RILS LAB 세미나
(VLA)

RILS LAB

Robot Intelligence Learning & System

송재민

2026. 08. 14

목차

어디를 봐야 하는가

01

ReconVLA: Reconstructive Vision-Language-Action Model as Effective Robot Perceiver

Wenxuan Song[1], Ziyang Zhou[1], Han Zhao[2,3], Jiayi Chen[1], Pengxiang Ding[2,3], Haodong Yan[1], Yuxin Huang[1], Feilong Tang[4], Donglin Wang[2], Haoang Li[1]

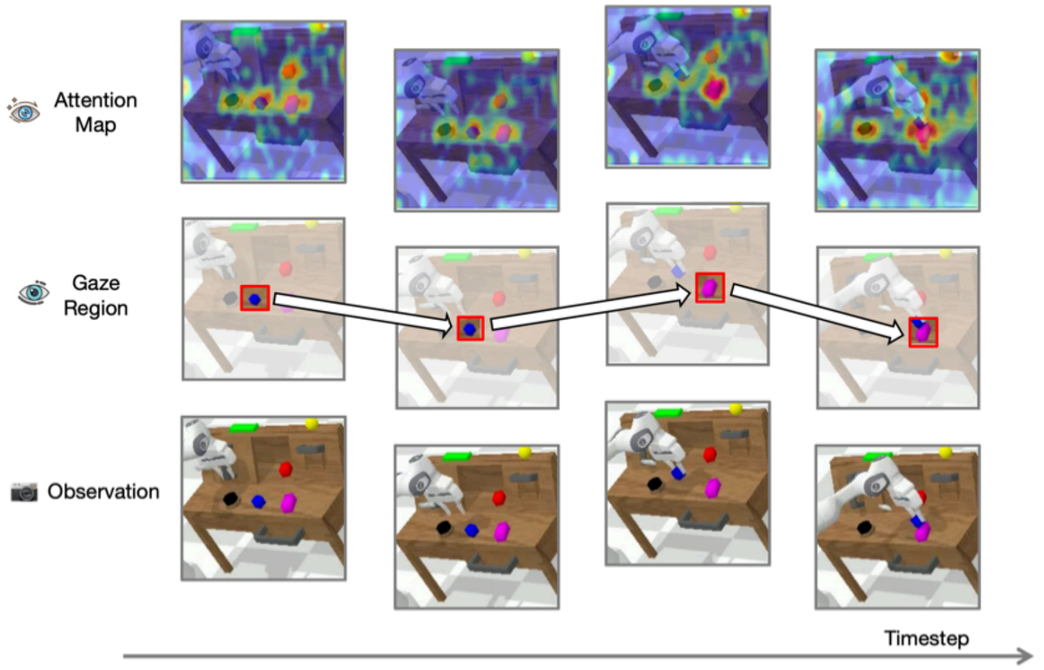
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]Westlake University [3]Zhejiang University [4]Monash University

AAAI/ 2026

문제상황

기존의 VLA 모델

기존 VLA는 행동 생성은 가능하지만, 조작해야 할 target object에 visual attention이 정확히 집중되지 않고 분산되는 문제가 있다.

