Carnegie Mellon University
Robotics Institute

Enhancing 2D VLA Models with 3D Spatial Awareness

Kailash Jagadeesh, Keerthi G V

Problem Statement

Current **VLA** models only **see the world in 2D**, which limits their ability to interpret depth, object geometry, and occlusions.

Although 3D-aware VLA models are emerging, they demand **significant compute** resources and **massive 3D dataset**, making them **impractical**.

Can we extract 3D scene representations from 2D input?

Research Objective

1. To check if "**software-only 3D**" can replace actual depth sensors.
2. To understand the quality **gap** between **real and predicted 3D.**
3. To test if VLAs benefit from **lightweight 3D cues.**

Pain Points Addressed

1 No Specialized 3D Hardware Needed

2 Leverages widely available 2D datasets

3 Makes VLAs capable of geometric reasoning

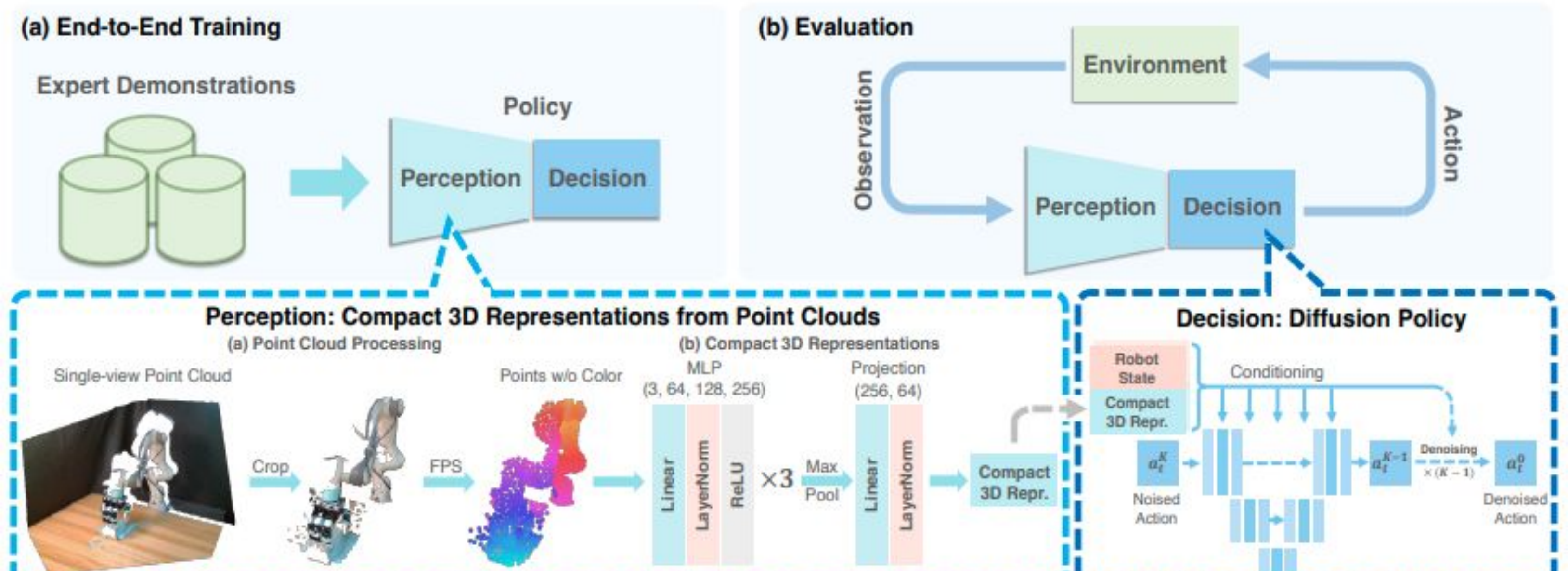
4 Avoids overfitting to perfect simulator depth

