RILS LAB 세미나
(Reasoning in VLA Models)

RILS LAB

Robot Intelligence Learning & System

송재민

2026. 07. 23

VLA 소개

VLA는 크게 두 축으로 구별 가능

1.최종 액션을 어떻게 생성하는가

Autoregressive action generation
Diffusion-based action generation
Flow-matching-based action generation

2.액션 전에 무엇으로 reasoning하는가

Direct observation-to-action
Text reasoning
Visual reasoning
Action-space reasoning

01

Diffusion-VLA: Generalizable and Interpretable Robot Foundation Model via Self-Generated Reasoning

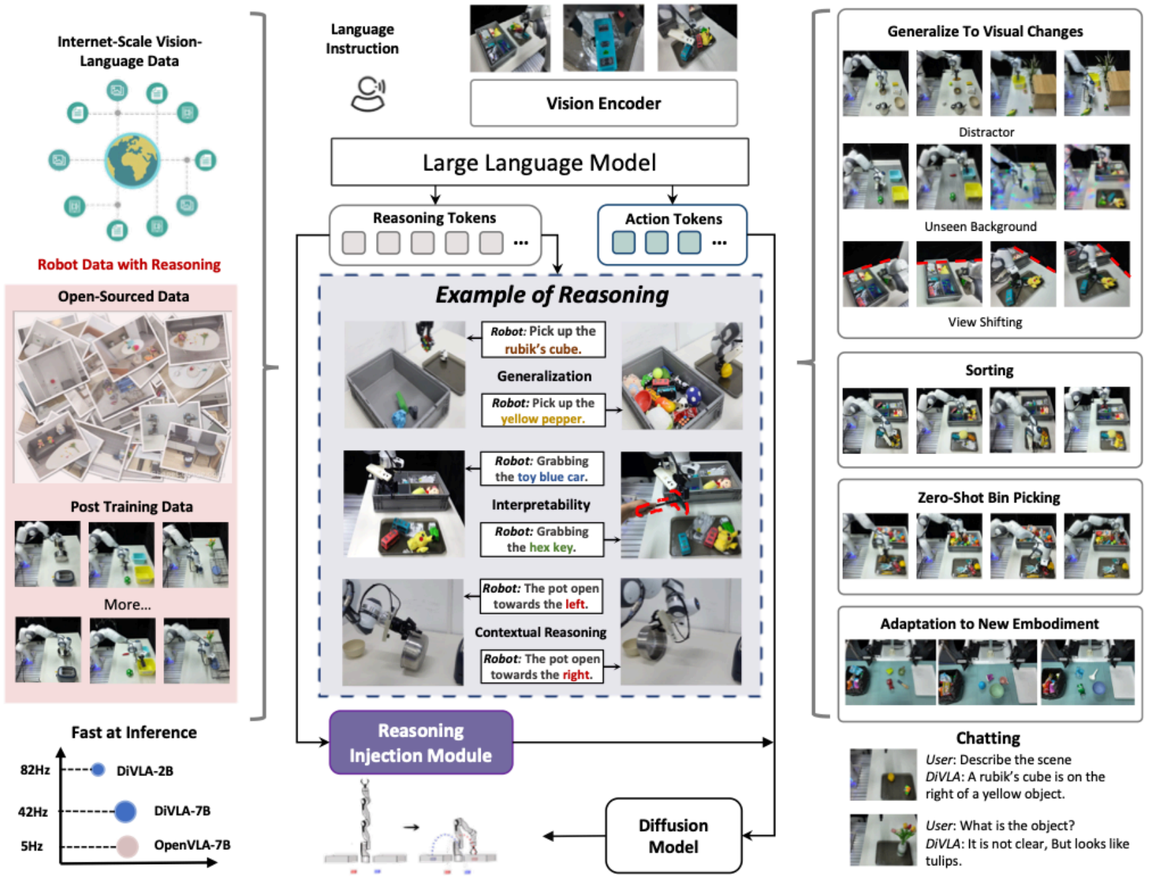
Junjie Wen, Minjie Zhu, Yichen Zhu, Zhibin Tang, Jinming Li, Zhongyi Zhou, Chengmeng Li, Xiaoyu Liu, Yaxin Peng, Chaomin Shen, Feifei Feng

ICML/ 2025

Diffusion-VLA

Diffusion-VLA는 어떤 모델인가?