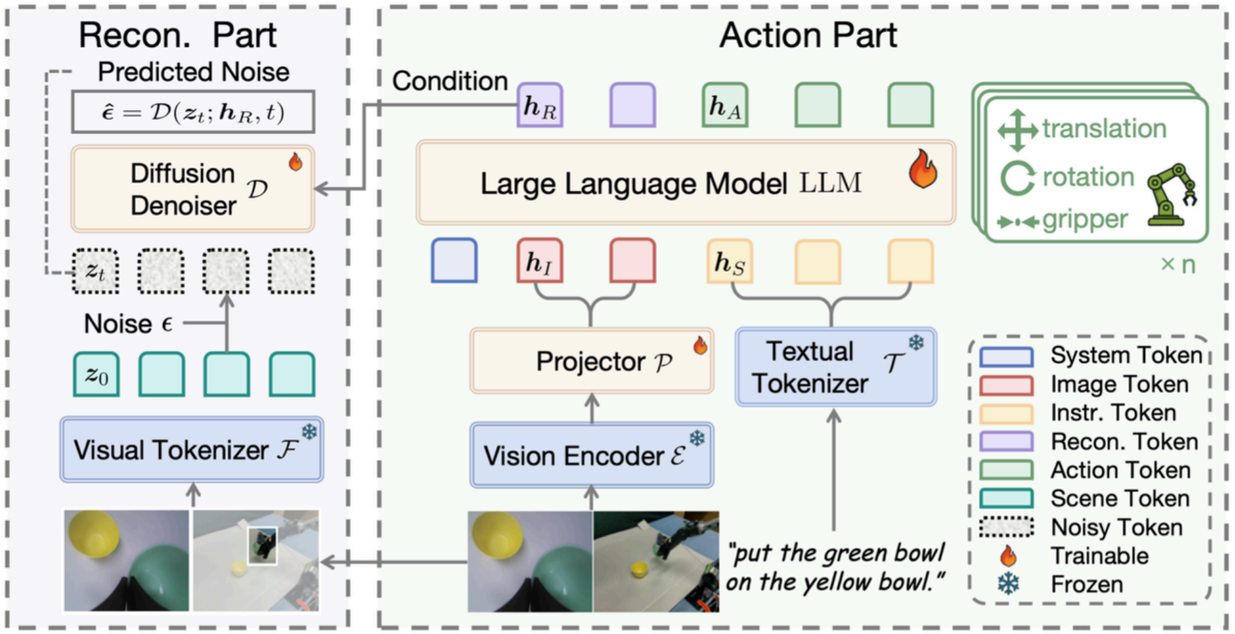


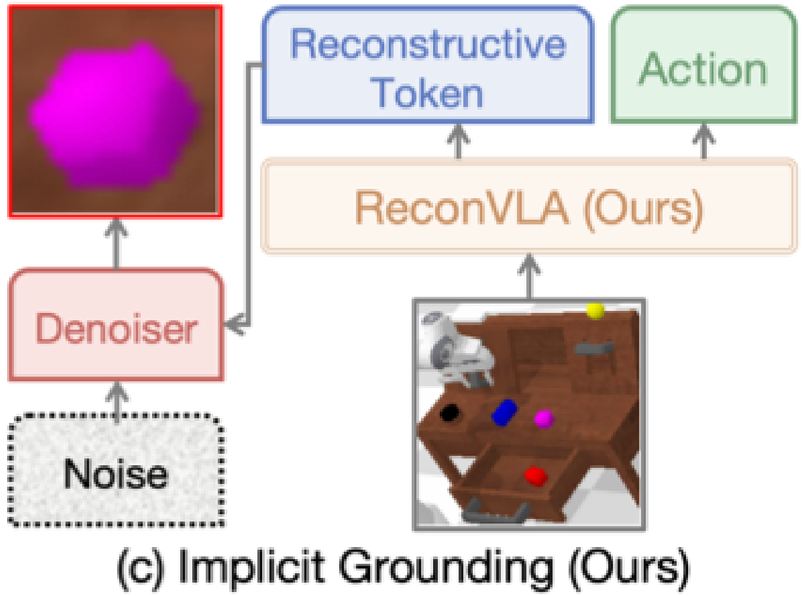
Action만 잘 예측하도록 학습시키지 말고, VLA 자체가 조작 대상에 제대로 집중하도록 만들자

핵심 아이디어

Gaze Region을 복원 하도록 학습

① Image + Instruction 입력 → ② 두 종류의 출력 생성 → ③ 학습 시 정답 Gaze Region 준비 →
④ Diffusion Denoiser가 Gaze Region 복원



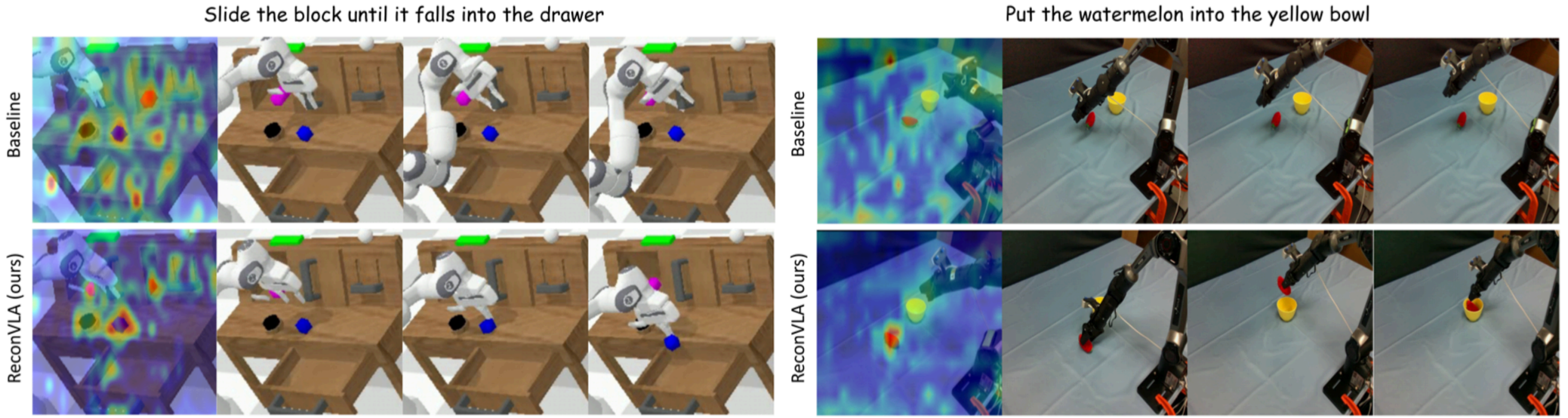


VLA가 스스로 중요한 영역에 집중하도록 학습

핵심 기여

더 정확한 Visual Grounding → 더 정확한 Manipulation 가능

Bounding box를 출력하거나 crop image를 추가 입력하지 않고, reconstruction supervision만으로 VLA 내부 visual grounding 능력을 강화했다. 아래 그림에서 baseline의 attention은 분산되어 있지만 ReconVLA는 실제 조작 대상에 집중



