ESPADA: Execution Speedup via Semantics-Aware Demonstration Data Downsampling for Imitation Learning

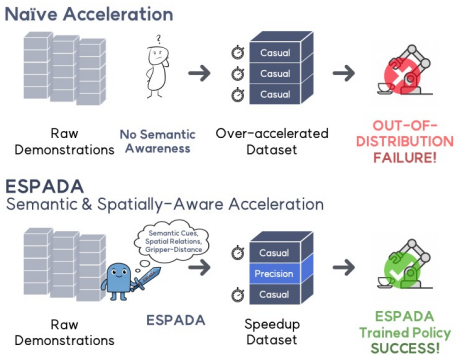
Byungju Kim[1,2]* Jinu Pahk[1,2]* Chungwoo Lee[1]* Jaejoon Kim[1,2]* Jangha Lee[1,2]* Theo Taeyeong Kim[1,2] Kyuhwan Shim[2] Jun Ki Lee[2] Byoung-Tak Zhang[1,2]†

[1] Tommoro Robotics [2] Seoul National University * Equal contribution † Corresponding author

IROS 2026 PITTSBURGH
IEEE/RSJ IROS 2026

1 Motivation

Behaviour-cloned policies inherit the slow tempo of their demonstrations. Replaying faster or subsampling uniformly pushes the policy out of distribution.



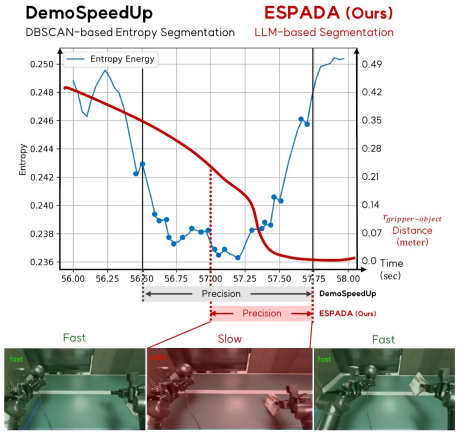
2 Challenges

Prior accelerators segment with motion statistics, not task intent — and action entropy is not a proxy for precision. It fails both ways:

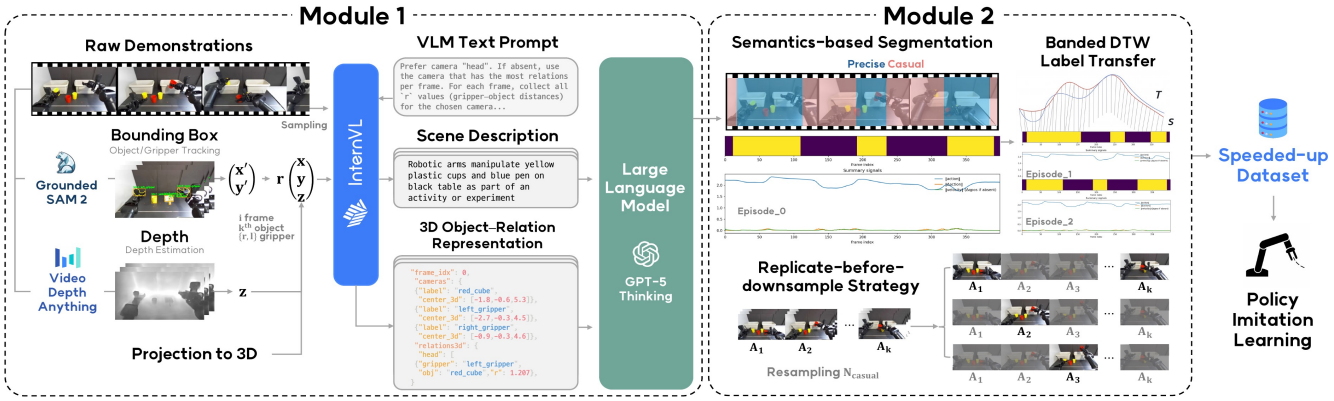
High entropy, yet precision-critical. On *Conveyor* the arm waits for the object; DemoSpeedup calls it casual and descends early › **1/20**.