로봇이 자기 주도적으로 Text reasoning을 생성하고 디퓨전 기반 정책과 직접 결합하는 모델



기존 연구와 한계

기존 end-to-end VLA모델

복잡한 환경에서 취약하고 실패 원인과 행동 의도가 불분명해진다.

모델 내부에 추론을 추가

↓

그중에서도 **텍스트 기반의 설명을** 활용하는 방향으로 발전

여러가지 문제 발생

Reasoning 학습 데이터는 LLM을 이용해 자동 구축하고, 추론 시에는 VLM이 reasoning을 직접 생성

Diffusion-VLA의 핵심 아이디어

언어 추론은 autoregressive VLM이 담당하고, 정밀한 로봇 행동 생성은 diffusion policy가 담당하게 한 뒤, VLM이 생성한 추론 정보를 실제 행동 생성 과정에 직접 주입한 모델

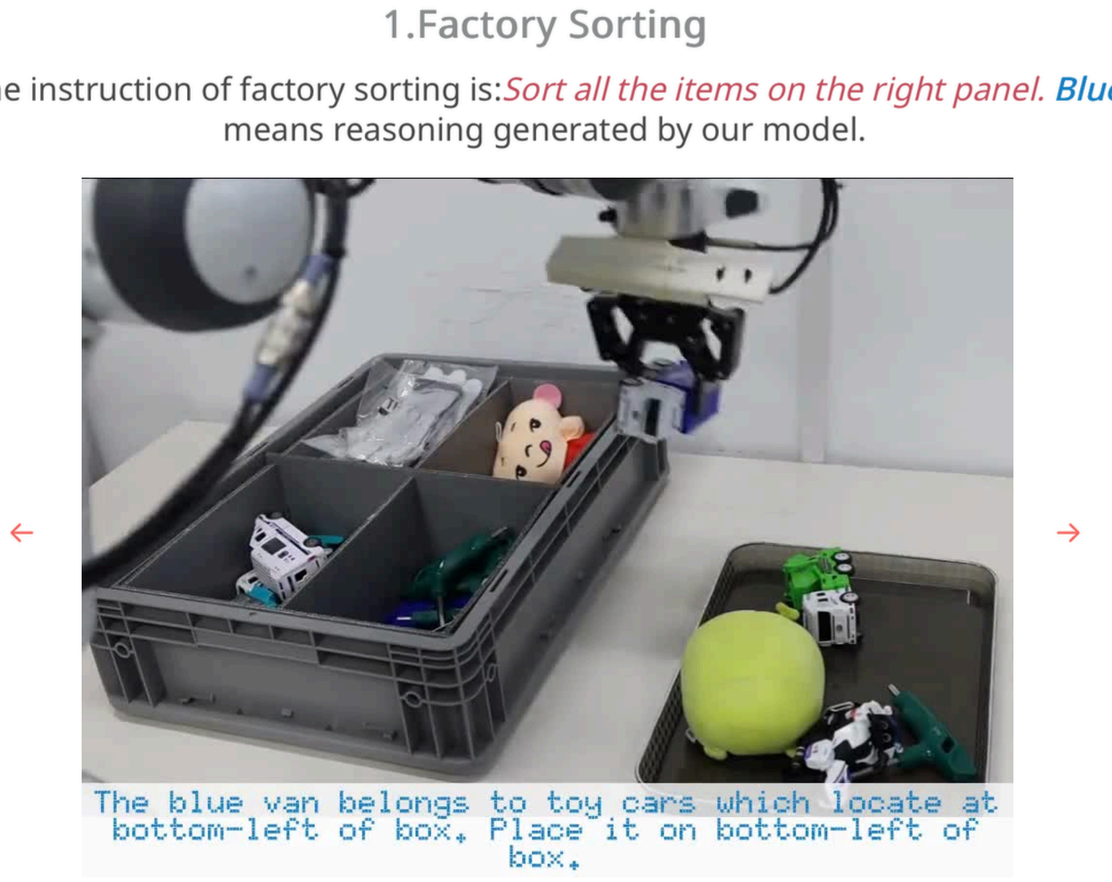
추론과 행동 생성을 서로 잘하는 모델에 분담

Autoregressive VLM

언어 명령 이해
장면과 물체 인식
과제 분해
텍스트 기반 reasoning 생성

