

Advanced Vision-Language-Action Models (Part 1): Enhancing Visual Understanding and Reasoning

DMQA Open Seminar (2026. 05. 08)

허종국

발표자 소개

허종국 (Jong Kook, Heo)

- Data Mining & Quality Analytics Lab
- Ph.D. Student (2021.03~)
- 지도 교수 : 김성범 교수님

Research Interest

- Reinforcement Learning with Human Feedback
- Vision-Language-Action Models

Contact

- hjks01406@korea.ac.kr



Introduction

Importance of Physical AI

❖ Direction of Future AI

- 사람은 실제 물체와의 상호 작용을 통해 더 적극적으로 주의하고 행동
- 로봇과의 물리적 상호 작용은 화면 기반 것보다 훨씬 건강하고 효과적

Direction of Future AI?

가상 세계에 기반한 화면 속 애플리케이션

→ 실제 세계를 인지하고 인간과 물리적으로 상호 작용하는 실체 (embodied agent)

Approach:

- "Puppeteer" the baby by demonstrations from a "baby-sized & baby-shaped robot."
- Robot moves to maximize visual contact with the baby, the baby is most likely to imitate same movement.
- Challenge & reward (e.g., kicking height, bell chime) adapt to facilitate learning.

The robot's movement drives the baby's attention, which drives the baby's action. (Can't be done from a screen.)



Vision-Language Model (VLM) →

Vision-Language-Action Model (VLA)

NO, PHYSICAL EMBODIMENT IS CRITICAL

> 50 years of neuroscience evidence supporting embodied interactions leading to improved learning, training, and health outcomes

the simplest of embodiments is processed differently by your brain, than a screen. Tons and tons of data. Literally, your brain is engaged differently. Did

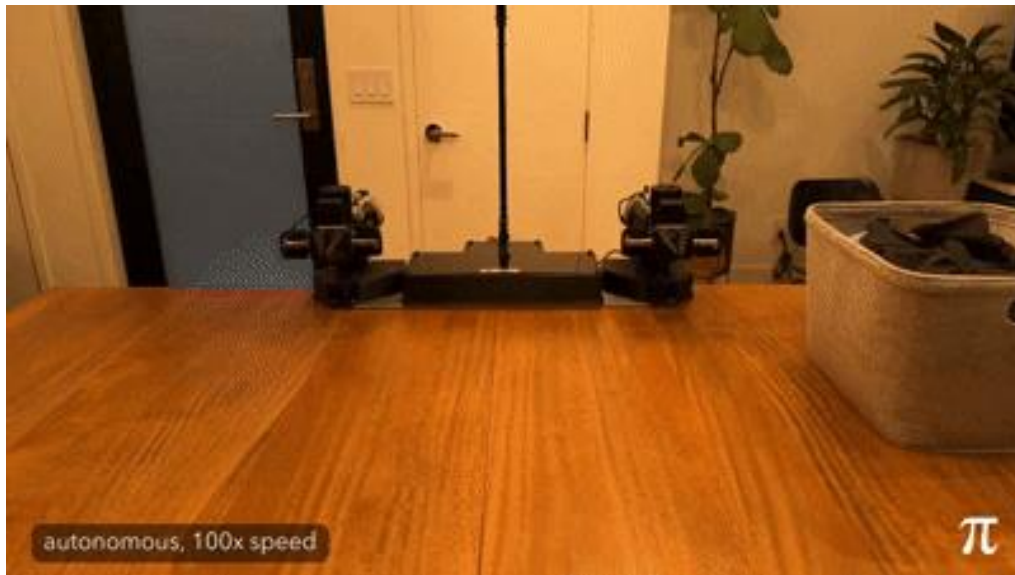
the work we've done. Here's an example of something that you absolutely cannot do with the screen ever. Do it. Okay. We were interested in working with really young

Preliminaries

Vision-Language-Action Models

❖ What is vision-language-action models (VLA)??

- 시각적 관측 정보(vision)와 자연어 명령(language)을 이해하고, 실제 환경에서 수행할 행동 (action)을 예측하는 모델
- 간단히 물체를 집는 것을 넘어 **복잡한 협업 및 고난이도의 정밀한 태스크를 수행**

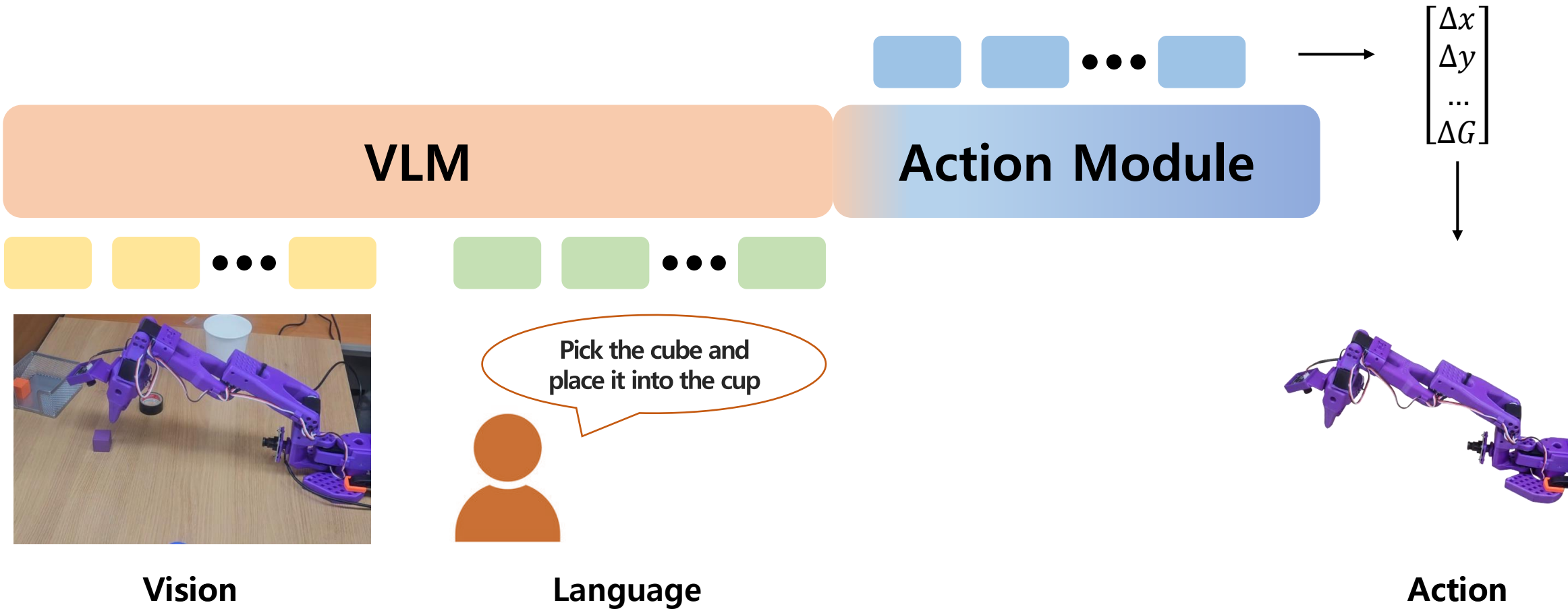


Preliminaries

Vision-Language-Action Models

❖ What is vision-language-action models (VLA)??

- 사전 학습된 VLM/MLLM 을 기반으로 시각/언어 표현을 추출하고, action head를 통해 로봇 제어 명령을 예측

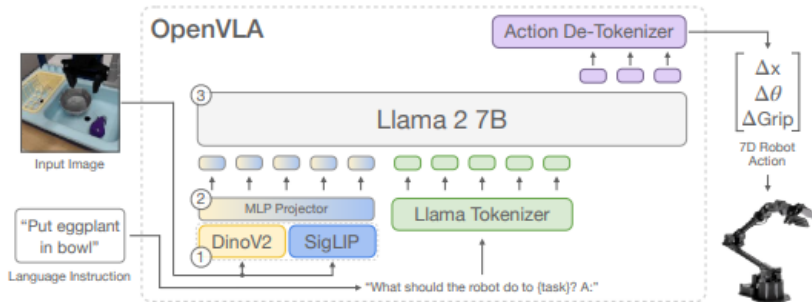


Preliminaries

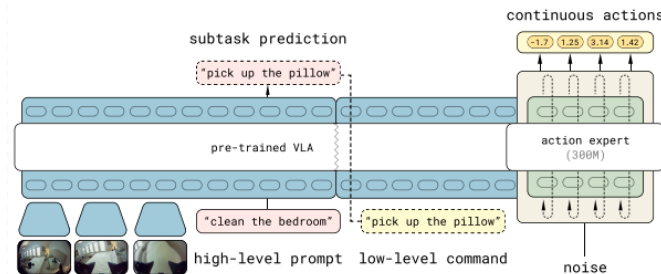
Vision-Language-Action Models

❖ Initial Challenges of Vision-Language-Action Models

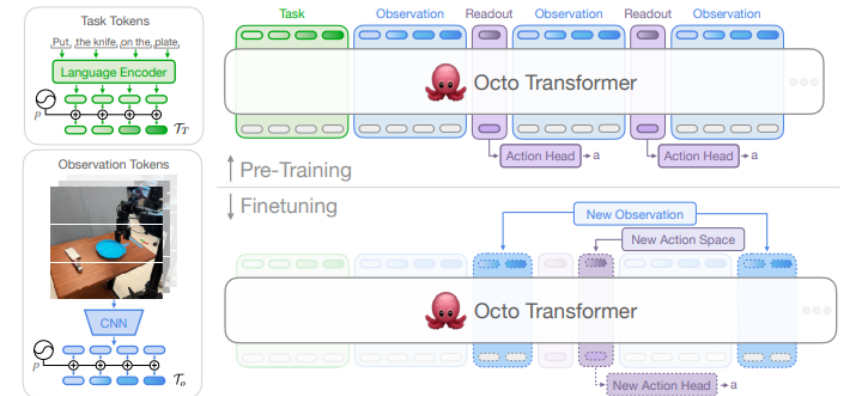
- 모델 구조를 어떻게 디자인할까?
- 액션은 어떤 형태로 예측할까?
- 서로 다른 관절을 가진 로봇에 범용적으로 적용할 수 있을까?



