(0, \mathbf{I})$ is a Gaussian noise. Features at varying resolutions retain information across distinct frequency domains. By using lower-resolution features for earlier stages and gradually refining the predictions with higher-resolution features, the model benefits from both the stability of coarse representations and the precision of fine details.

Training. To train the denoising network $\epsilon_\theta^{(t)}$, we randomly sample an observation-action pair $((I, q), a^0) \in \mathcal{D}$ and noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. The network is optimized to predict ϵ given a noisy action conditioned on the final feature map $\hat{f}_K$ and robot pose q, via the objective: $\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{a^0,\epsilon,t}\left[\|\epsilon_\theta^{(t)}(\sqrt{\alpha_t}a^0 + \sqrt{1-\alpha_t}\epsilon|\hat{f}_K, q) - \epsilon\|^2\right]$.

3. Experiments

3.1. Simulation Experiments

3.1.1. Experiment setup

To sufficiently verify the effectiveness of H³DP, we evaluate H³DP on **5** simulation benchmarks, encompassing a total of **44** tasks [1, 6, 11, 13, 21]. To comprehensively assess the performance of H³DP, we compare it against three baselines: *Diffusion Policy* [3], *Diffusion Policy (w/ depth)* and *DP3* [23].

Table 2. **Instance generalization results.**

Method \ Tasks	Place Bottle			Sweep Trash			**Average**
	coke bottle	sprite	can	$8\,\text{cm}^3$	$64\,\text{cm}^3$	$216\,\text{cm}^3$	
H³DP	**67**	**49**	**53**	**75**	**86**	**67**	**66.2**
Diffusion Policy	45	36	40	52	72	60	50.8

3.1.2. Simulation performance

As shown in Table 1, the simulation experiment results exhibit that H³DP outperforms or achieves comparable performance among the whole simulation benchmarks. Our method outperforms DP3 by a relative average margin of $+\mathbf{27.5}\%$. Notably, DP3 requires manual segmentation of the point cloud to remove background and task-irrelevant elements. In contrast, benefiting from our design, H³DP obtains superior performance using only raw RGB-D input, without the need for segmentation and human effort.

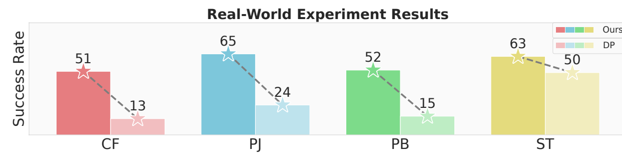

Figure 2. **Success rate in real-world.**

3.2. Real-world Experiments

In terms of real-world experiments, we choose Galaxea R1 robot as our platform. We use the ZED camera to acquire the depth image with 60Hz running frequency. The demonstrations are collected by Meta Quest3.

We design four diverse challenging real-world tasks to evaluate the effectiveness of our method: Clean Fridge (CF), Pour Juice (PJ), Place Bottle (PB), Sweep Trash (ST). Regarding the two long-horizon tasks, both the baseline and our method incorporate the pre-trained ResNet18 [8] encoders for RGB modality to enhance the policy's perceptual capabilities in real-world environments.

3.2.1. Experiment Results

Spatial generalization: As shown in Figure 2, H³DP significantly outperforms the baseline across all four real-world tasks, achieving an average improvement of $+\mathbf{32.3}\%$. H³DP demonstrates superior perceptual and decision-making capabilities compared to alternative algorithms. Meanwhile, it should be noted that in terms of the point cloud based method DP3, it requires precise segmentation and high-fidelity depth sensing, resulting in it being less effective in handling our four cluttered real-world scenes that we designed.

Instance generalization: Regarding instance generalization, we evaluate the model on two real-world tasks by varying the size and shape of bottles or trash items. As shown in Table 2, after replacing the objects with variants of differing sizes and shapes, H³DP maintains strong generalization capabilities attributable to its ability to hierarchically model features at multiple levels of granularity, and consistently outperforms baseline approaches across all settings.

References

[1] Chen Bao, Helin Xu, Yuzhe Qin, and Xiaolong Wang. Dexart: Benchmarking generalizable dexterous manipulation with articulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21190–21200, 2023. 2

[2] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. 1

[3] 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

[4] Sander Dieleman. Diffusion is spectral autoregression, 2024. 1, 2

[5] Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. *arXiv preprint arXiv:2410.12557*, 2024. 1

[6] Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, et al. Maniskill2: A unified benchmark for generalizable manipulation skills. *arXiv preprint arXiv:2302.04659*, 2023. 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015. 1

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[9] Seungjae Lee, Yibin Wang, Haritheja Etukuru, H Jin Kim, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. Behavior generation with latent actions. *arXiv preprint arXiv:2403.03181*, 2024. 1

[10] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 1

[11] 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

[12] Aaditya Prasad, Kevin Lin, Jimmy Wu, Linqi Zhou, and Jeannette Bohg. Consistency policy: Accelerated visuomotor policies via consistency distillation. *arXiv preprint arXiv:2405.07503*, 2024. 1

[13] 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 preprint arXiv:1709.10087*, 2017. 2

[14] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 2

[15] Severi Rissanen, Markus Heinonen, and Arno Solin. Generative modelling with inverse heat dissipation. *arXiv preprint arXiv:2206.13397*, 2022. 1, 2

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 1

[17] Nur Muhammad Shafiullah, Zichen Cui, Ariuntuya Arty Altanzaya, and Lerrel Pinto. Behavior transformers: Cloning k modes with one stone. *Advances in neural information processing systems*, 35:22955–22968, 2022. 1

[18] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2

[19] Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. Ddt: Decoupled diffusion transformer. *arXiv preprint arXiv:2504.05741*, 2025. 1, 2

[20] Zhendong Wang, Zhaoshuo Li, Ajay Mandlekar, Zhenjia Xu, Jiaojiao Fan, Yashraj Narang, Linxi Fan, Yuke Zhu, Yogesh Balaji, Mingyuan Zhou, et al. One-step diffusion policy: Fast visuomotor policies via diffusion distillation. *arXiv preprint arXiv:2410.21257*, 2024. 1

[21] 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 *Conference on robot learning*, pages 1094–1100. PMLR, 2020. 2

[22] Zhecheng Yuan, Tianming Wei, Shuiqi Cheng, Gu Zhang, Yuanpei Chen, and Huazhe Xu. Learning to manipulate anywhere: A visual generalizable framework for reinforcement learning. *arXiv preprint arXiv:2407.15815*, 2024. 1

[23] 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. 1, 2

[24] Renrui Zhang, Han Qiu, Tai Wang, Ziyu Guo, Ziteng Cui, Yu Qiao, Hongsheng Li, and Peng Gao. Monodetr: Depth-guided transformer for monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9155–9166, 2023. 1

[25] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023. 1

[26] Haoyi Zhu, Yating Wang, Di Huang, Weicai Ye, Wanli Ouyang, and Tong He. Point cloud matters: Rethinking the impact of different observation spaces on robot learning. *arXiv preprint arXiv:2402.02500*, 2024. 1