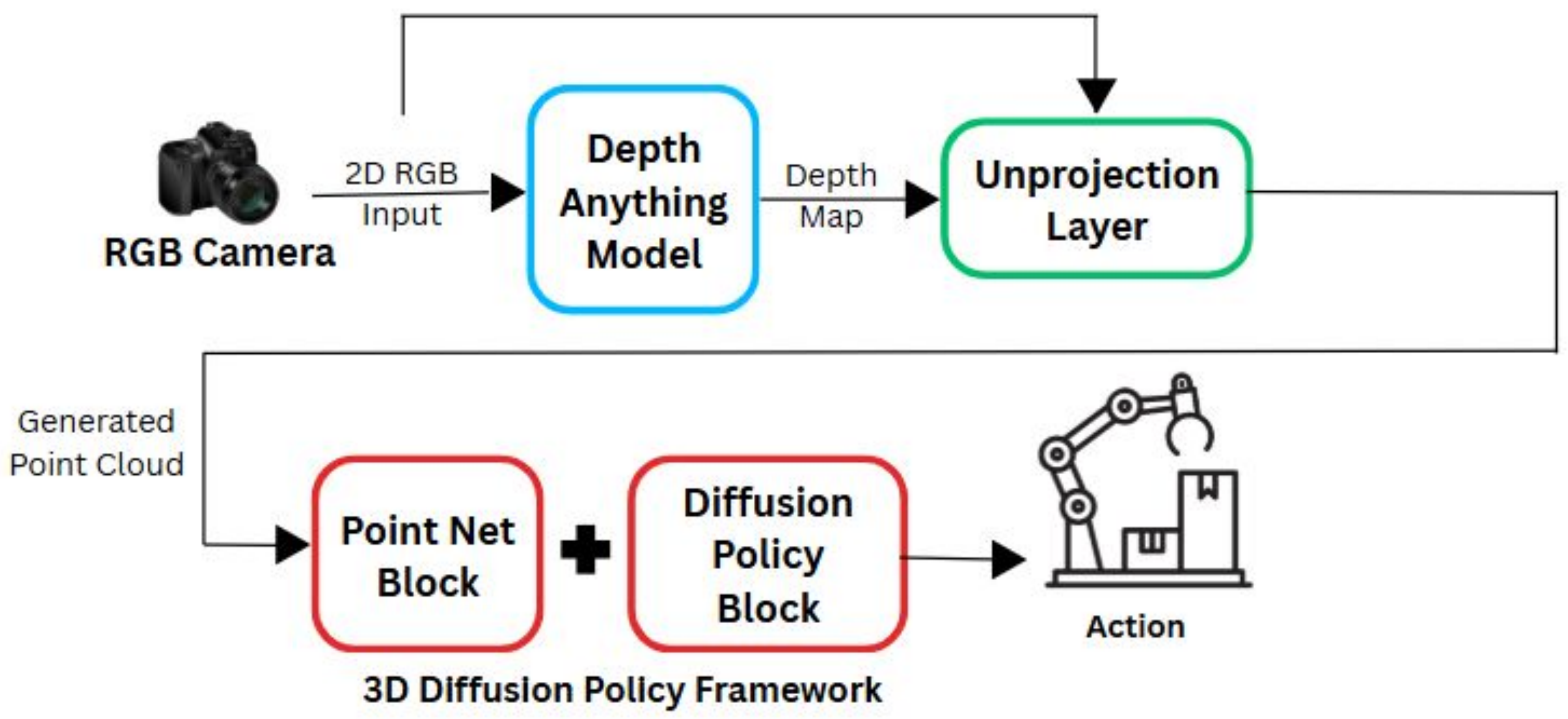
5 Enables real-world scalability

Our Solution

This project investigates the practicality of **converting 2D RGB inputs** into rich **3D scene** representations for Vision-Language-Action models.

Our baseline architecture is the **3D Diffusion Policy**, a robot control framework that uses a **diffusion model** to learn and generate actions directly from **3D point clouds**.