OpenVLA



$\pi_{0.5}$



Octo

Preliminaries

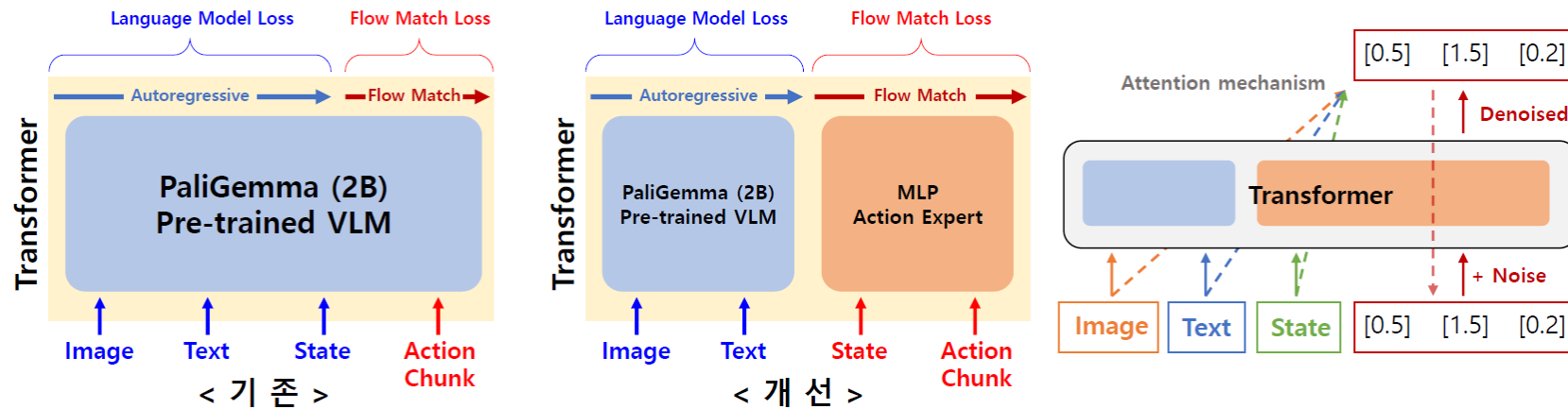
Reference

❖ Details

- Introduction to Robot Learning
 - ✓ 초기 로봇 파운데이션 모델 소개 : RT1, RT2, π_0
 - ✓ Action chunking, action discretization, flow matching 기반 action sequence 생성

❖ π_0 : Mixture-of-Expert

- π_0 는 Mixture-of-Expert에서 제안한 한 구조 내 여러 가중치 집합(expert)을 활용하여 각 가중치가 특화된 역할을 학습함
- VLM은 사전학습에 쓰인 modality인 이미지와 텍스트만을 처리하며 새롭게 추가한 action expert는 로봇의 관절 상태와 행동 값을 처리함
 - VLM에서 발생할 수 있는 distribution shift 문제를 완화 → 모델의 성능 향상
 - Action expert는 VLM보다 크기가 작음 → 모델의 연산 효율성 증가



Preliminaries

Research Trends

❖ Emerging Research Areas in VLA

- 장기 태스크 수행 및 forgetting 현상 해결을 위한 memory module 고도화
- 복잡한 상황에 대한 시각적 이해 및 구체적인 계획 수립을 위한 CoT 기반 reasoning
- 비싼 robot demonstration 의존성을 줄이기 위한 world-model 혹은 data-efficient training 기법

Memory
for Long Horizon Tasks

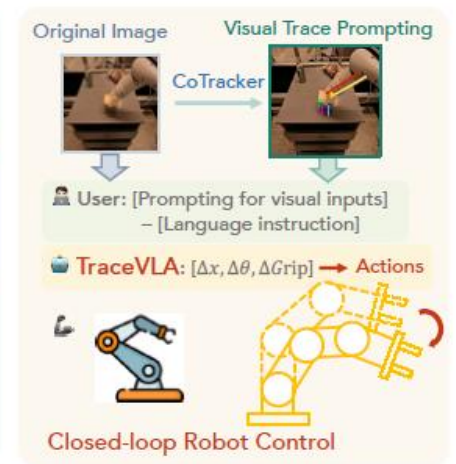
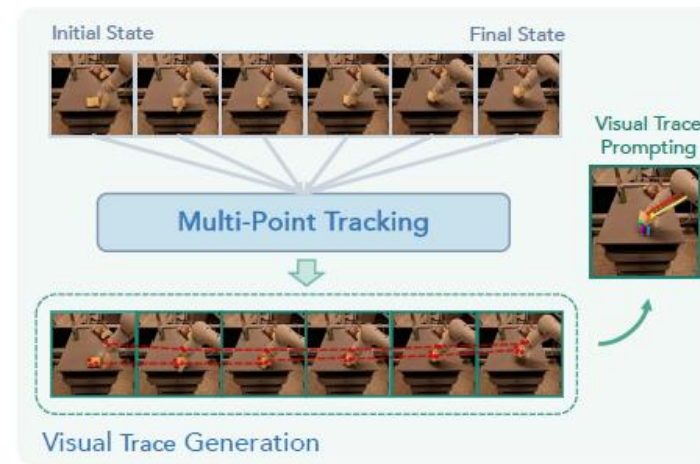
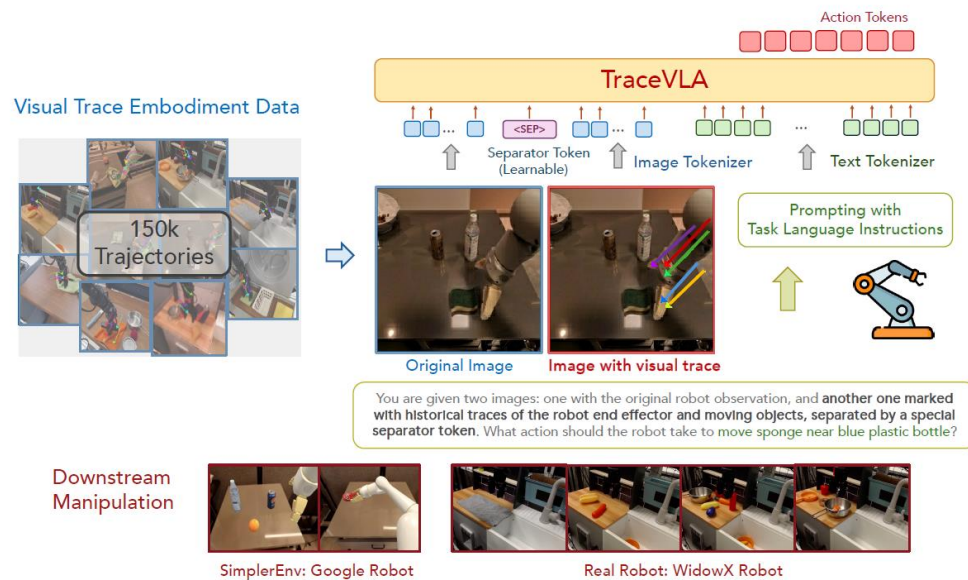
Reasoning
for Complex-Decision
Making

Data Efficiency
for Scalable VLA Training

Advanced Methods

TraceVLA

- ❖ TraceVLA: Visual Trace Prompting Enhances Spatial-Temporal Awareness For Generalist Robotic Policies (Zheng et al., ICLR 2025)
 - 현재 관측에 주로 의존하는 기존 VLA는 과거 움직임에 대한 **시공간 인지 능력(spatial-temporal awareness)**이 부족
 - **로봇의 end-effector 및 움직이는 물체에 대한 이동 궤적을 표현하여**, 과거 상태-행동 흐름을 시각적 prompt로 제공

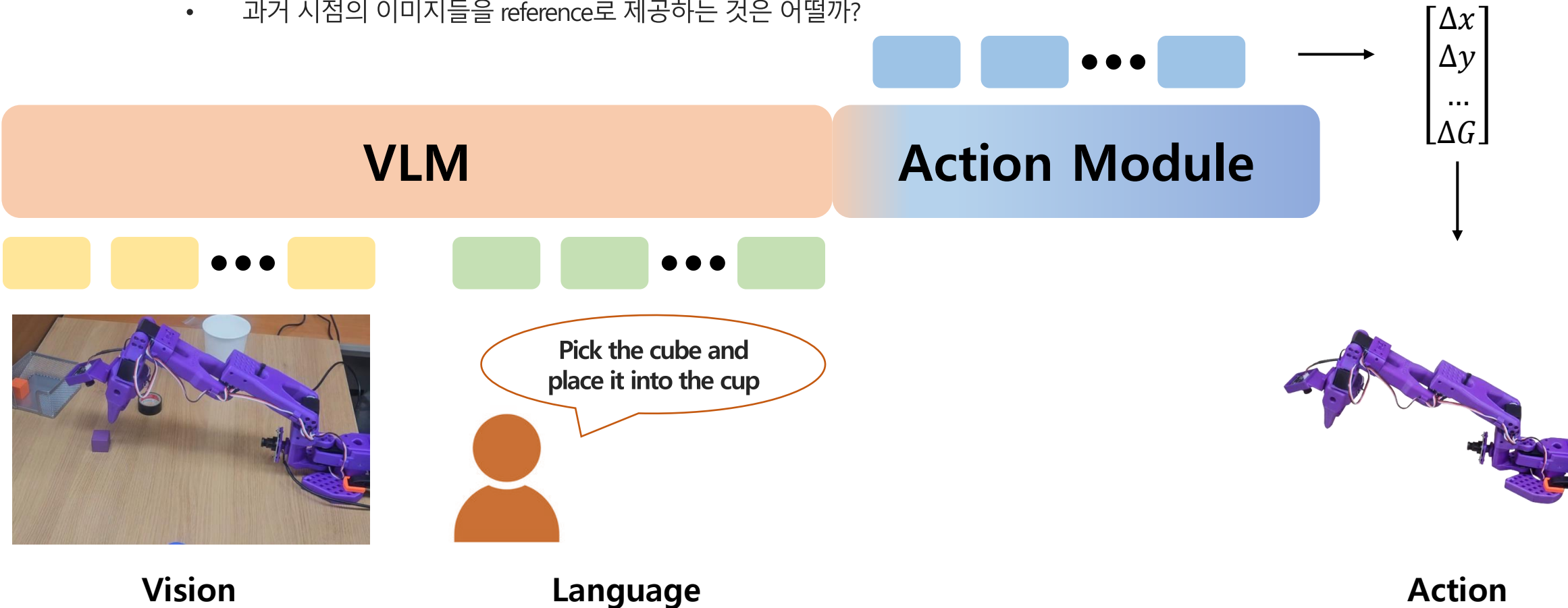


Advanced Methods

TraceVLA

❖ Lack of Spatial-Temporal Awareness in VLA

- 현재 시점에 대한 시각적 정보만 제공할 경우, **로봇이 시공간적 맥락(spatial-temporal context)**을 포착하기 어려움
- 과거 시점의 이미지들을 reference로 제공하는 것은 어떨까?

