Injecting Point-Cloud Geometry into Frozen pi0 for Robust Real-World Manipulation

Tersoo Samuel Awai | Enkhbaatar Batbayar | HeeKong Lee

4DiVISION, Seoul, South Korea

IEEE ARM 2026 Workshop SSCR

AI-Driven Manipulation with 3D Vision at 4DiVISION

- 4DiVISION in-house 3D cameras - RGB + point cloud, eye-in-hand and eye-to-hand
- Goal: AI-driven industrial arm manipulation using VLA models
- Phase 1: Point-cloud VLA validation on SO-101
- Phase 2: Transfer to Yaskawa MH5LF
- Phase 3: CBF safety layer for human-adjacent deployment

Introduction and Problem Statement

The Problem with RGB-only VLAs:

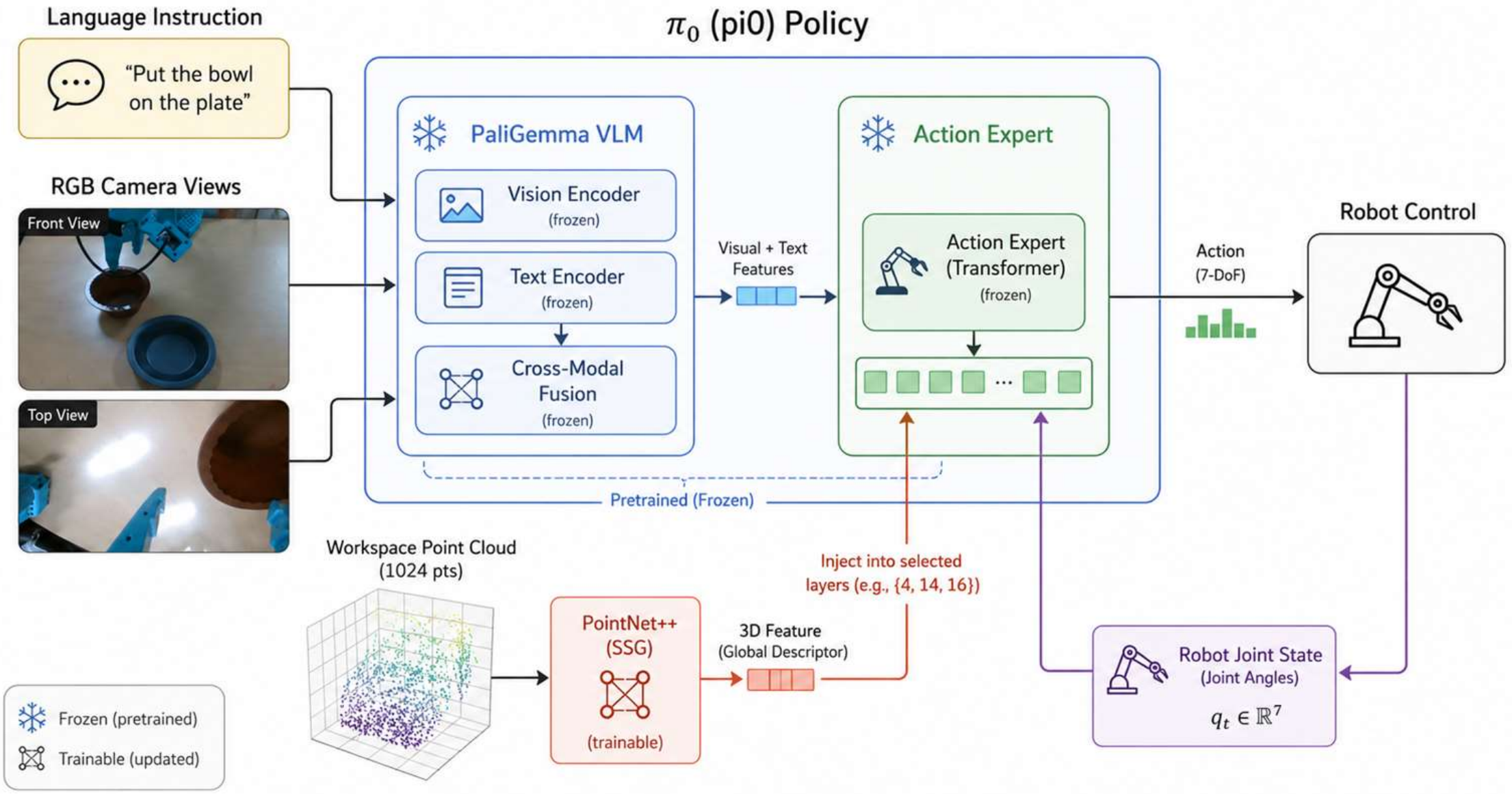
- Pi0 encodes 3D geometry implicitly from RGB
- Brittle under viewpoint shifts, lighting change, and objects near camera frame edges.
- No explicit 3D spatial awareness

Our Approach

- Adapt the PointVLA injection strategy to pi0

(A) SO-101 follower arm
(B) RealSense D435i(fixed-stand, RGB+depth, world view)
(C) End-effector webcam (gripper view)

Pi0 Policy Architecture



Main Results

The 3D model **matches or beats** the 2D baseline on every condition, **with the largest gain at frame-edge positions (60% → 90%)**.

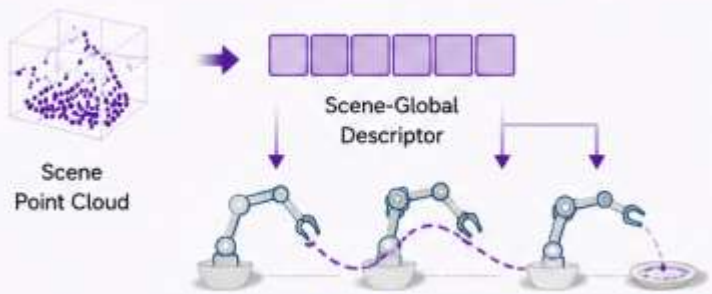
The 3D model **underperforms** the 2D baseline on three of five conditions for **flat placement**.

CONDITION	ANY SUCCESS (↑)		FLAT PLACEMENT (↑)	
	2D (RGB only)	+3D	2D (RGB only)	+3D
CROSS-SCENE				
Bowl position	13 / 15	13 / 15	13 / 15	9 / 15
Lighting	12 / 15	14 / 15	12 / 15	11 / 15
Bowl colour	14 / 15	14 / 15	14 / 15	8 / 15
Object height	9 / 10	10 / 10	9 / 10	8 / 10
IN-SCENE, FRAME-EDGE				
Edge positions	6 / 10	9 / 10	6 / 10	7 / 10

Better than 2D baseline Worse than 2D baseline Similar to 2D baseline

MECHANISM (BRIEF)

A **single scene-global descriptor**, broadcast uniformly across the action chunk, improves **where** the gripper reaches but carries no per-region information to constrain **how** the wrist orients at release.



Thank you for listening