UniVEP | Unifying Velocity and Endpoint Prediction for Robot Manipulation

Complementary Strengths of X and V Prediction

Schematic of stage-dependent strengths

	High noise	Low noise
X pred (endpoint)	Stable recovery	Overwrites input
V pred (velocity)	Noise sensitive	Precise refinement

Multiple Benchmarks



- **LIBERO** — Different layouts, same objects
- **LIBERO-Plus** — Varied camera views, layouts, backgrounds, lighting, etc.
- **RoboTwin** — Dual-arm tasks with rich diversity

StarVLA · FastWAM · LiLa-WAM

Joint training



Observations → Shared policy → X mode ↕ Consistency ↕ V mode

Joint objective

$$\mathcal{L}_{joint} = \mathcal{L}_x + \mathcal{L}_v + \mathcal{L}_{cons}$$

$\mathcal{L}_x$: X supervision · $\mathcal{L}_v$: V supervision · $\mathcal{L}_{cons}$: Endpoint consistency

Staged inference



Noise → X recovery → $\sigma_i = \sigma^*$ Switch → V refinement → Action



Higher success rate

baselines: 16 steps

X-pred · V-pred · X / V-loss

StarVLA · RoboTwin Clean

2 steps

Near-peak performance



0% extra FLOPs

latency change ↓

	X / V-loss	UniVEP
LiLa-WAM	16 steps	2 steps
	100.98 ms	23.16 ms

↑ +3.36 pp Success

↓ −77.1% Latency