Robot-DIFT: Correspondence-Sensitive Diffusion Features for Contact-Rich Robot Manipulation

Yu Deng[1,4*] Yufeng Jin[1,6*] Xiaogang Jia[2] Jiahong Xue[1] Gerhard Neumann[2,3] Georgia Chalvatzaki[1,4,5]

[1] TU Darmstadt [2] KIT [3] FZI [4] Hessian.AI [5] Robotics Institute Germany [6] Honda Research Institute Europe GmbH

* Equal contribution

yudeng321.github.io/Robot-DIFT/

CoRL 2026 · Paper 1930

Website

Robot-DIFT preserves local correspondence cues in a reusable diffusion backbone for precise robot manipulation.

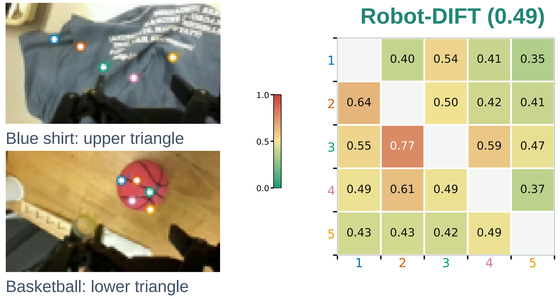
1) High-precision manipulation

Which visual details does precise control require?

- Object recognition can miss **precise alignment**.
- Small pose offsets can require **opposite actions**.
- Goal: retain **spatial cues** for precise control.

Local point discrimination

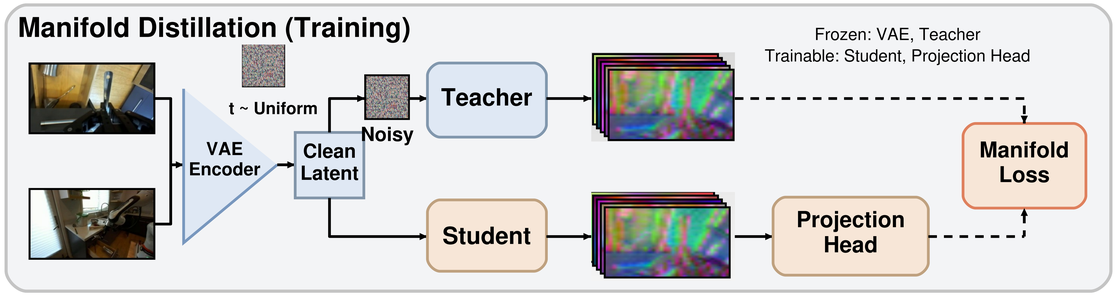
- Five surface points per object.
- Lower similarity: more **distinct local features**.



2) Robot-DIFT: Manifold Distillation + S2-FPN

DROID [3] pretraining with Manifold Distillation

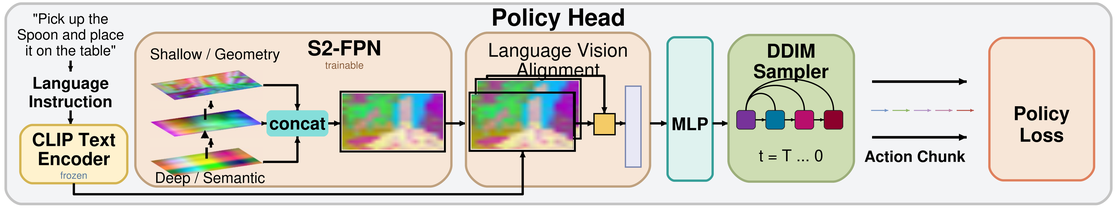
- **Imitation learning** adapts the Student to robot observations and actions.
- **Annealed feature alignment** preserves spatial cues from the frozen diffusion Teacher.



Imitation loss + λ(t) × feature alignment loss

Diffusion features: DIFT [1]; SD2.1 Teacher [4]; alignment follows CleanDIFT [2].

S2-FPN: coarse semantics + fine spatial detail



- Fuse **three decoder scales** before policy pooling.
- **Freeze the Student**; train a new readout and policy per task.
- **Single-pass Student**; no Teacher at deployment.

3) RoboCasa and LIBERO-10

- **RoboCasa**: controlled representation comparison.
- **LIBERO-10**: language-conditioned transfer.

Average task success (%)

RoboCasa

Method	Success
CLIP	19
SigLIP	27
VGGT	22
FastVGGT	23
DINOv2	28
DINOv3	33
SiT	29
DIFT	38
Robot-DINO	40
Robot-DIFT	**48 ± 1**

LIBERO-10

Method	Success
DP-C	53
DP-T	58
OpenVLA	54
π_0	85
π_0-FAST	60
UniPi	0
UVA	90
DIFT-VLA	68
Robot-DIFT	**92 ± 1**

Robot-DIFT: 3-seed mean ± SD; baselines: 1 seed.
50 rollouts/task.

- **Highest average success** among the evaluated methods on both benchmarks.
- Gains are strongest where **precise local alignment** matters.

Selected references

[1] Tang et al. DIFT. NeurIPS 2023.

[2] Stracke et al. CleanDIFT. CVPR 2025.

[3] Khazatsky et al. DROID. RSS 2024.

[4] Rombach et al. LDM. CVPR 2022.

[5] Siméoni et al. DINOv3. arXiv 2025.

[6] Zhai et al. SigLIP. ICCV 2023.