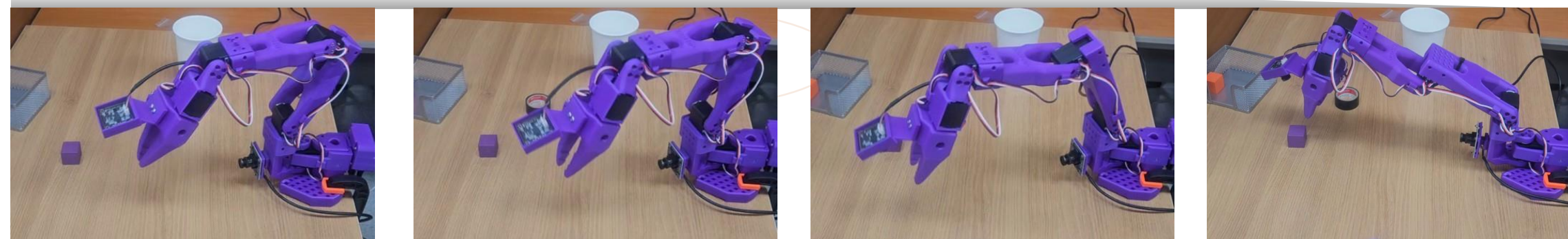
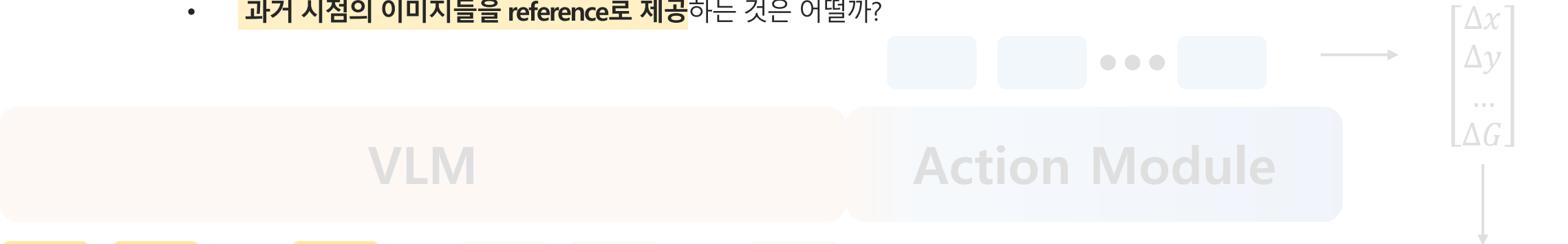


Advanced Methods

TraceVLA

❖ Lack of Spatial-Temporal Awareness in VLA

- 현재 시점에 대한 시각적 정보만 제공할 경우, 로봇이 시공간적 맥락(spatial-temporal context)을 포착하기 어려움
- 과거 시점의 이미지들을 reference로 제공하는 것은 어떨까?



Vision

Previous observations($o_{t-N:t-1}$)

Current observation(o_t)

Advanced Methods

TraceVLA

❖ Lack of Spatial-Temporal Awareness in VLA

- 현재 시점에 대한 시각적 정보만 제공할 경우, 로봇이 시공간적 맥락(spatial-temporal context)을 포착하기 어려움
- **중요한 물체들의 궤적만 이미지에 표시하는 것은 어떨까?**

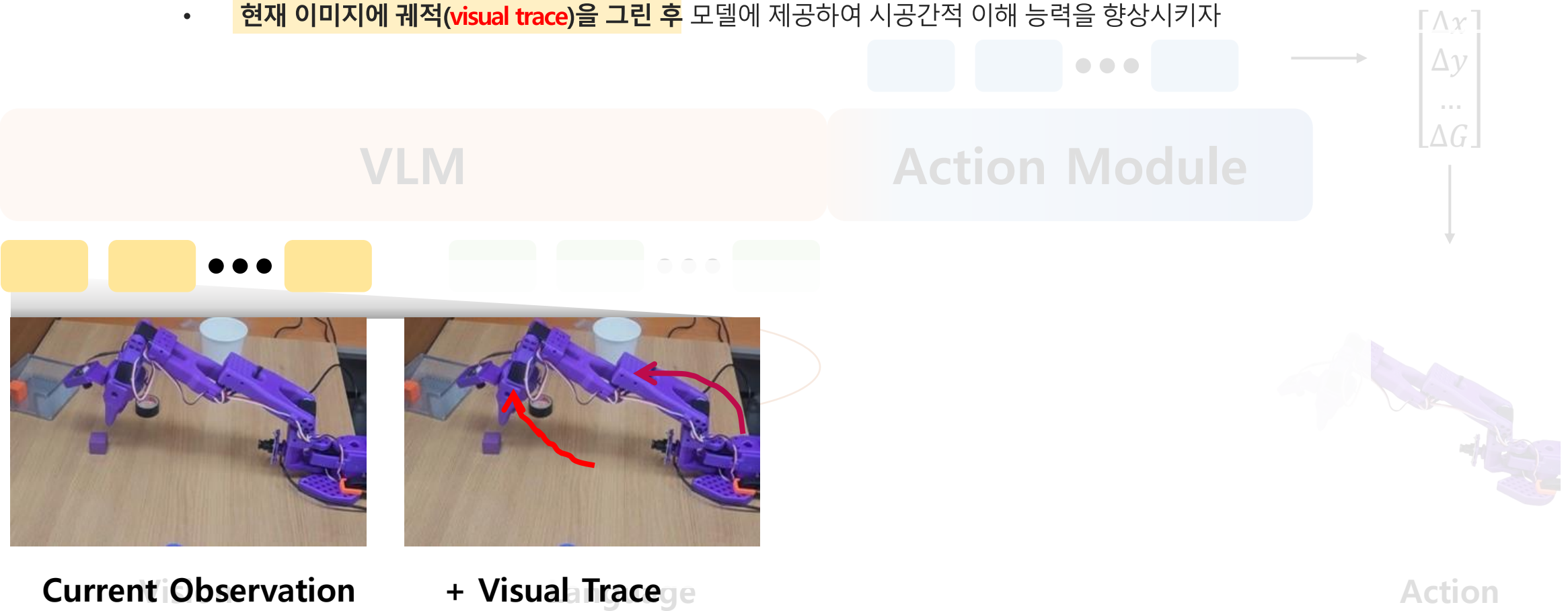


Advanced Methods

TraceVLA

❖ Visual Trace Prompting

- 현재 시점에 대한 시각적 정보만 제공할 경우, 로봇이 시공간적 맥락(spatial-temporal context)을 포착하기 어려움
- 현재 이미지에 궤적(visual trace)을 그린 후 모델에 제공하여 시공간적 이해 능력을 향상시키자



Current Observation

+ Visual Trace

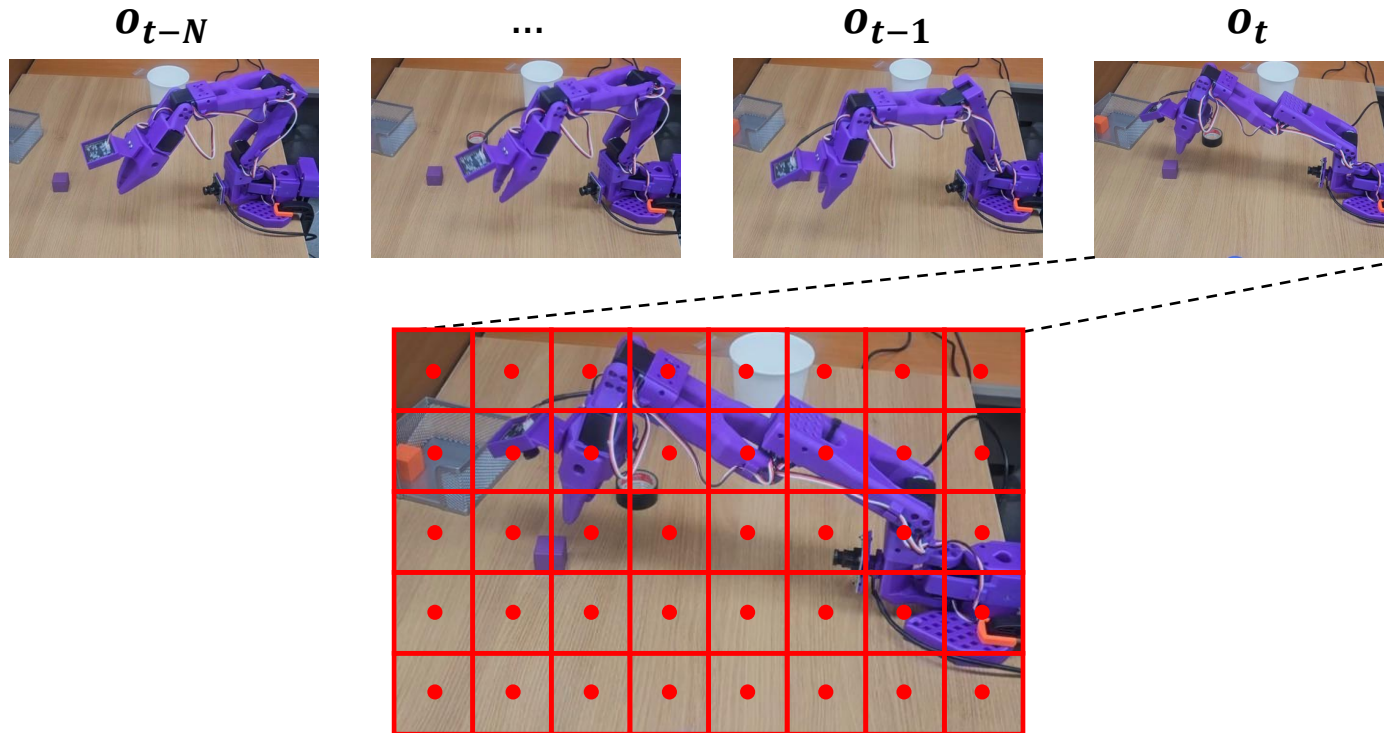
Action

Advanced Methods

TraceVLA

❖ 어떻게 Visual Trace를 추출할까?

