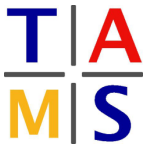
# Recent Diffusion-Based Methods for Imitation Learning

## Intelligent Robotics Seminar SS25

Sven Schreiber

TAMS Group
Department of Informatics
MIN Faculty
University of Hamburg

June 26, 2025





Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

# Table of Contents

# Motivation

- Learning complex behaviors has many challenges
- Diffusion models enable:
  - Robustness to unstructured data
  - Multimodal action space
- Recent methods improve performance with
  - 3D visual representations
  - Hierarchical perception and action prediction

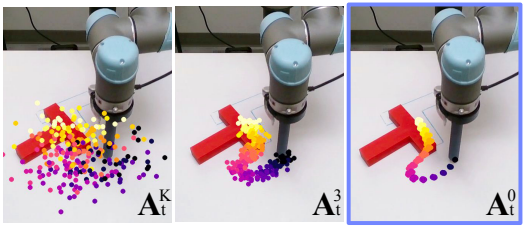



Figure: Chi et al.[1]

[1] Chi et al. 'Diffusion Policy: Visuomotor Policy Learning via Action Diffusion', RSS'23

# 3D Diffusion Policy (DP3)[1]

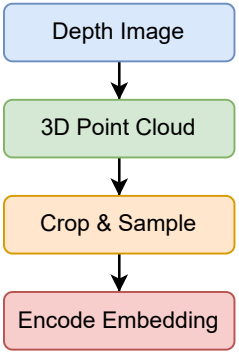
- Builds on *Diffusion Policy*[2]
- Makes use of 3D point clouds
- Condition a diffusion model
- Denoise action

[1] Ze et al. '3D Diffusion Policy: Generalizable Visuomotor Policy Learning via Simple 3D Representations', RSS'24

[2] Chi et al. 'Diffusion Policy: Visuomotor Policy Learning via Action Diffusion', RSS'23

# Perception Module

- Single-view camera setup
- **Input:** Depth images of size $84 \times 84$
- Transform to 3D point cloud
- Crop-out distractions
- Encode to compact representation
- **Output:** 3D visual features



# Depth Image to Point Cloud

Figure: Hoegner et al.[1]

[1] Hoegner et al. 'Co-registration of Time-of-Flight (TOF) camera generated 3d point clouds and thermal infrared images (IR)', DGPF'13

# Point Cloud Processing

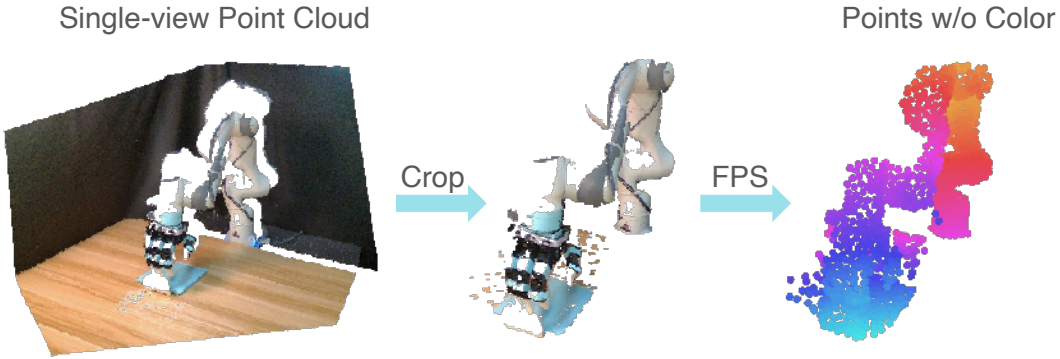



Figure: Ze et al.[1]

[1] Ze et al. '3D Diffusion Policy: Generalizable Visuomotor Policy Learning via Simple 3D Representations', RSS'24

# Point Cloud Sampling


Figure: Furthest Point Sampling[1]