Diffusion policy

연속적인 로봇 action sequence 생성
정밀하고 부드러운 trajectory 생성
높은 주파수의 실시간 제어



핵심 기여 및 한계

1.Autoregressive reasoning과 diffusion action을 하나의 VLA로 통합

2.명령, 실행이아닌 명령 - 사고 - 실행의 구조를 통해 해석 가능성과 일반화 성능 확보

3.텍스트 형태 reasoning은 실제 로봇의 위치나 물체의 자세와 같은 시각적 정보를 모두 표현하는 데에는 한계

02

CoT-VLA: Visual Chain-of-Thought Reasoning for Vision-Language-Action Models

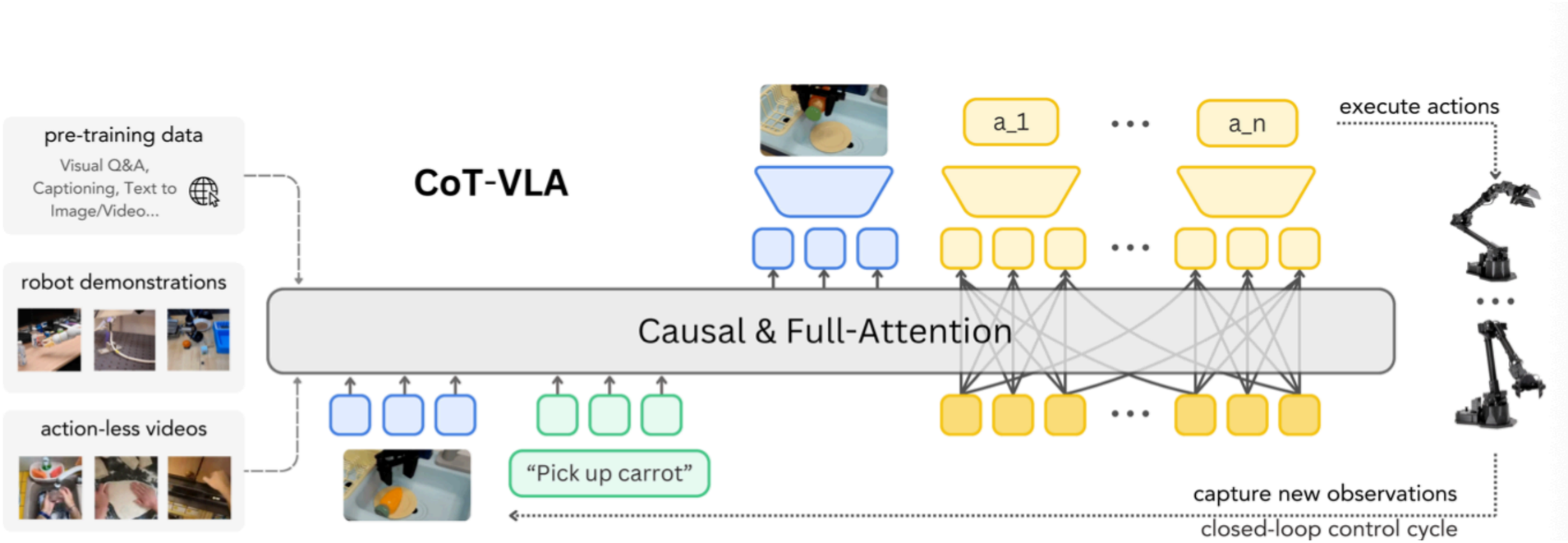
Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, Ankur Handa, Ming-Yu Liu, Donglai Xiang, Gordon Wetzstein, Tsung-Yi Lin