- Cotracker (Karaev et al., ECCV 2024)를 통해 현재 시점 이미지에서 $K \times K$ grid의 중심점의 과거 경로를 추출
- 위치 변화가 큰 경로 (active point)중 M개의 점을 선별하여 일정 시간 동안 tracking
- 상황 변화 및 누적 오차를 방지하기 위해 주기적으로 새롭게 tracking point 선별

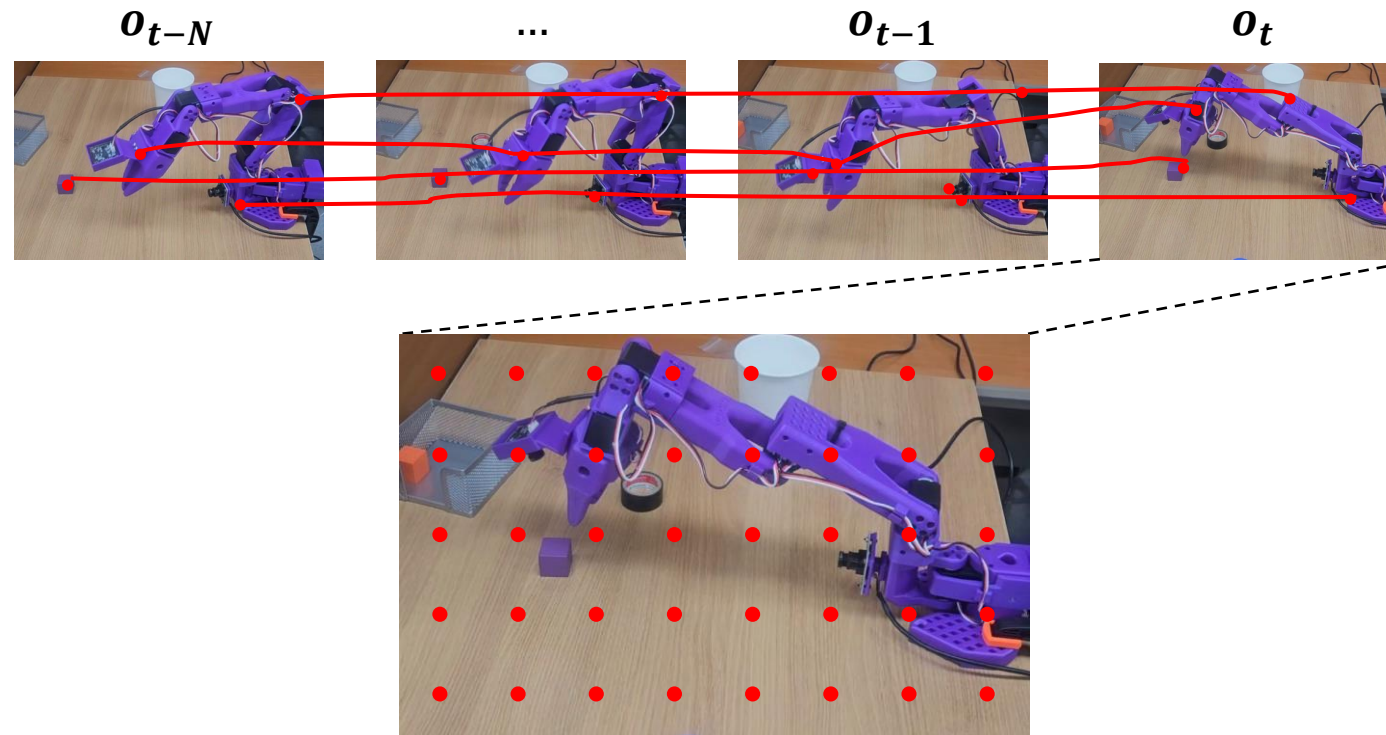


Advanced Methods

TraceVLA

❖ 어떻게 Visual Trace를 추출할까?

- Cotracker (Karaev et al., ECCV 2024)를 통해 현재 시점 이미지에서 $K \times K$ grid의 중심점의 과거 경로를 추출
- 위치 변화가 큰 경로 (active point)중 M개의 점을 선별하여 일정 시간 동안 tracking
- 상황 변화 및 누적 오차를 방지하기 위해 주기적으로 새롭게 tracking point 선별

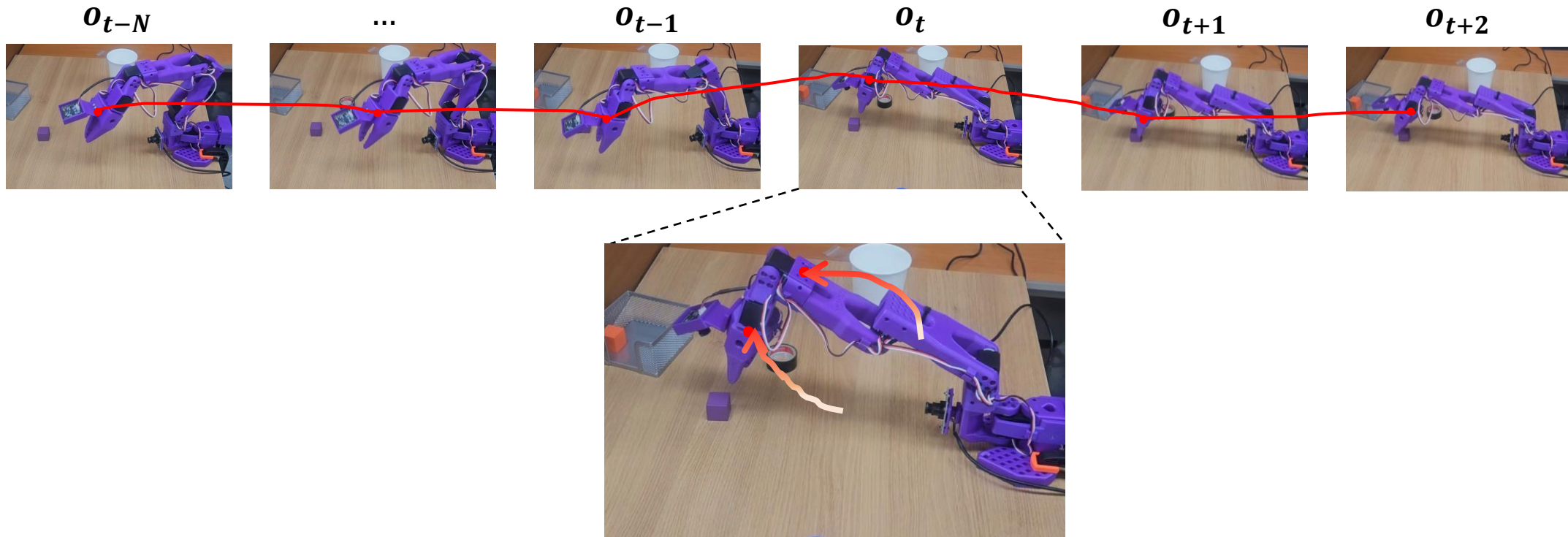


Advanced Methods

TraceVLA

❖ 어떻게 Visual Trace를 추출할까?

- Cotracker (Karaev et al., ECCV 2024)를 통해 현재 시점 이미지에서 $K \times K$ grid의 중심점의 과거 경로를 추출
- 위치 변화가 큰 경로 (active point)중 M개의 점을 선별하여 일정 시간 동안 tracking
- 상황 변화 및 누적 오차를 방지하기 위해 주기적으로 새롭게 tracking point 선별

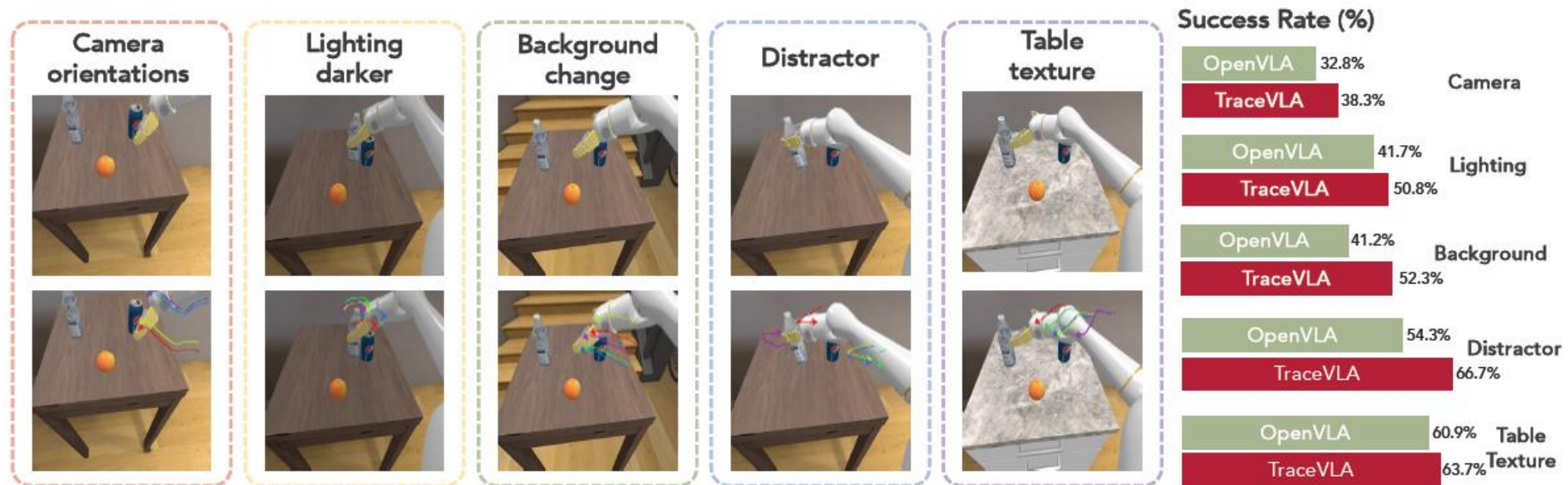


Advanced Methods

TraceVLA

❖ Experiments

- OpenVLA 7B/4B 모델에 visual trace prompt를 활용해 fine-tuning (150k robot demonstrations from BridgeData-v2, Google Robot, WidowX250)
- 시뮬레이션 환경 (SimplerEnv)와 실제 로봇 (WidowX-250)에 대해 실험
- 조명, 카메라 각도, 배경 변화에도 강건한 성능 입증