[1] https://minibatchai.com/2021/08/07/FPS.html

# Point Cloud Encoding

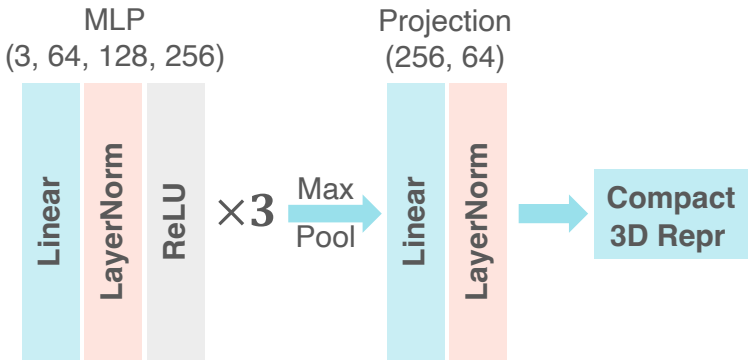



Figure: Ze et al.[1]

[1] Ze et al. '3D Diffusion Policy: Generalizable Visuomotor Policy Learning via Simple 3D Representations', RSS'24

# Decision Module

- Denoising diffusion model
- Denoises random Gaussian noise into action
- **Input:**
  - 3D visual features
  - Robot poses
- **Output:** denoised action

# Denoising Diffusion Model

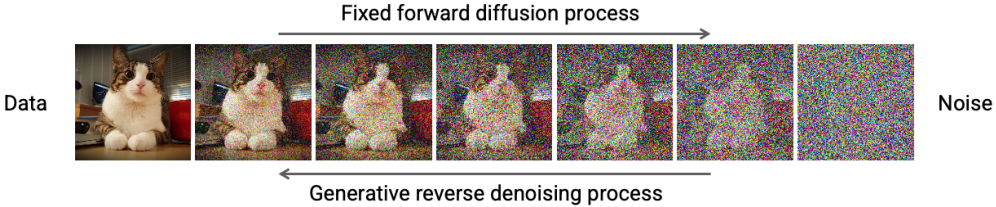



Figure: Song et al.[1]

[1] Song et al. 'Denoising Diffusion Models: A Generative Learning Big Bang', https://cvpr2023-tutorial-diffusion-models.github.io/

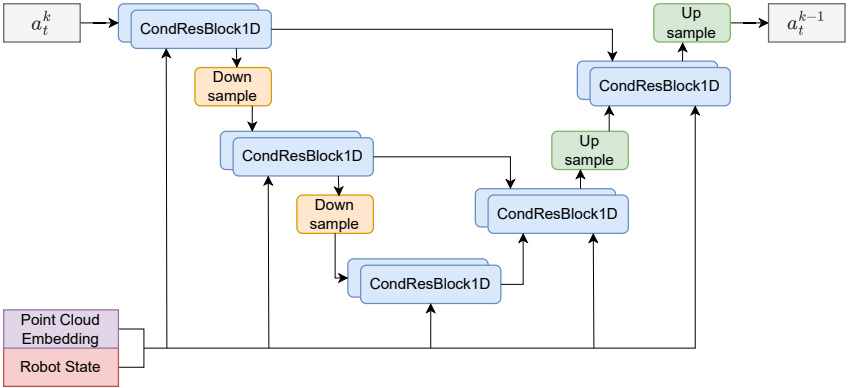
# Denoising Diffusion Model (cont.)

- $v$: 3D visual features
- $q$: robot poses
- $a^k$: action
- $\epsilon_\theta$: denoising network
- $\alpha_k, \gamma_k, \sigma_k$: denoising coefficients

## Denoising Diffusion Model

$$a^{k-1} = \alpha_k(a^k - \gamma_k \epsilon_\theta(a^k, k, v, q)) + \sigma_k \mathcal{N}(0, \mathbf{I})$$