CVPR / 2025

CoT-VLA

CoT-VLA는 어떤 모델인가?

Text기반 reasoning 대신 이미지 기반 reasoning을 거쳐 액션을 생성하는 모델



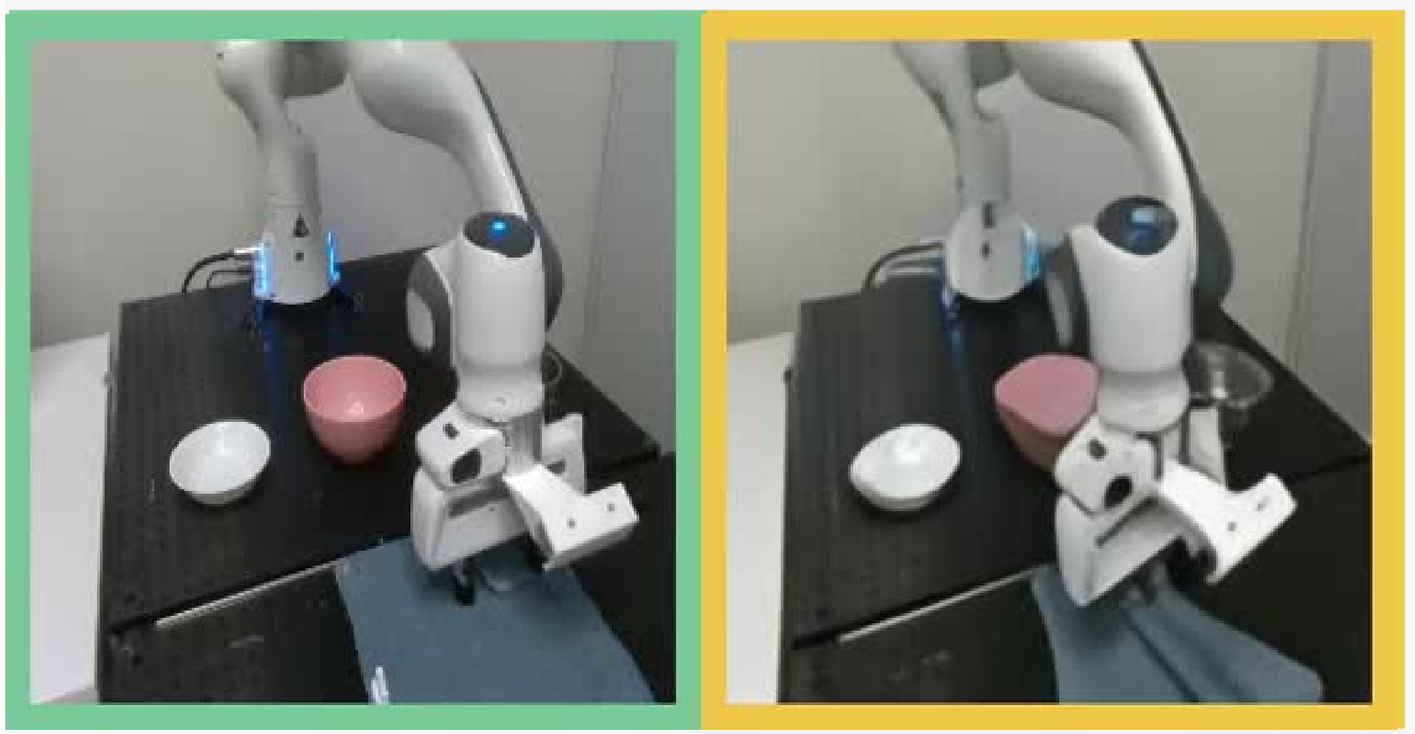
자연어 중간 추론이 아닌 미래 이미지를 reasoning으로 활용

CoT-VLA의 핵심 아이디어

행동하기 전에 미래 장면을 먼저 생성하고 그 이미지를 바탕으로 짧은 액션 시퀀스를 생성한다.

Subgoal image 활용

텍스트 기반 reasoning은
이미지 입력을 텍스트로 변환시켜
단계적으로 할 일을지정하는데
이미지를 글로 바꿀시
이미지의 정보 손실이 발생