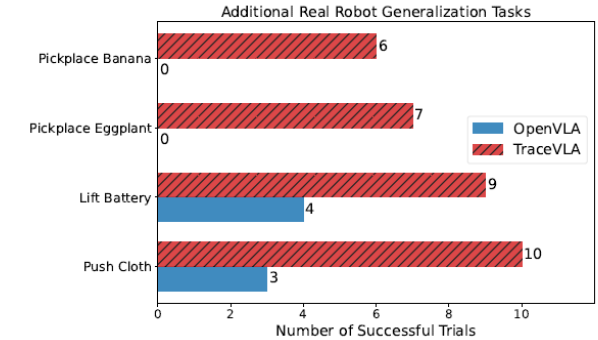
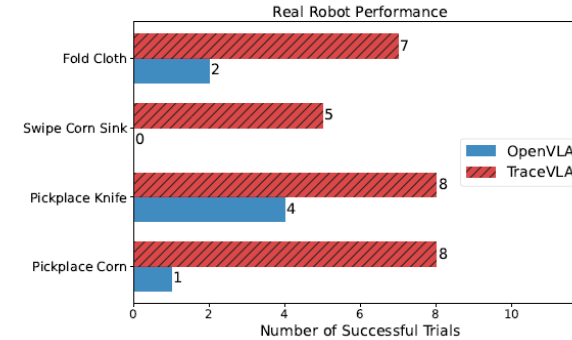


Advanced Methods

TraceVLA

❖ Experiments

- OpenVLA 7B/4B 모델에 visual trace prompt를 활용해 fine-tuning (150k robot demonstrations from BridgeData-v2, Google Robot, WidowX250)
- 시뮬레이션 환경 (SimplerEnv)와 실제 로봇 (WidowX-250)에 대해 실험
- 학습 태스크 뿐만 아니라 보지 못했던 태스크에서도 높은 일반화 성능 입증

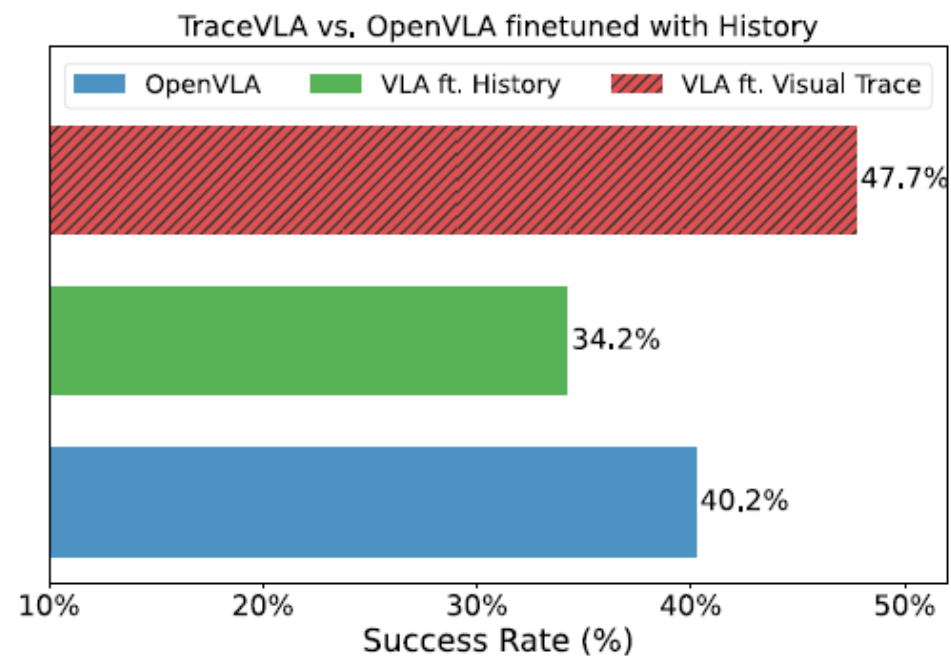
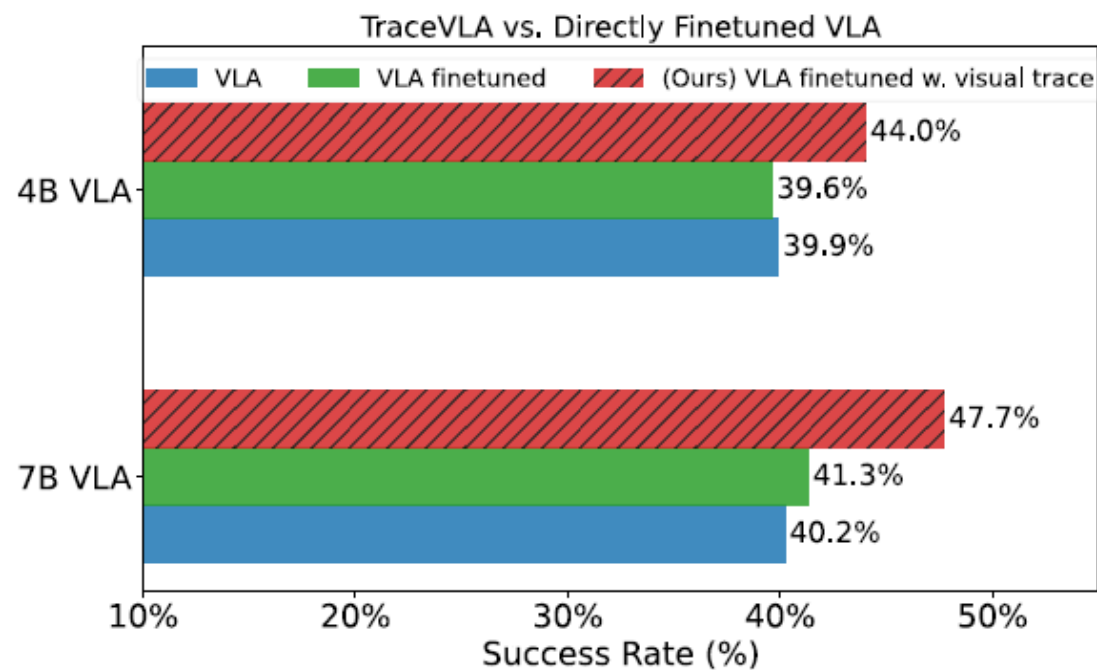


Advanced Methods

TraceVLA

❖ Ablation

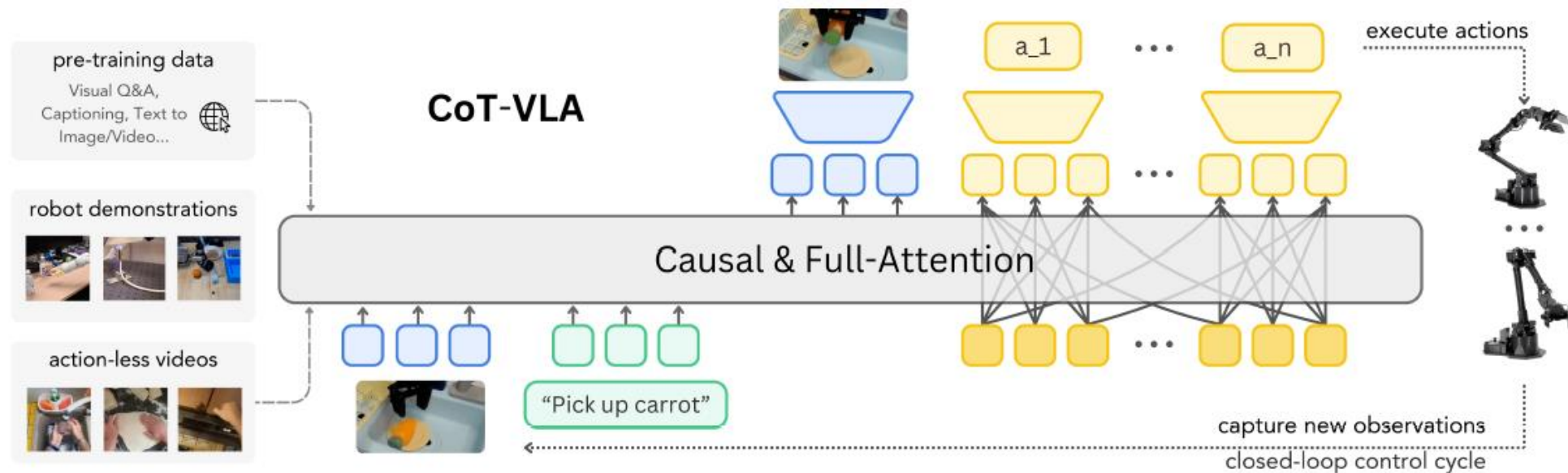
- 성능 향상의 원인이 visual trace 때문인가 아니면 추가 학습 때문인가?
- 단순히 과거 시점의 프레임을 넣는 것은 어떠한가?