# Denoising Model Architecture



# Results

| Algorithm \ Task | Adroit | | | MetaWorld | | | DexArt | | | | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Hammer | Door | Pen | Assembly | Disassemble | Stick-Push | Laptop | Faucet | Toilet | Bucket | |
| **DP3** | **100±0** | **62±4** | **43±6** | **99±1** | **69±4** | **97±4** | **83±1** | **63±2** | **82±4** | **46±2** | **74.4** |
| Diffusion Policy | 48±17 | 50±5 | 25±4 | 15±1 | 43±7 | 63±3 | 69±4 | 23±8 | 58±2 | **46±1** | 44.0 |
| BCRNN | 0±0 | 0±0 | 9±3 | 3±4 | 32±12 | 45±11 | 3±3 | 1±0 | 5±5 | 0±0 | 9.8 |
| BCRNN+3D | 8±14 | 0±0 | 8±1 | 1±5 | 11±6 | 0±0 | 29±12 | 26±2 | 38±10 | 24±11 | 14.5 |
| IBC | 0±0 | 0±0 | 9±2 | 0±0 | 1±1 | 16±2 | 3±2 | 7±1 | 14±1 | 0±0 | 5.0 |
| IBC+3D | 0±0 | 0±0 | 10±1 | 18±9 | 3±5 | 50±6 | 1±1 | 7±2 | 15±1 | 0±0 | 10.4 |

Figure: Ze et al.[1]

[1] Ze et al. '3D Diffusion Policy: Generalizable Visuomotor Policy Learning via Simple 3D Representations', RSS'24

# Generalizability

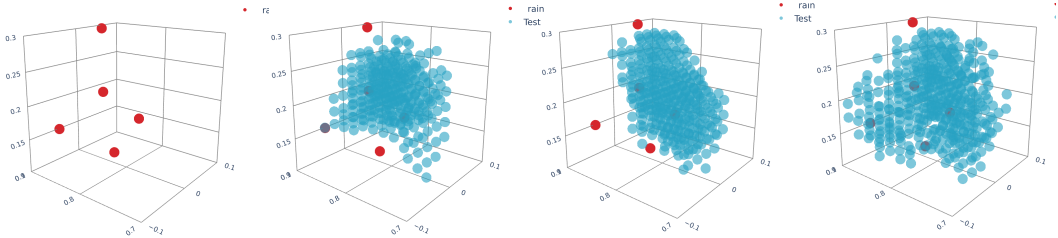



Figure: Ze et al.[1]

[1] Ze et al. '3D Diffusion Policy: Generalizable Visuomotor Policy Learning via Simple 3D Representations', RSS'24

# Discussion

- Performant and sample efficient
  - For some tasks, 5 demonstrations suffice
- No RGB input
  - Good for generalization
  - But, for some tasks RGB is required
- Needs manual scene segmentation

# Triply-Hierarchical Diffusion Policy ($H^3DP$)[1]

- Incoorporates hierarchical structure into:
  - Visual perception
  - Action prediction
- Levels of hierarchy
  - Depth-aware input layering
  - Multi-scale visual representations
  - Hierarchically conditioned diffusion

[1] Lu et al. '$H^3DP$: Triply-Hierarchical Diffusion Policy for Visuomotor Learning', arXiv'25

# Depth-Aware Layering

- $m$: layer index
- $N$: number of layers
- $d$: depth value
- $d_{min}, d_{max}$: depth bounds

## Linear-Increasing Discretization

$$m = \left\lfloor a\sqrt{b \cdot N \frac{d - d_{\text{min}}}{d_{\text{max}} - d_{\text{min}} + \epsilon}} \right\rfloor$$