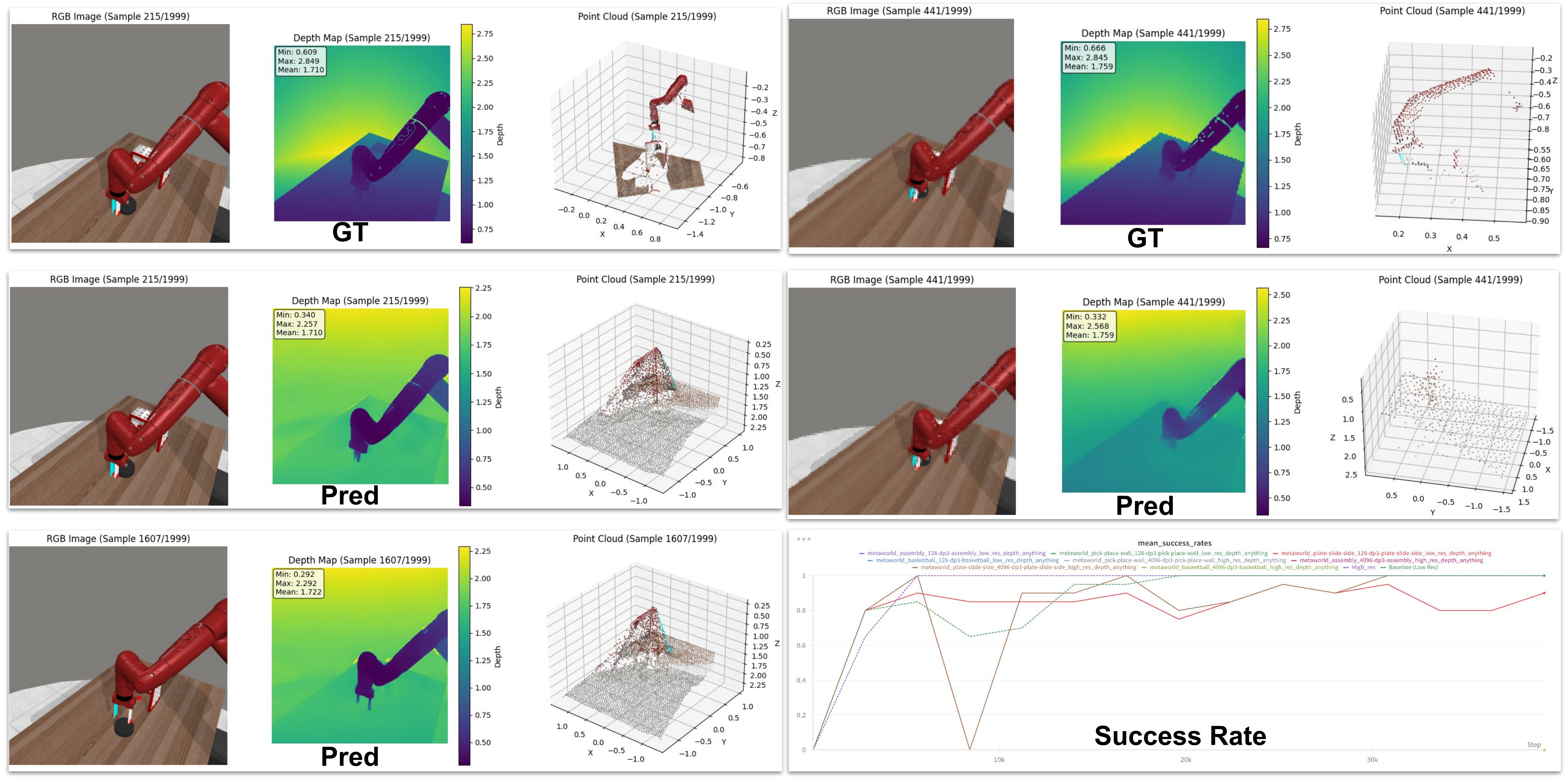


Results

PIPELINE 1 : Depth Anything V3

Task: Slide Plate High Res (512*512;4096 points)

Task: Slide Plate Low Res (128*128;512 points)



Task: Assembly High Res (512*512;4096 points)

Task: Assembly Low Res (128*128;512 points)