어디서 봐야 하는가

02

ActiveVLA: Injecting Active Perception into Vision-Language-Action Models for Precise 3D Robotic Manipulation

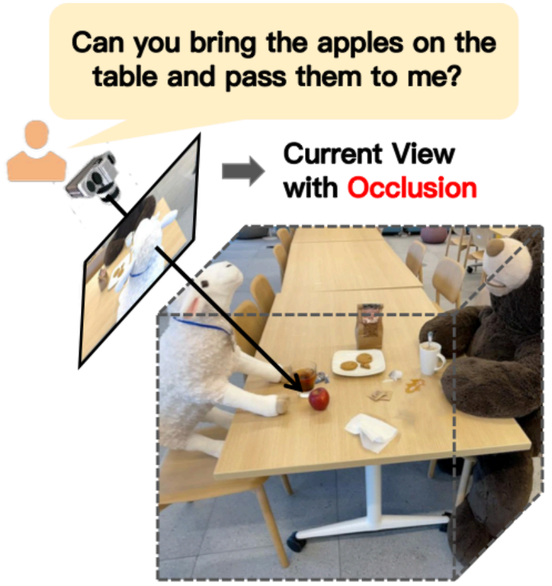
Zhenyang Liu[1,2], Yongchong Gu[1], Yikai Wang[3*], Xiangyang Xue[1†], Yanwei Fu[1,2†]
[1]Fudan University [2]Shanghai Innovation Institute [3]Nanyang Technological University

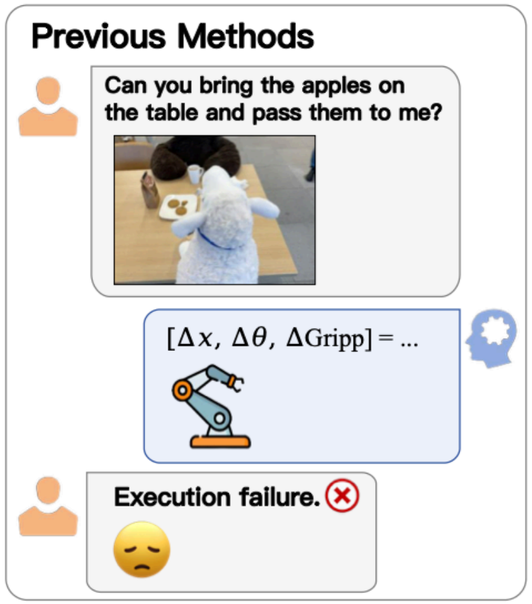
CVPR/ 2026

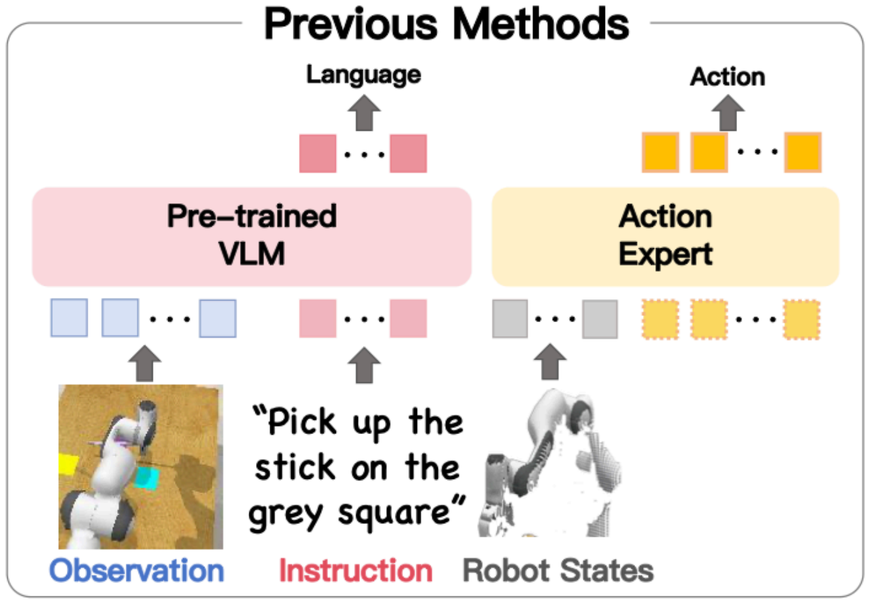
문제상황

기존의 VLA 모델

기존 VLA는 대체로 고정된 카메라나 wrist camera로 들어온 영상을 그대로 입력으로 사용하여 action을 출력한다.