Advanced Methods

CoT-VLA

- ❖ CoT-VLA: Visual Chain-of-Thought Reasoning for Vision-Language-Action Models (Zhao et al., CVPR 2025)
 - 행동을 예측하기 전에 가까운 미래에 달성해야 할 **sub-goal** 이미지를 우선 생성하여 visual reasoning을 수행
 - “현재 상태 → 목표 상태 상상 → 행동 생성” 구조를 통해 reasoning before acting 구현
 - Action-less video data를 활용해 visual reasoning 강화

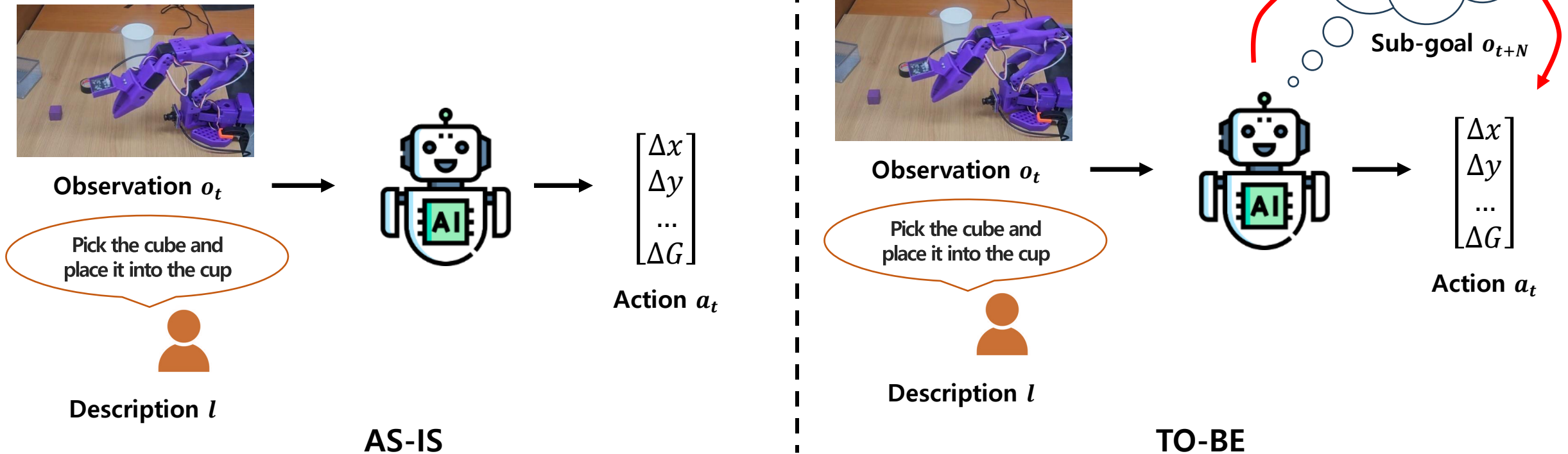


Advanced Methods

CoT-VLA

❖ Imagine Before Acting

- 기존 VLA는 현재 관측과 언어 지시를 action으로 직접 매핑 → 명시적인 중간 추론단계가 부족
- “가까운 상태에 우선 무엇을 달성해야 하는가?”를 시각적으로 계획 (intermediate visual CoT)



Advanced Methods

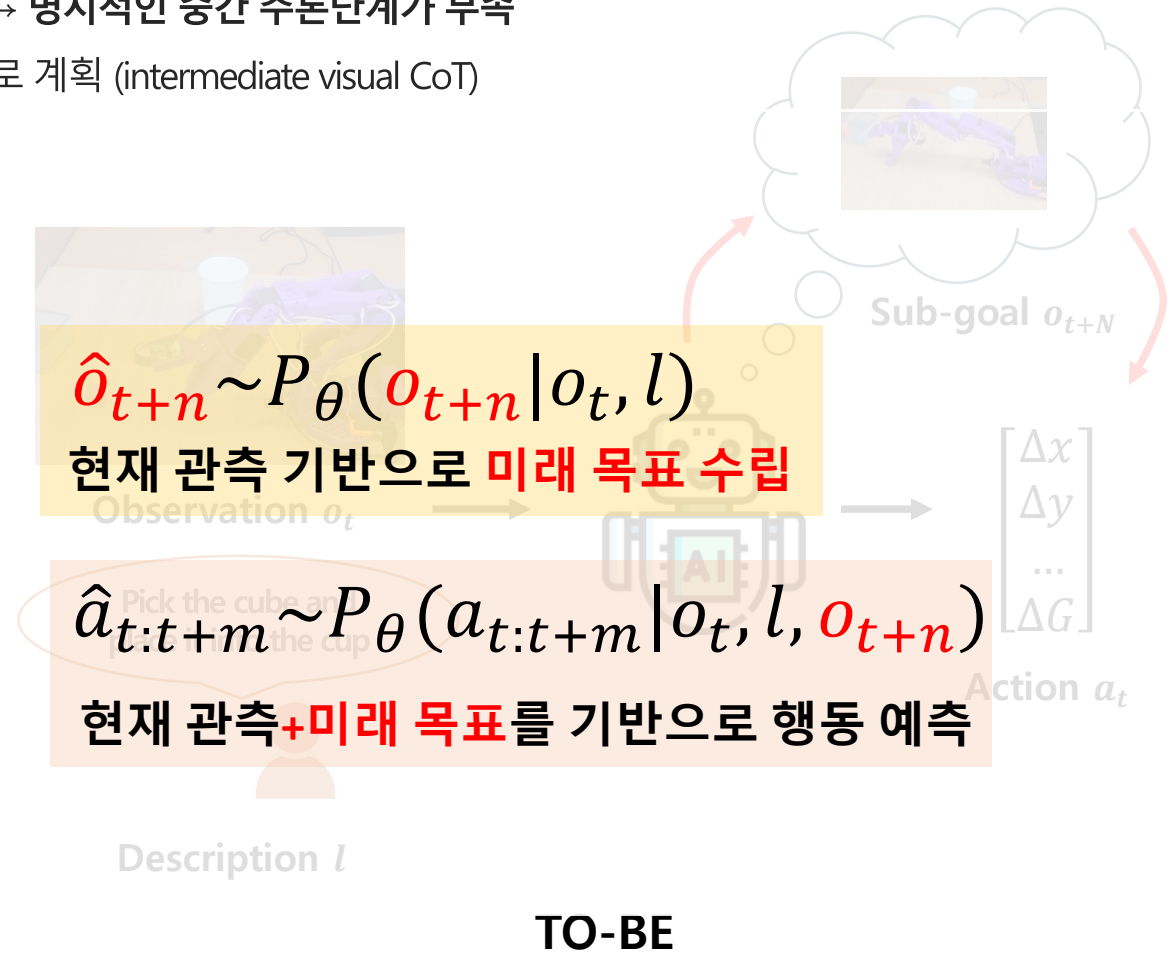
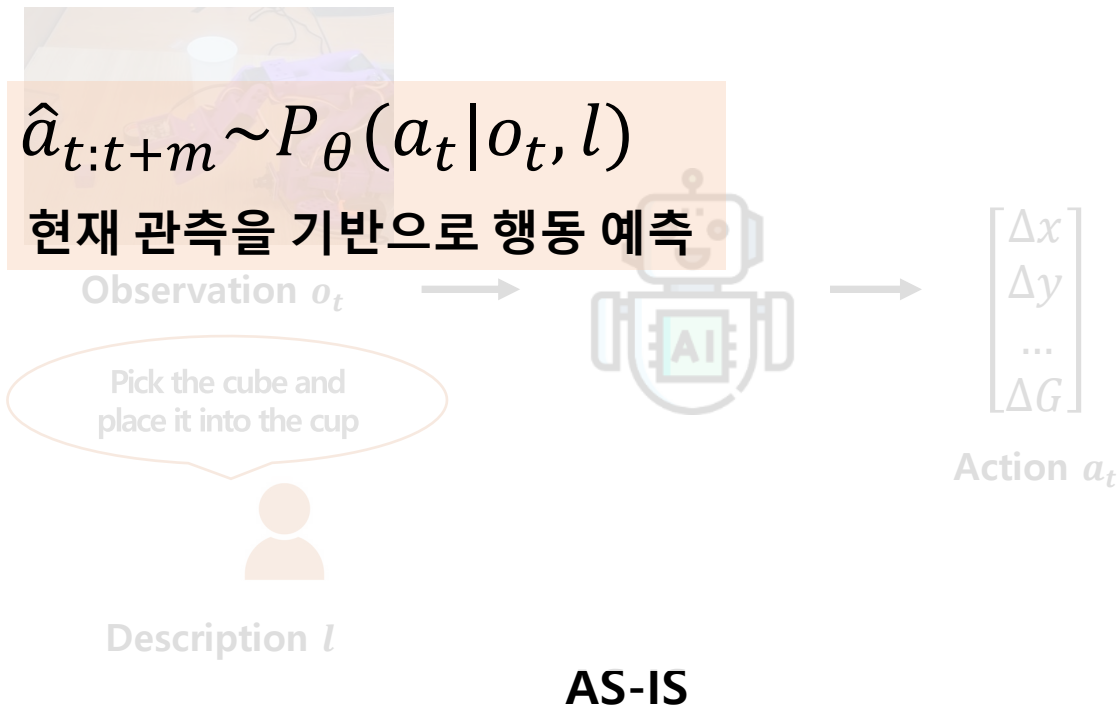
CoT-VLA

❖ Imagine Before Acting

- 기존 VLA는 현재 관측과 언어 지시를 action으로 직접 매핑 → 명시적인 중간 추론단계가 부족
- “가까운 상태에 우선 무엇을 달성해야 하는가?”를 시각적으로 계획 (intermediate visual CoT)