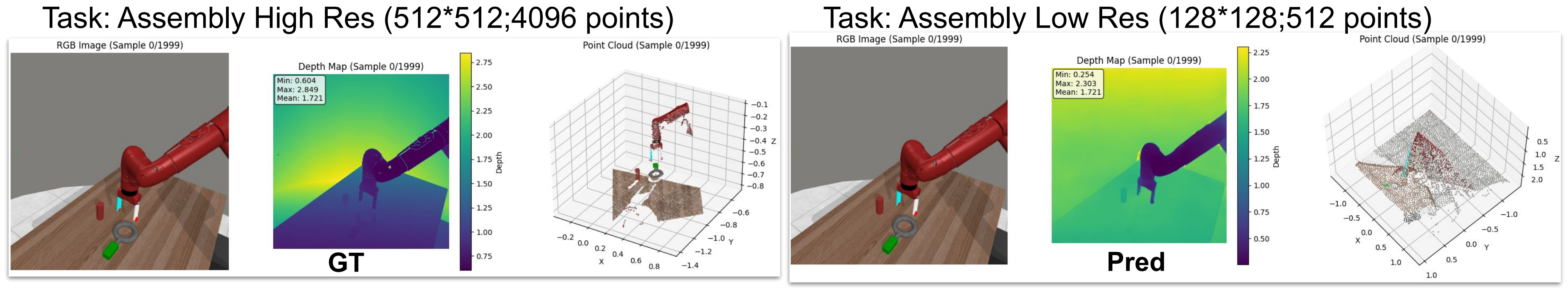
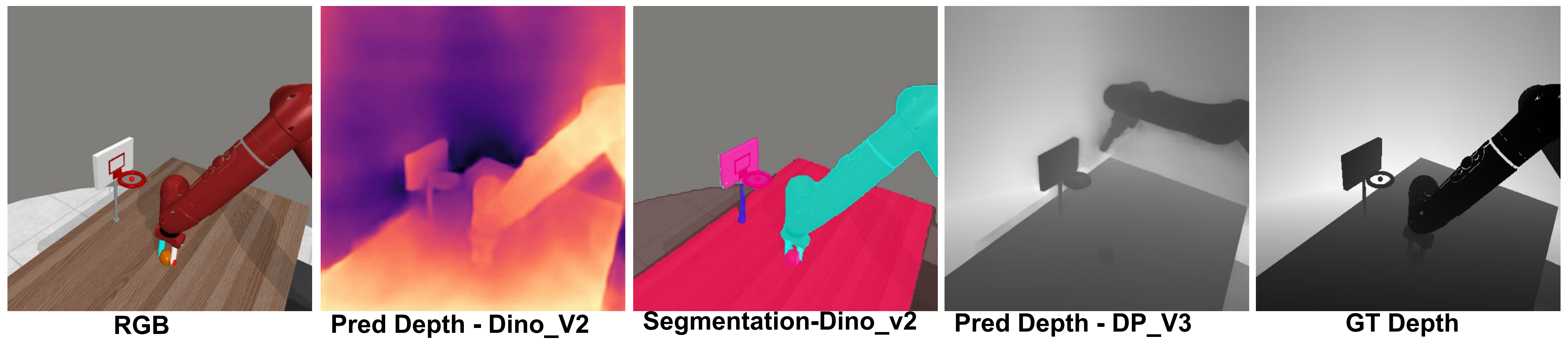