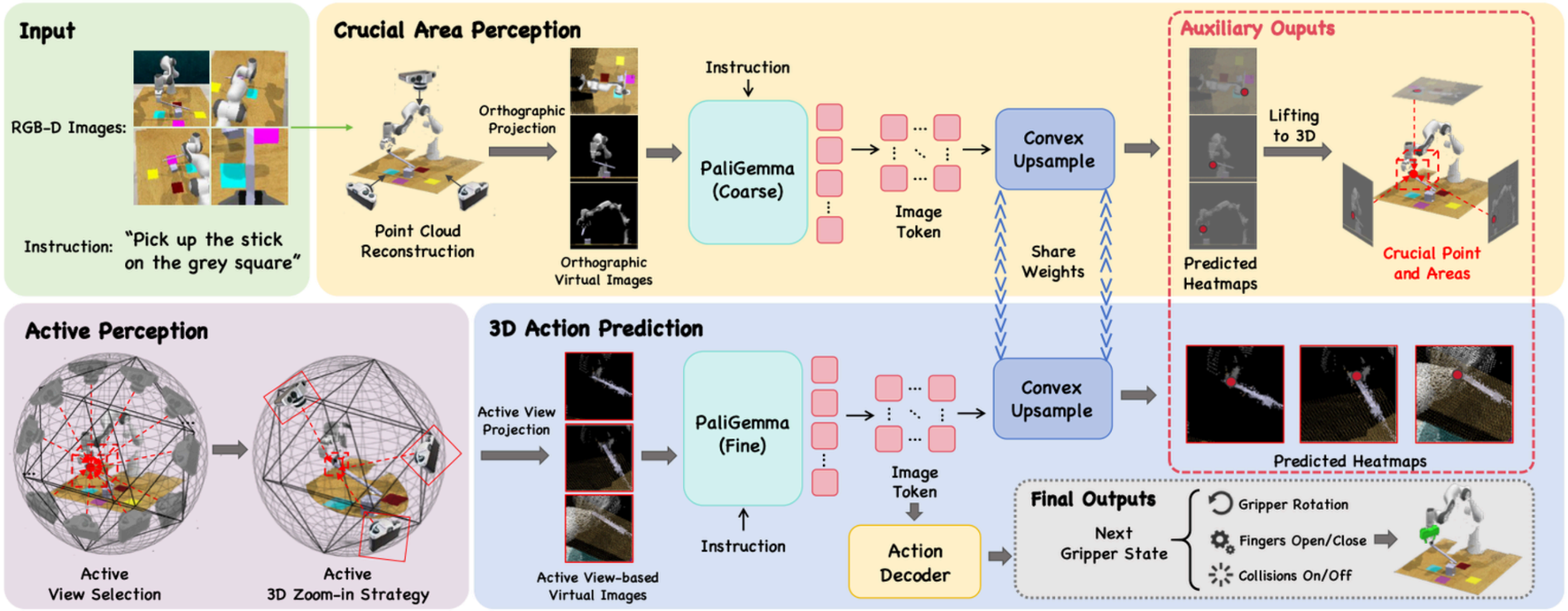


지금 보고 있는 화면이 애초에 안 좋으면, 로봇이 스스로 더 잘 보이는 시점을 찾아가자

핵심 아이디어

최종적으로 VLA에 들어가는 건 가상 카메라에서 렌더링한 2D 이미지

① (RGB-D images → 3D Point Cloud) → ② 3개 view로 렌더링 → ③ 세 이미지와 명령어를 PaliGemma에 넣음 →
④ 3개의 2D heatmap을 만들고 다시 3D로 합침 → ⑤ Active Viewpoint Selection



카메라가 찍어준 걸 먼저 3D로 만들고, 그 3D에서 내가 보기 좋은 이미지를 새로 만들어서 행동하자

핵심 기여

3D 공간에서 중요한 영역을 찾고 최적 가상 시점 선택

고정된 관찰에 의존하던 VLA를 넘어, 3D 장면에서 task-relevant 영역을 찾고 최적의 가상 시점을 능동적으로 선택해 정밀 조작 성능을 높였다

VLA에 적용하는 Reasoning

03

$LaST_0$: Latent Spatio-Temporal Chain-of-Thought for Robotic Vision-Language-Action Model