Low entropy, yet casual. Repetitive scooping stays locked as precision, so the available speed-up is never taken.

Method	Segmentation signal	Segmentation · real-world outcome
Uniform 2×	one rate for every phase	out-of-distribution failure
SAIL	waypoint-based clustering	strong compression, lowest success (BiGym 61% @ 2.4×)
DemoSpeedup	proxy-policy action entropy	seg. IoU 0.114 · real ACT 36.0% @ 2.59×
ESPADA (Ours)	**3D relation + VLM scene semantics**	**seg. IoU 0.261 · real ACT 90.0% @ 2.26×**



3 Overview of ESPADA



TL;DR — acceleration as **semantic phase segmentation**. A 3D scene graph and a VLM summary are read by an LLM to split every demo into **casual** and **precision** spans; banded DTW propagates one seed episode; RBD compiles the sped-up data. No extra data, no extra sensors, no policy change.

4 Method

1 3D scene graph

Grounded-SAM 2 tracks grippers and objects; **Video Depth Anything** gives depth. Back-projection yields the relation $r_t(g,o)$.

2 VLM episode summary

InternVL-3.5 8B reads **4-8 sampled frames** › strict-JSON episode summary: objects, activity, stage order.

3 LLM segmentation

GPT-5 Thinking reads the r_t series *together with* the VLM descriptor. Near-contact plateau + micro-adjustment ⇒ **precision**; monotonic approach or retreat ⇒ **casual**. Gripper events are forced to precision.

4 Banded DTW label transfer — 1 seed › whole dataset

Only episode 0 is annotated. Proprioception/action features are aligned under a Sakoe–Chiba band; the monotone map transfers every boundary. **CPU-only, no proxy policy.**

5 RBD + geometric consistency

Replicas at offsets $j = 0 \cdots N-1$ keep every original frame, so state coverage survives compression. **2× precision / 4× casual**, with $K' \approx K/2$ preserving per-chunk displacement.

Both cues are necessary

Ablated cue	Insertion IoU	Transfer Cube IoU
w/o 3D relation r_t	0.0224	0.2791
w/o VLM description	0.1024	0.0584
w/o both	0.1111	0.0693
Full model (Ours)	**0.5166**	**0.3064**

Table 2. Segmentation IoU vs. ground truth — dropping either cue collapses one task.

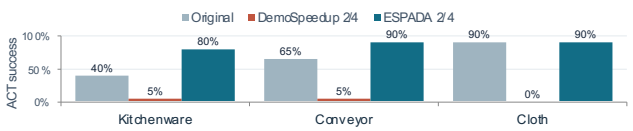
5 Experiments

ACT and DP on five real-world AI-Worker tasks, two ALOHA and seven BiGym tasks. Q = success × normalised speed.



5.1 Real-world execution

Policy / data	Success	Speed	Q
ACT Original	76.3%	1.00×	0.76
ACT + DS 1/3	52.7%	1.34×	0.71
ACT + DS 2/4	36.0%	2.59×	0.93
ACT + ESPADA 1/3	81.7%	1.35×	1.10
ACT + ESPADA 2/4	**90.0%**	**2.26×**	**2.03**
DP Original	38.0%	1.00×	0.38
DP + DS 2/4	69.7%	2.52×	1.76
DP + ESPADA 2/4	**82.3%**	**2.59×**	**2.13**



Per-task ACT success at matched 2×/4× — Kitchenware 16/20 vs 1/20, Conveyor 18/20 vs 1/20.

5.2 Simulation comparison (ACT)

Policy / data	ALOHA Q	BiGym S	Speed	Q
Original	0.47	74%	1.0×	0.74
Uniform 2×	0.75	66%	1.7×	1.12
SAIL	—	61%	2.4×	1.44
DemoSpeedup	1.13	78%	2.1×	1.64
ESPADA (Ours)	**1.12**	**73%**	**2.3×**	**1.68**