TDSR-VLA: Transition-aware Denoising Sequence Representations for Vision-Language-Action

Dongwoo Kim[1,*], Keunho Song[1,*], Seungmin Lee[1,*], Hwanhee Ju[1], Eun Sang Cha[1,†], Daekyum Kim[1,2,†]

[1]Department of Mechanical Engineering, Korea University, [2]School of Smart Mobility, Korea University

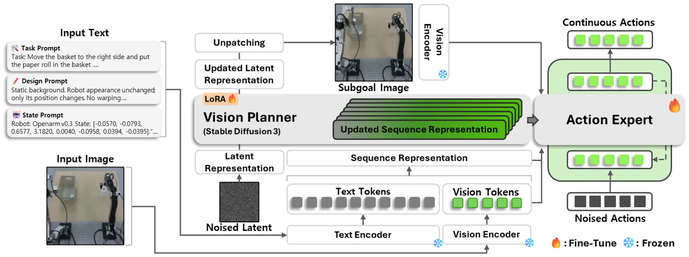
[*] Equal contribution. [†] Corresponding authors.

ECCV 2026 | MALMÖ | SEPT 8–12

Motivation

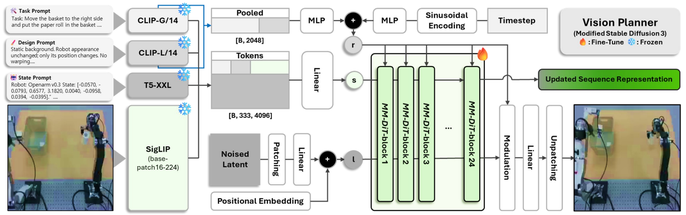
- VLA policies become **brittle under distribution shift** when only limited target-domain demonstrations are available.
- A subgoal image specifies **what** future scene to reach — not **how** the robot should transition toward it.
- TDSR-VLA reuses the Vision Planner's denoising states (**SR & USR**) as transition-aware cues for action generation.
- **SR (Sequence Representation)** grounds action generation in the current observation; **USR (Updated Sequence Representation)** is refined at each denoising step, encoding how to transition toward the subgoal.

Framework Overview

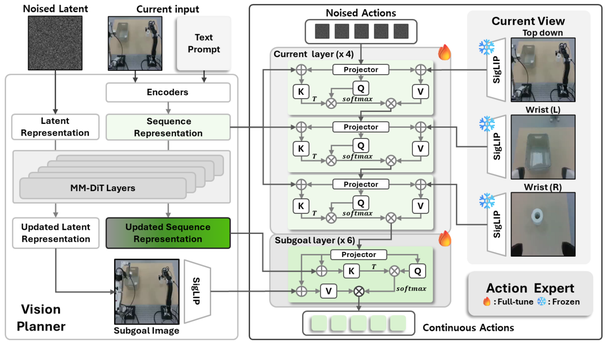


The **SD3-based Vision Planner** predicts a subgoal image Δ = 10 frames ahead; the **flow-matching Action Expert** generates 50-step action chunks via layer-wise prefix KV conditioning.

Vision Planner



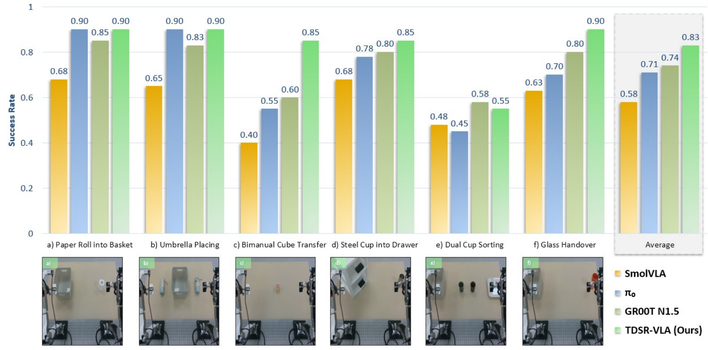
Action Expert



Current layers (×4) attend to SR grounded in the current observation; **subgoal layers (×6)** attend to USR refined during denoising — phase-specific conditioning for corrective, long-horizon control.

Real-World Results

Best overall OpenArm score **0.83** — vs GR00T N1.5 (0.74), π_0 (0.71), SmolVLA (0.58); **+0.18** on object-cooperative tasks.



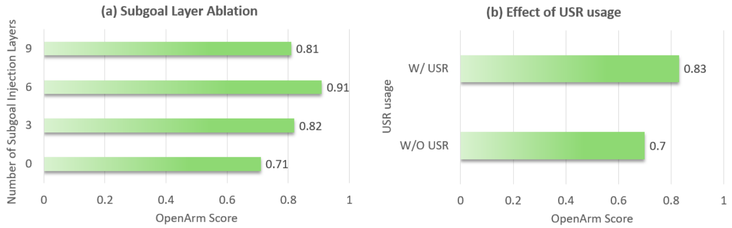
LIBERO Benchmark

Table 1: LIBERO Benchmark Results.

Models	Pre-training	LIBERO Goal	LIBERO Long
OpenVLA	Yes	79.2	53.7
π_0	No	89.0	48.0
TDSR-VLA (Ours)	No	82.0	56.0

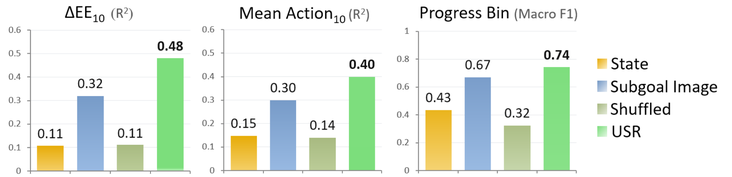
Without robot-demonstration pretraining: **82.0 on Goal** and **56.0 on Long** — above OpenVLA on both suites and π_0 on Long.

Ablation Studies



3–6 subgoal layers are optimal; using **USR** as subgoal-layer keys lifts OpenArm **0.70 → 0.83**.

Why USR Works



USR beats State, Subgoal-Image and Shuffled-USR probes → **timestep-aligned, action-useful cues**.