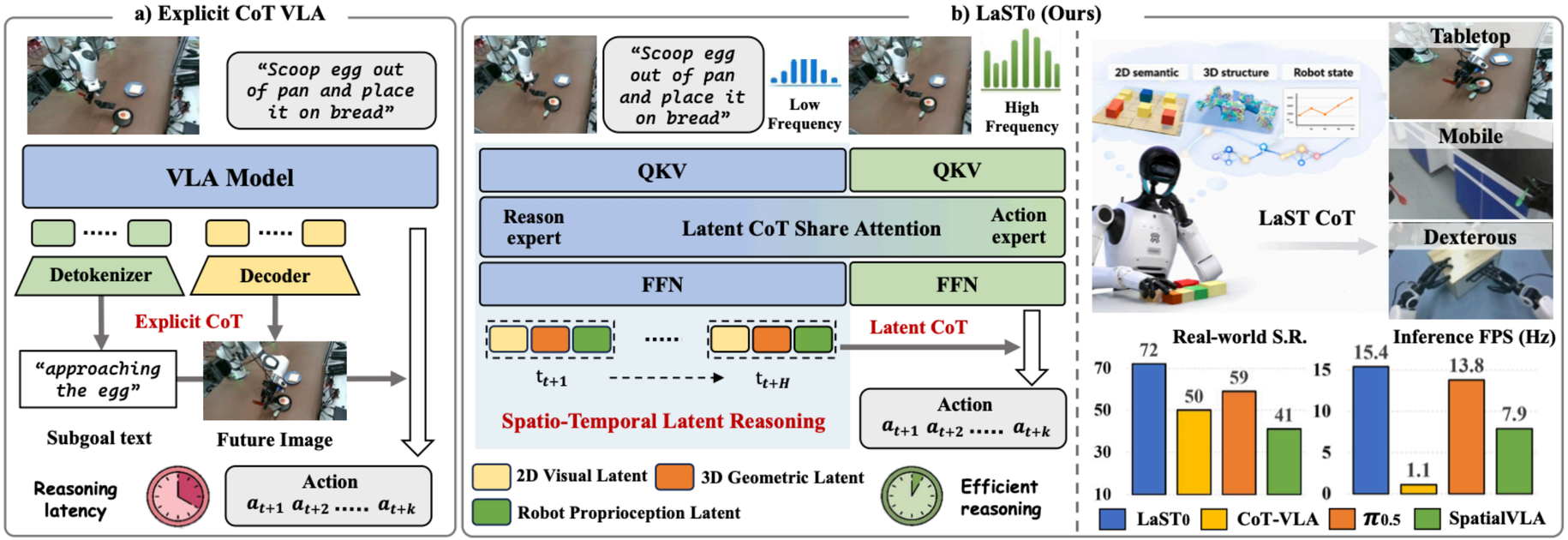
Zhuoyang Liu [*1] Jiaming Liu [*†1] Hao Chen [*2] Jiale Yu [1] Ziyu Guo [2] Chengkai Hou [13] Chenyang Gu [14]
Xiangju Mi [14] Renrui Zhang [2] Kun Wu [3] Zhengping Che [†3] Jian Tang [3] Pheng-Ann Heng [2]
Shanghang Zhang [✉1]

ICML/ 2026

문제상황

행동 전에 Textual CoT를 시키는 연구

로봇은 빠르게 액션을 내보내야 하는데 매 순간 자연어 CoT를 생성하면 지연이 발생하고,
자연어 표현만 으로는 로봇이 생각해야 하는 것들을 전부 표현할 수 없다.