핵심 기여

1.Visual Chain-of-Thought를 VLA에 도입

2.명령, 실행이아닌 명령 - 사고 - 실행의 구조를 통해 해석 가능성과 일반화 성능 확보

03

ACoT-VLA: Action Chain-of-Thought for Vision-Language-Action Models

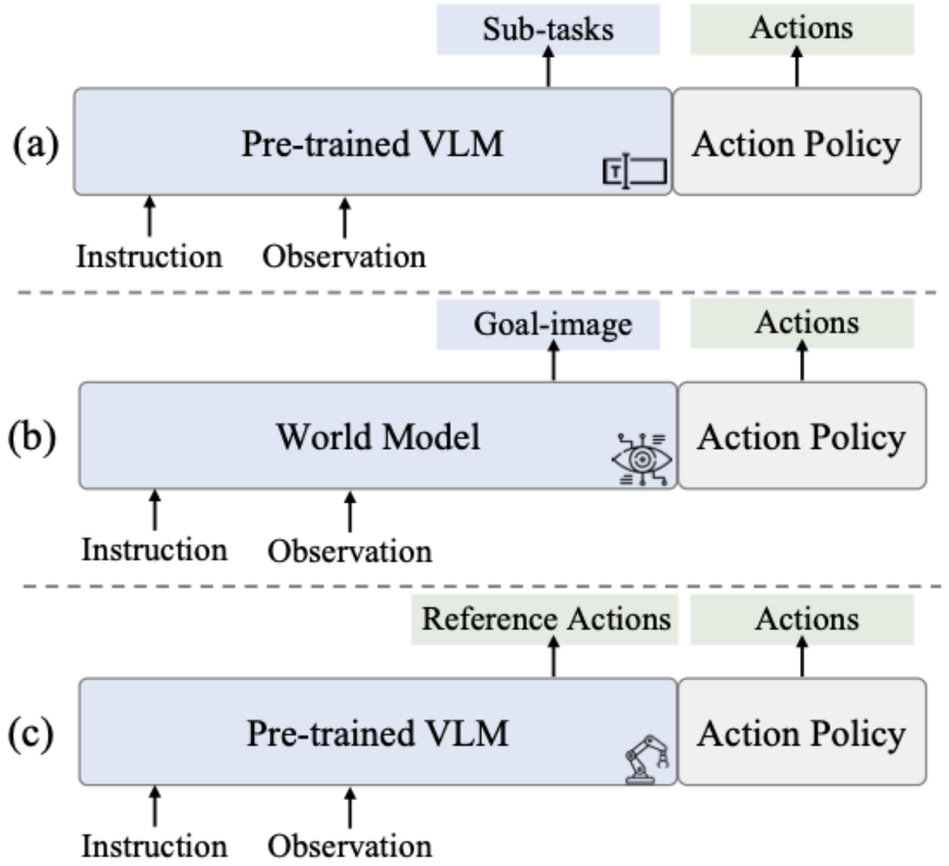
Linqing Zhong, Yi Liu, Yifei Wei, Ziyu Xiong, Maoqing Yao, Si Liu, Guanghui Ren

CVPR/ 2026

ACoT-VLA

ACoT-VLA는 어떤 모델인가?

