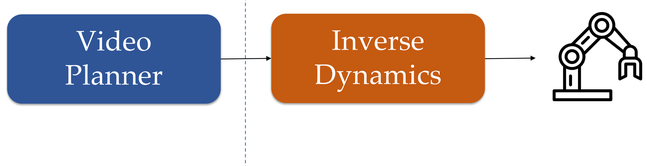
Learning Skill Abstraction from Action-Free Videos

Hung-Chieh Fang[1]*, Kuo-Han Hung[1]*, Chu-Rong Chen[1], Po-Jung Chou[1], Chun-Kai Yang[1],
Po-Chen Ko[1], Yu-Chiang Frank Wang[12], Yueh-Hua Wu[2], Min-Hung Chen[2], Shao-Hua Sun[1]

[1]National Taiwan University [2]NVIDIA

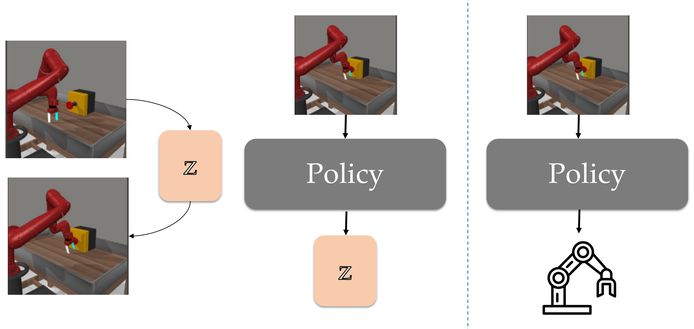
Learning from Videos

Video for Decision Making [1]



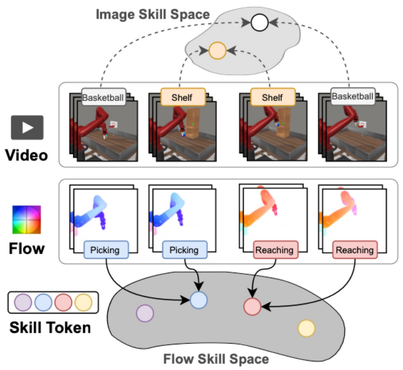
- Hard to translate to low-level actions
- Generating video plans is slow
+ Can train video planner on large-scale videos

Latent Action [2]



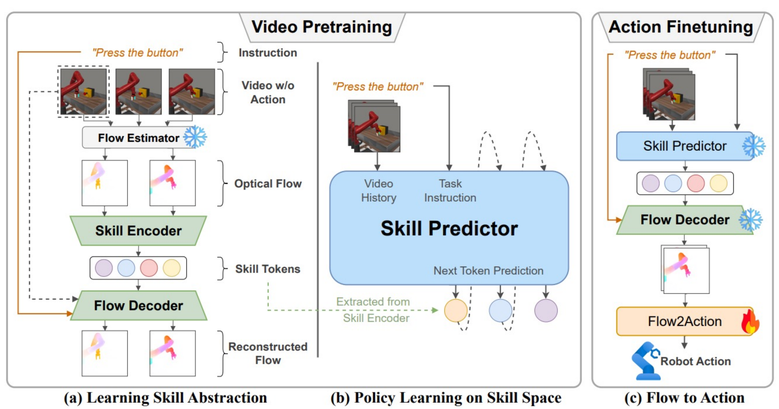
- Hard to capture fine-grained motion
- Single-step actions struggle with long-horizon
+ Better translates to low-level actions

Skills as Shared Structure for Videos



Flow as *action-surrogate* repr. for skill learning

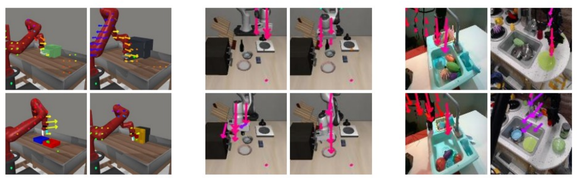
+ Better long-horizon planning with skills
+ Flow allows easier translation to low-level actions



(a) Learning Skill Abstraction (b) Policy Learning on Skill Space (c) Flow to Action

Experiments

Skill Plan Analysis



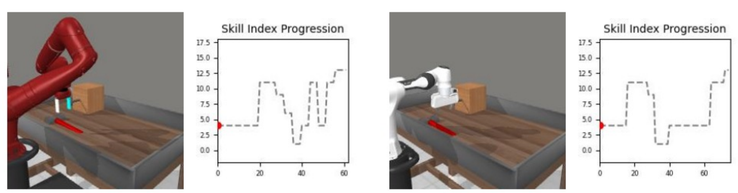
(a) Different Tasks & Scenes (b) Different Objects & Positions (c) Real World

Multi-task Learning

	door-open	door-close	bin-picking	box-close	drawer-open
BC	0.64 ± 0.06	**1.00** ± 0.00	0.00 ± 0.00	**0.20** ± 0.07	0.63 ± 0.02
DP	0.00 ± 0.00	0.84 ± 0.05	0.00 ± 0.00	0.00 ± 0.00	1.00 ± 0.00
AVDC	0.84 ± 0.04	0.92 ± 0.04	0.00 ± 0.00	0.04 ± 0.00	0.02 ± 0.02
LAPA	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
SOF (Ours)	**0.98** ± 0.03	**1.00** ± 0.00	**0.24** ± 0.07	0.12 ± 0.07	**0.78** ± 0.04

	faucet-close	faucet-open	handle-press	assembly	**Overall**
BC	**0.78** ± 0.04	**1.00** ± 0.00	**0.87** ± 0.03	0.00 ± 0.00	0.57 ± 0.01
DP	0.06 ± 0.02	0.86 ± 0.07	0.00 ± 0.00	0.00 ± 0.00	0.31 ± 0.01
AVDC	0.24 ± 0.04	0.78 ± 0.02	0.72 ± 0.04	0.00 ± 0.00	0.42 ± 0.02
LAPA	0.17 ± 0.08	0.28 ± 0.11	0.65 ± 0.11	0.12 ± 0.04	0.14 ± 0.02
SOF (Ours)	0.62 ± 0.06	**0.99** ± 0.02	0.69 ± 0.06	**0.82** ± 0.07	**0.69** ± 0.02

Cross-embodiment Learning



(a) Sawyer (b) Panda

Reference

[1] Du et al. "Learning Universal Policies via Text-Guided Video Generation." NeurIPS'23
[2] Ye et al. "Latent Action Pretraining from Videos." ICLR'25