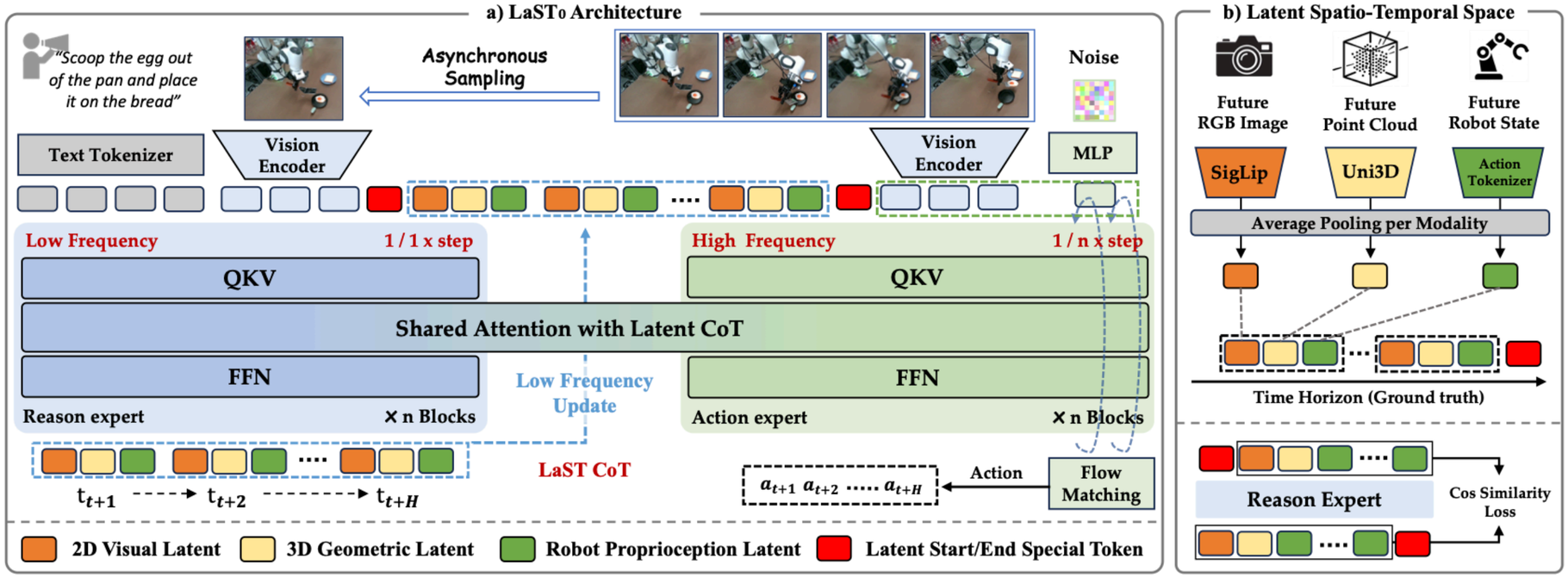


굳이 로봇의 생각을 자연어로 만들지 말자

핵심 아이디어

앞으로의 상태를 latent로 연쇄 예측

실제 trajectory의 미래 visual·3D·robot-state feature를 supervision으로 사용해서, 현재 상태로부터 미래 물리 상태의 latent chain을 학습하고, 그 latent chain을 action generation의 reasoning context로 사용

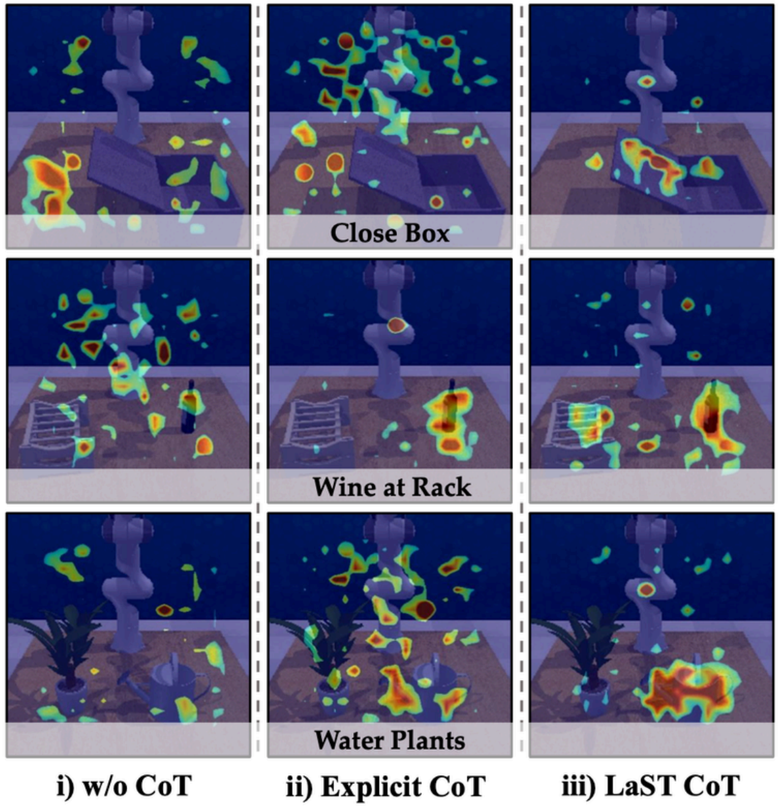


미래 physical state를 latent CoT로 구성해서 action 전에 reasoning하게 만든 모델

핵심 기여

언어 대신 latent에서 CoT reasoning

기존 Text CoT처럼 문장을 생성하지 않고, 미래 visual + 3D geometry + robot state를 latent token으로 예측하며 reasoning 해서 로봇 제어에 필요한 공간적·동적인 정보를 더 직접적으로 표현