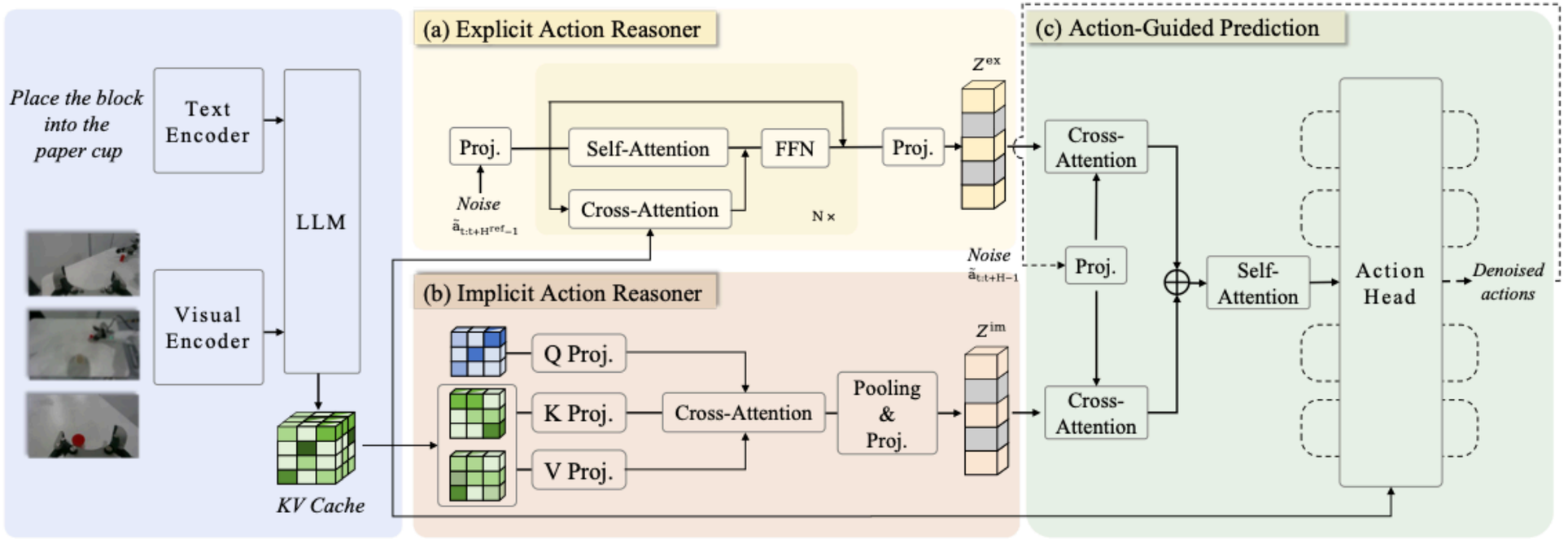
중간reasoning을 이제는 text도 image도 아닌 대략적인 행동으로 만드는 모델



ACoT-VLA의 핵심 아이디어

행동하기 전에 대략적인 거친 행동을 먼저 생성하고 그 액션을 바탕으로 액션 시퀀스를 생성한다.

현재 이미지와 명령을 보고, 대략적인 참조 행동을 먼저 만드는 것



VLM이 장면을 보며 이해한 내용 중에서, 실제 움직임과 관련된 정보만 뽑는 것

핵심 기여

1.CoT의 공간을 언어·시각에서 행동으로 이동

2.행동 정보를 명시적·암묵적으로 분리

로봇에서 무엇을 reasoning이라고 부를 것인가를 action-space intermediate plan으로 재정의

Thank you

Question & Disscusion

RILS LAB