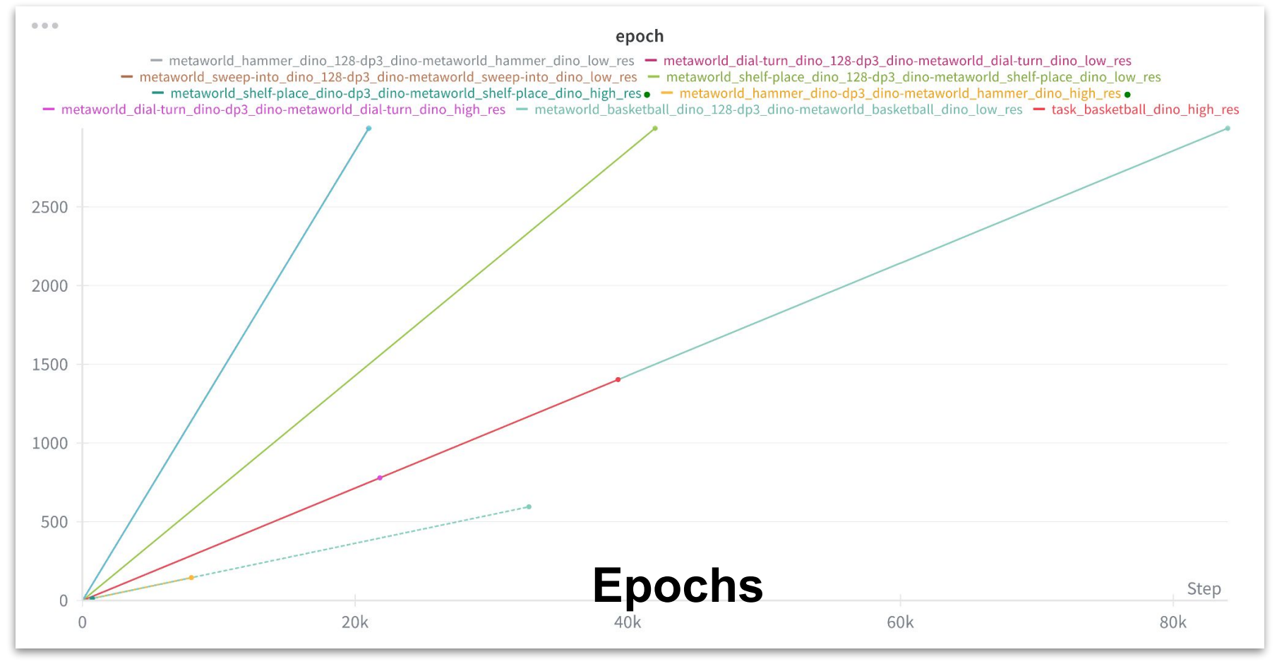
Table 1: Depth Anything v3 Pipeline (PointNet Encoder + Pseudo-3D Point Clouds)

Task	Res	Image Size	Num Points	Success Rate (%)
plate-slide-side	High	512×512	4096	100.0
plate-slide-side	Low	128×128	512	90.0
assembly	High	512×512	4096	0.0
assembly	Low	128×128	512	0.0
pick-place-wall	High	512×512	4096	0.0
pick-place-wall	Low	128×128	512	0.0
basketball	High	512×512	4096	0.0
basketball	Low	128×128	512	0.0

PIPELINE 2: Dino V2 Encoder

Task: Assembly High Res (512*512;4096 points)