VLM을 효율적으로 VLA로 전환

04

VITA: VISION-TO-ACTION FLOW MATCHING POLICY

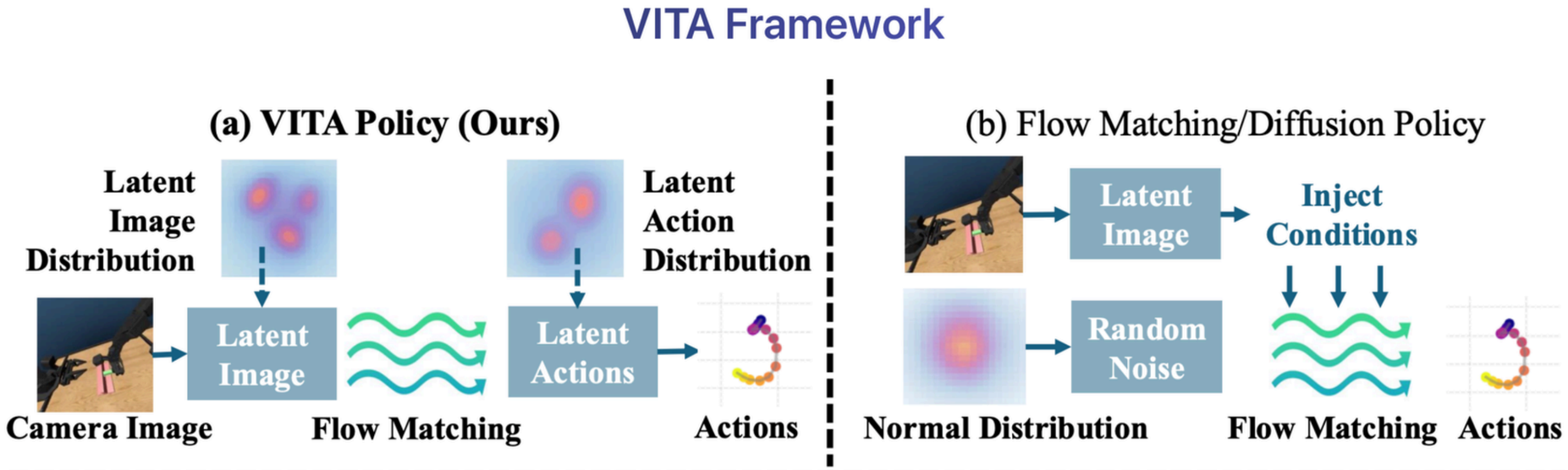
Dechen Gao[1], **Boqi Zhao**[1], **Andrew Lee**[1], **Ian Chuang**[2], **Hanchu Zhou**[1], **Hang Wang**[1], **Zhe Zhao**[1], **Junshan Zhang**[1], **Iman Soltani**[1]
[1]University of California, Davis, [2]University of California, Berkeley

ICLR/ 2026

문제상황

기존 Diffusion Policy나 Flow Matching Policy

액션을 생성할 때마다 현재 이미지 정보를 잊으면 안 되니까, 매 flow step마다 이미지 feature를 conditioning으로 다시 넣는데 매 step마다 이미지 정보를 주입하니 계산량과 메모리 사용량이 커진다.

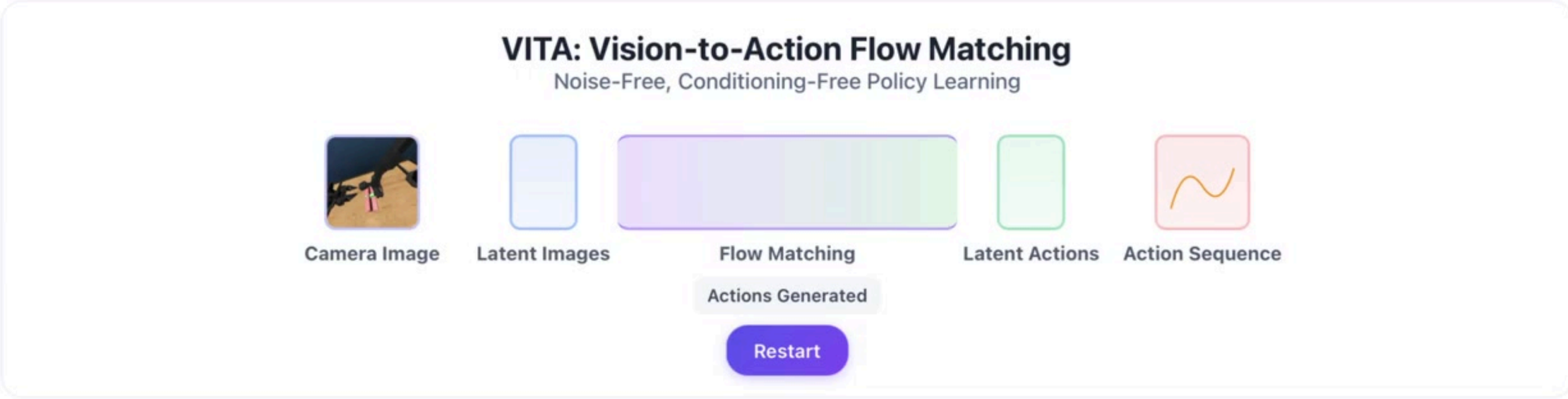


굳이 Noise에서 출발해서 Vision을 condition으로 계속 넣지 말자

핵심 아이디어

Vision 정보를 출발점 자체에 두자

Flow Matching의 시작점(source)을 Gaussian noise가 아니라 현재 이미지의 latent representation으로 사용한다.