이미지 + 행동 정보가 모두 있는 로봇 데이터(D_r)로 학습

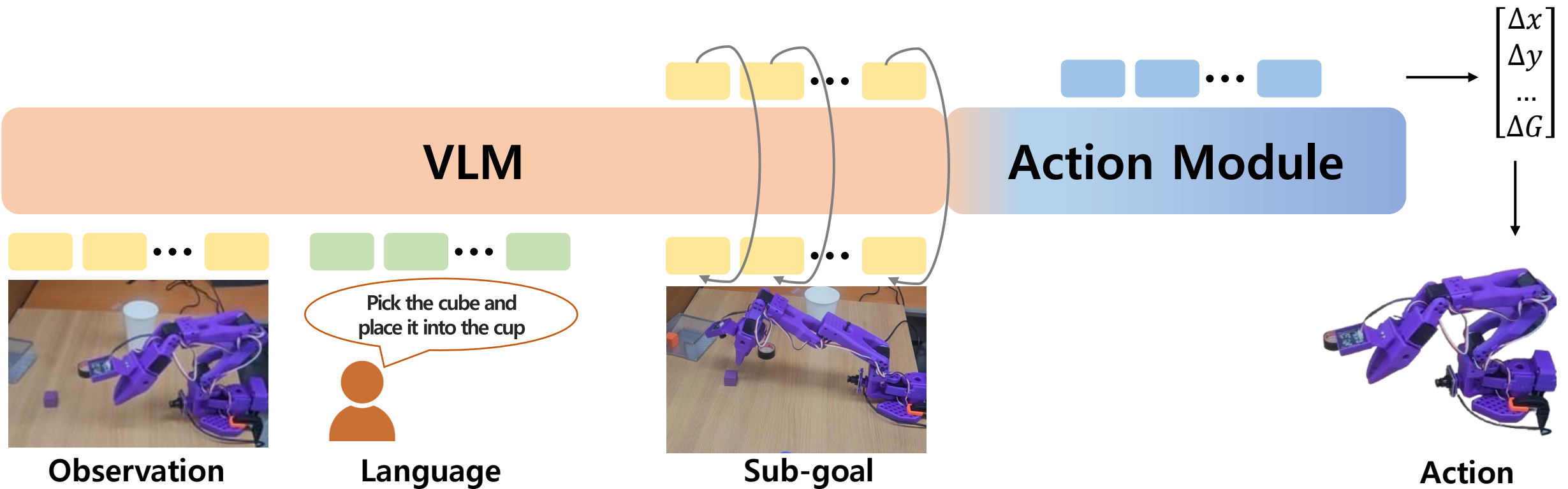
이미지만 있는 일반 비디오 데이터(D_v)로 학습



Advanced Methods

CoT-VLA

- ❖ Training with generative VLM (VILA-U)
 - **Visual Token Prediction** : sub-goal 이미지를 패치(토큰) 단위로 생성 (auto-regressive)
 - **Action Token Prediction** : sub-goal 이미지를 참고하여 행동 시퀀스 생성 (bi-directional)



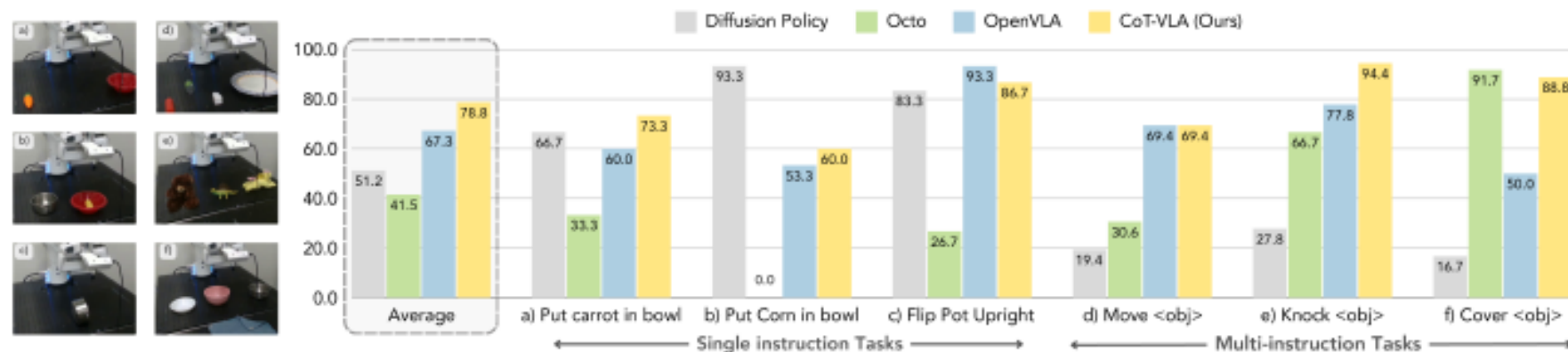
Advanced Methods

CoT-VLA

❖ Experiments

- 사전 학습 데이터 : 로봇 시연 데이터 (Open X Embodiment) + 일반 영상 데이터 (EPIC-KITCHENS + Something-Something-V2)
- 시뮬레이션 환경 (LIBERO) 및 실제 로봇 (WidowX, Franka-Tabletop) 환경에 대해 전이 학습 후 평가

	Average (↑)	Spatial (↑)	Object (↑)	Goal (↑)	Long (↑)
Diffusion Policy	72.4 ± 0.7%	78.3 ± 1.1%	92.5 ± 0.7%	68.3 ± 1.2%	50.5 ± 1.3%
Octo fine-tuned	75.1 ± 0.6%	78.9 ± 1.0%	85.7 ± 0.9%	84.6 ± 0.9%	51.1 ± 1.3%
OpenVLA fine-tuned	76.5 ± 0.6%	84.7 ± 0.9%	88.4 ± 0.8%	79.2 ± 1.0%	53.7 ± 1.3%
CoT-VLA-7B (ours)	81.13 ± 0.6 %	87.5 ± 1.4%	91.6 ± 0.5%	87.6 ± 0.6%	69.0 ± 0.8%



Advanced Methods

CoT-VLA

❖ Better Visual Reasoning Helps

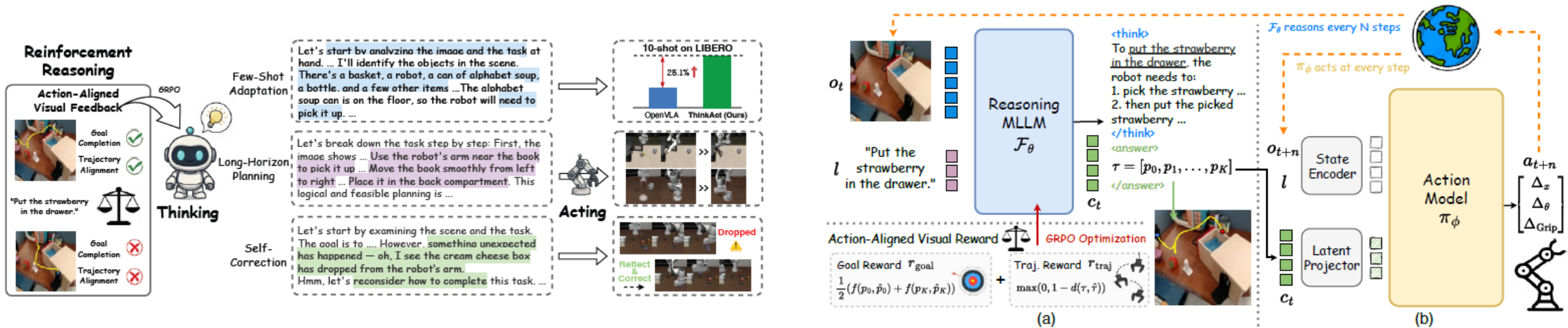
- Franka-Tabletop 로봇 환경에서 학습 때 보지 못한 장기 태스크 (long-horizon task)를 평가
- 생성된 sub-goal에 기반한 행동 추론과 실제 sub-goal에 기반한 행동 추론 비교
- **Sub-goal image generation 품질이 향상되면 행동 예측 성능도 향상될 수 있음**을 시사

	Sub-task 1	Sub-task 2
Generated Goal Images	20%	0%
Ground-truth Goal Images	60%	40%

Advanced Methods

ThinkAct

- ❖ ThinkAct: Vision-Language-Action Reasoning via Reinforced Visual Latent Reasoning (Huang et al., NeurIPS 2025)
 - Reasoning VLM과 action model을 분리한 think-before-act 구조
 - **Action-aligned visual reward**와 **강화학습**을 통해 embodied reasoning 강화
 - Visual latent planning을 통해 **long-horizon planning** 수행



Advanced Methods

ThinkAct

❖ Reason Before Acting

- 기존 VLA는 현재 관측과 언어 지시를 action으로 직접 매핑 → 명시적인 중간 추론단계가 부족
- Visual reasoning (CoT-VLA)는 sub-goal image 생성 필요 → 속도가 느리고 확장성이 떨어짐
- Q&A 방식의 추론은 실제 로봇 행동과의 직접 연결되지 않음 → 로봇 행동 유도하는 planning signal로 활용하기 어려움



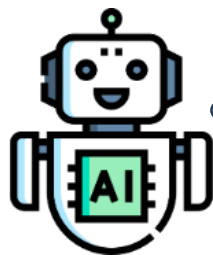
Observation

Pick the cube and place it into the cup

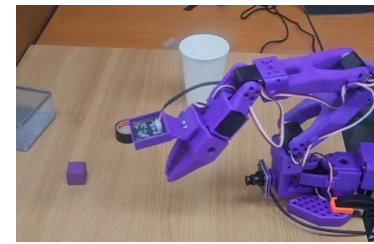


Description

Visual reasoning



Imagination as sub-goal (Image)