# Depth-Aware Layering (cont.)

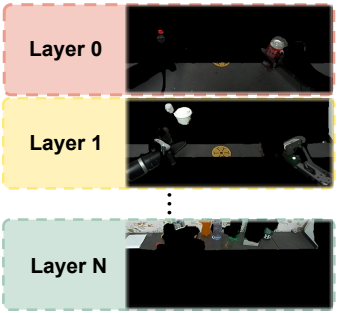
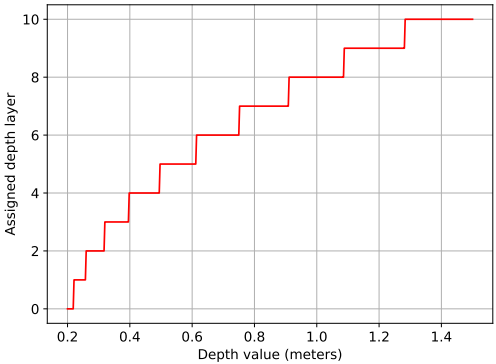



Figure: Lu et al.[1]




[1] Lu et al. '$H^3DP$: Triply-Hierarchical Diffusion Policy for Visuomotor Learning', arXiv'25

# Multi-Scale Visual Representation

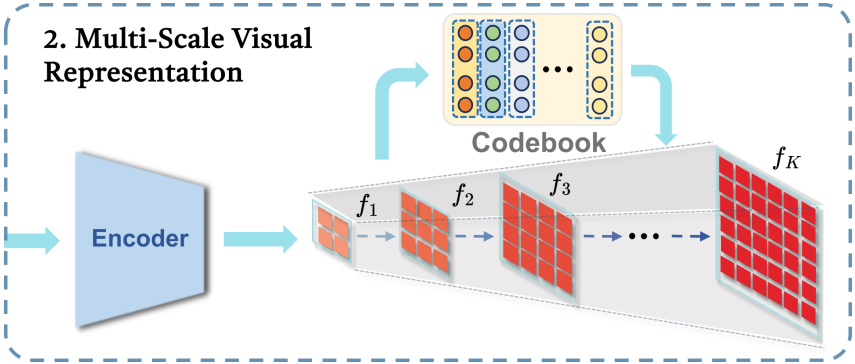



Figure: Lu et al.[1]

[1] Lu et al. '$H^3$DP: Triply-Hierarchical Diffusion Policy for Visuomotor Learning', arXiv'25

# Hierarchical Action Generation

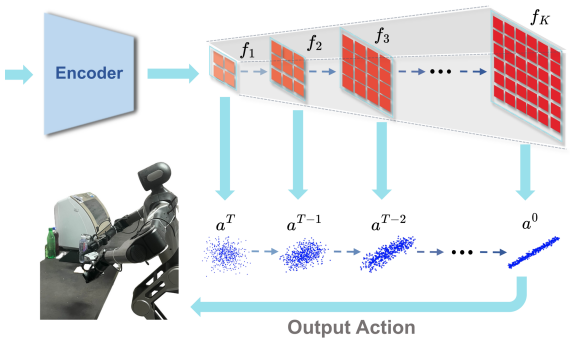



Figure: Lu et al.[1]

[1] Lu et al. '$H^3DP$: Triply-Hierarchical Diffusion Policy for Visuomotor Learning', arXiv'25

# 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**) |
|---|---|---|---|---|---|---|---|---|---|
| **$H^3DP$** | **98.3** | **87.8** | **95.8** | **59.3** | **65.3** | **87.3** | 53.3 | **57.4** | **75.6**$\pm$18.6 |
| DP | 78.2 | 52.6 | 58.0 | 22.3 | 27.5 | 79.0 | 44.3 | 22.8 | 48.1$\pm$23.1 |
| DP (w/ depth) | 77.7 | 57.2 | 71.2 | 44.5 | 40.8 | 76.0 | 42.0 | 12.6 | 52.8$\pm$22.2 |
| DP3 | 89.1 | 52.6 | 88.4 | 26.5 | 33.5 | 84.0 | **54.8** | 45.9 | 59.3$\pm$24.9 |

Figure: Lu et al.[1]

[1] Lu et al. '$H^3$DP: Triply-Hierarchical Diffusion Policy for Visuomotor Learning', arXiv'25

# Conclusion

- Diffusion-based policy learning shows to be:
  - Generalizable
  - Sample-efficient
- Visual features play a key role
- Cohesive vision-action coupling improves performance