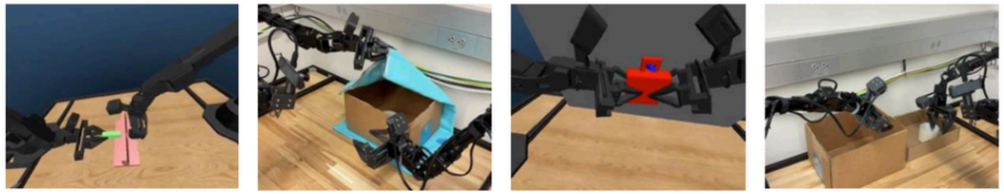


핵심 기여

Vision-to-Action Flow Matching 자체로서 높은 효율

Gaussian noise에서 출발하지 않고 visual latent를 flow의 source로 사용함으로써 vision → action을 직접 학습하여 효율성 또한 크게 높임

- **Simple:** Noise-Free & Conditioning-Free
- **Efficient:** MLP-Only & 1D Latent Spaces
- **High-Performing:** Sim & Real Experiments



World Model을 Robot Policy로 확장

05

COSMOS POLICY: FINE-TUNING VIDEO MODELS FOR VISUOMOTOR CONTROL AND PLANNING

Moo Jin Kim[1,2] **Yihuai Gao**[1,2] **Tsung-Yi Lin**[1] **Yen-Chen Lin**[1] **Yunhao Ge**[1] **Grace Lam**[1]
Percy Liang[2] **Shuran Song**[1,2] **Ming-Yu Liu**[1] **Chelsea Finn**[2] **Jinwei Gu**[1]
[1]NVIDIA [2]Stanford University

ICLR/ 2026

문제상황

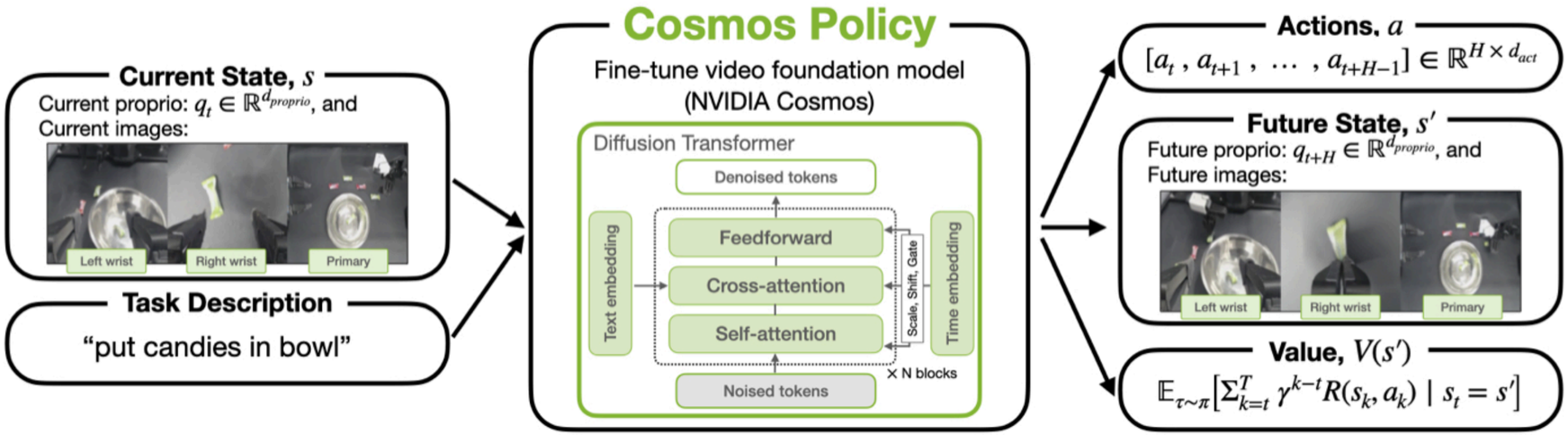
로봇 정책 학습에 적용하는 기존의 비디오 모델

일반적으로 여러 단계의 post-training을 필요로 하거나, 로봇 action을 생성하기 위한 새로운 아키텍처 구성요소를 추가해야 하는 복잡성이 존재한다.

핵심 아이디어

Video Model의 모든 것을 그대로 사용하여 하나의 Cosmos로 3가지 역할을 동시에 학습

Pretrained Video Model의 latent sequence에 Action, Robot State, Future State, Value를 모두 latent frame으로 삽입하여 하나의 모델이 Policy + World Model + Value Function 역할을 하도록 학습한다.



별도의 Action Head나 World Model Head를 추가하지 않는다

핵심 기여

① 구조 변경 없이 Video Foundation Model을 강력한 Robot Policy로 변환하고,
② Policy·World Model·Value Function을 하나의 diffusion model에 통합했으며,
③ 미래 상태를 직접 상상하고 평가하는 Model-Based Planning으로 실제 manipulation 성능을 향상시켰다.

Thank you

Question & Disscusion

RILS LAB