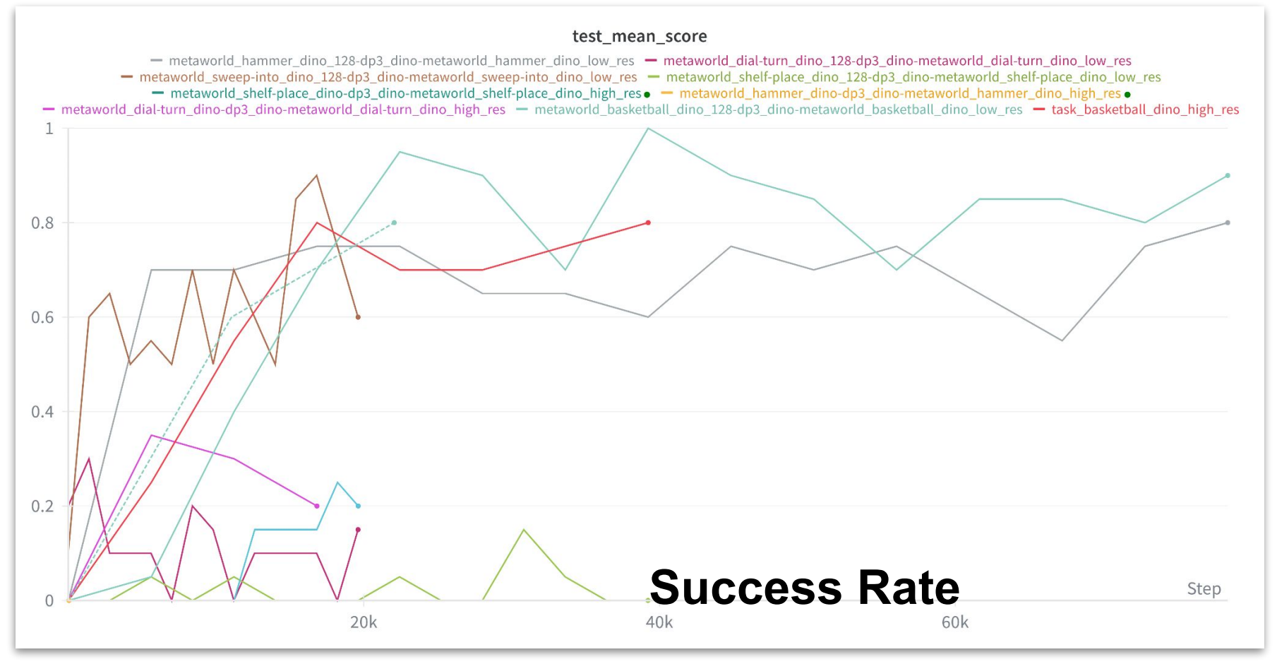
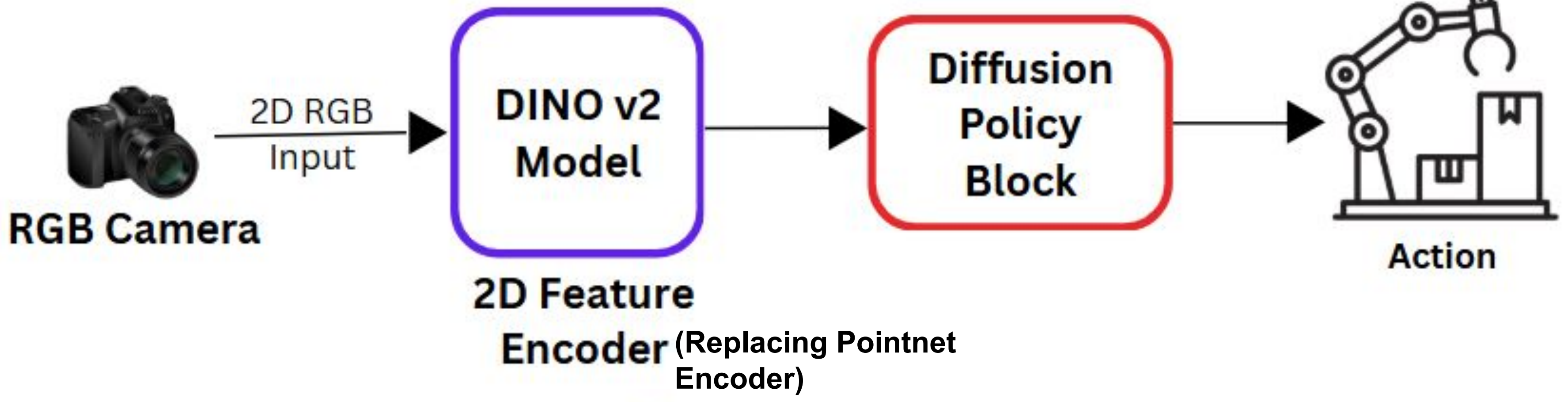


Table 2: DINO v2 Pipeline (dinov2_vits14 Backbone, 384-dim features)

Task	Res	Image Size	Success Rate (%)
hammer	Low	128×128	85.0
basketball	Low	128×128	80.0
basketball	High	512×512	50.0
sweep-into	Low	128×128	50.0
dial-turn	Low	128×128	15.0
shelf-place	Low	128×128	5.0
dial-turn	High	512×512	5.0
hammer	High	512×512	0.0

PIPELINE 2 :



Future Work

1. **DinoV2** delivered good results with single-view high-resolution images; hence, in the future, we would like to extend the architecture to **multi-view inputs**, as there is clear potential for **enhanced 3D reasoning**.

2. The **Depth Anything model** used to generate point clouds from RGB inputs produced sub-optimal results, which in turn degraded DP3's performance. This motivates us to explore stronger depth models such as **UniK3D, MiDaS, and Monodepth2.**