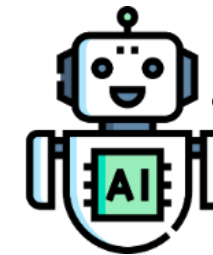
Observation

Pick the cube and place it into the cup



Description

Q&A reasoning



Q) What's the next step?
A) Move to the cube

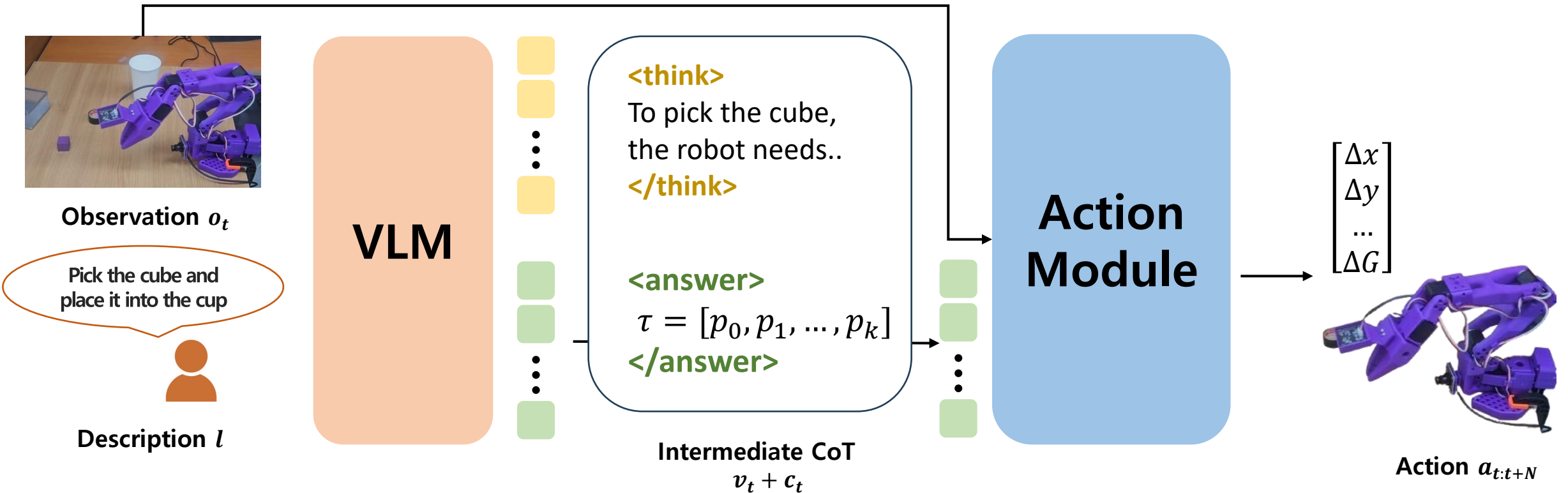
Plan as Answer (Text)

Advanced Methods

ThinkAct

❖ Reason Before Acting (Visual Latent Planning)

- 현재 이미지와 자연어 명령어를 입력받아 **추론 임베딩(v_t)**과 **경로 계획 임베딩(c_t)**를 함께 출력
- 경로 계획 임베딩을 함께 action module에 입력하여 행동 시퀀스를 생성

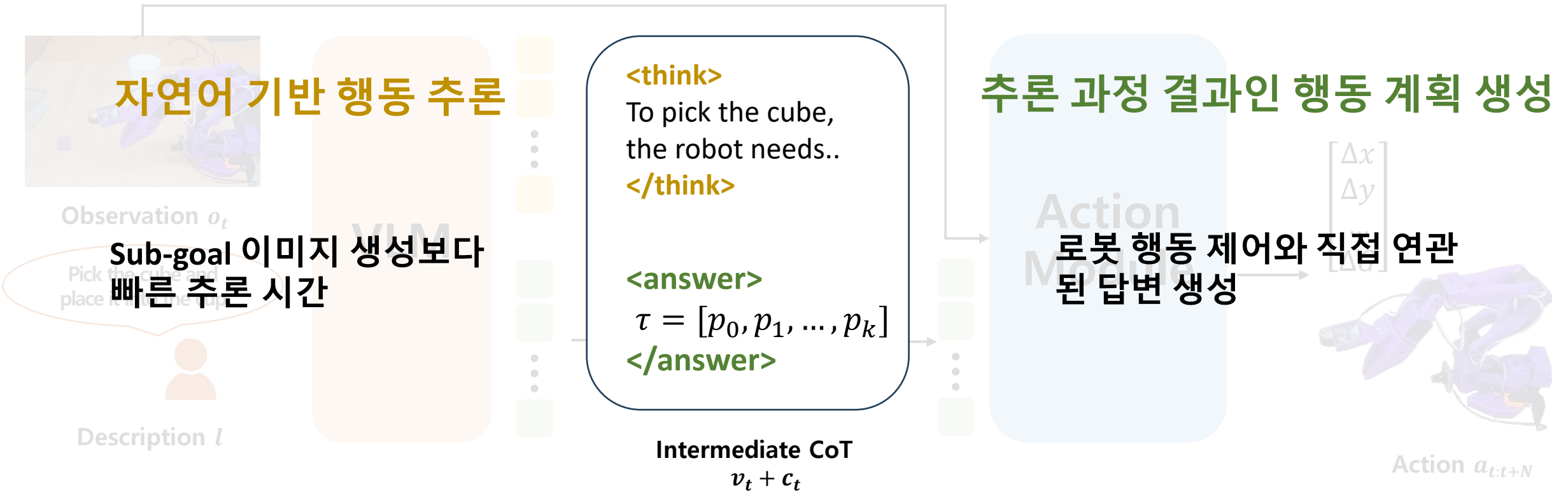


Advanced Methods

ThinkAct

❖ Reason Before Acting (Visual Latent Planning)

- 현재 이미지와 자연어 명령어를 입력받아 **추론 임베딩(v_t)**과 **경로 계획 임베딩(c_t)**를 함께 출력
- 경로 계획 임베딩을 함께 action module에 입력하여 행동 시퀀스를 생성

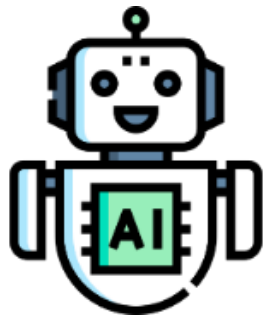


Advanced Methods

ThinkAct

❖ Reinforcement Fine-tuning for Eliciting Latent Planning

- 강화학습 알고리즘인 GRPO를 활용하여 VLM이 산출한 경로 계획 (c_t)을 평가
- 경로 계획이 실제 데이터에서 수행된 행동 시퀀스와 유사할 수록 (DTW distance) 높은 점수 부과



<think>

...

$\tau = [p_0, p_1, \dots, p_k]$

</answer>

<think>

...

$\tau = [p_0, p_1, \dots, p_k]$

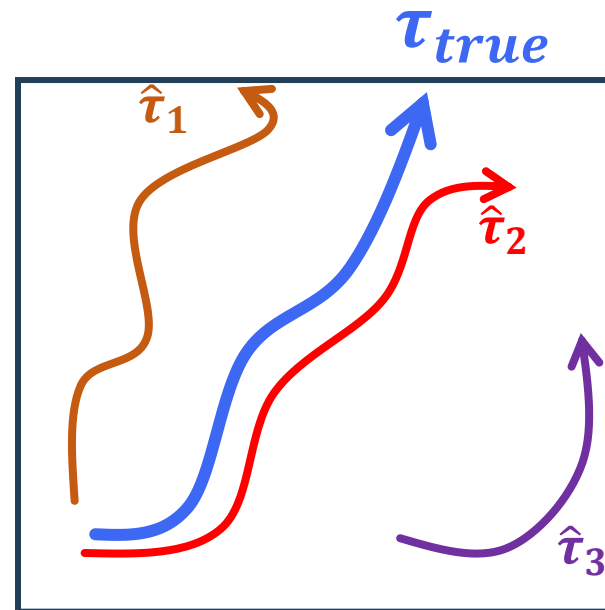
</answer>

<think>

...

$\tau = [p_0, p_1, \dots, p_k]$

</answer>



0.5



3.6



-1.8



VLM

Intermediate CoT

Trajectory Rendering

Reward

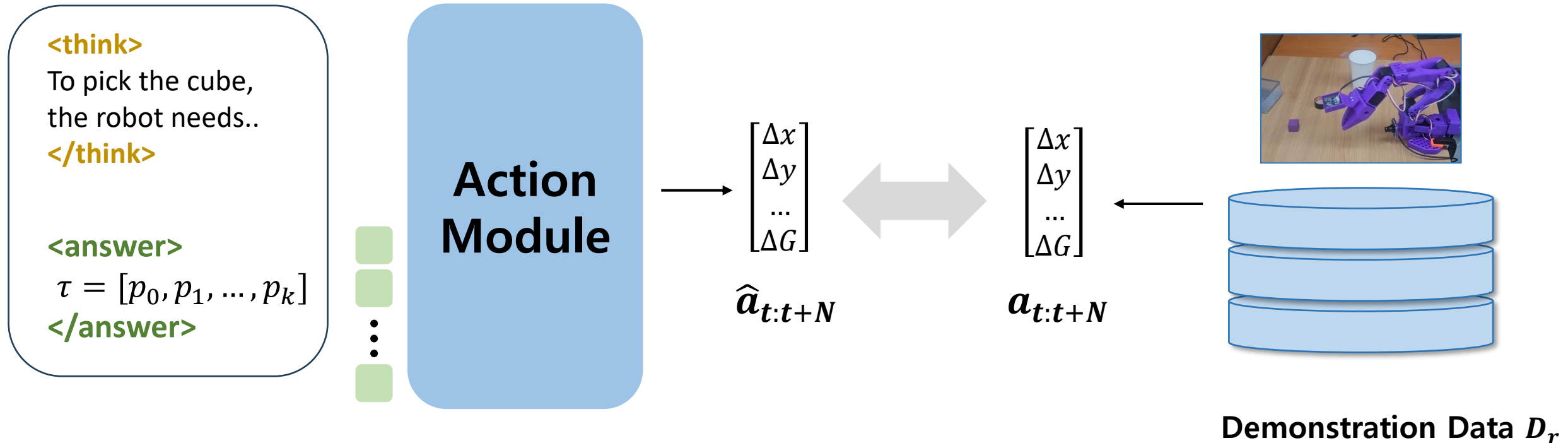
Likelihood

Advanced Methods

ThinkAct

❖ Reasoning-Enhanced Action Adaptation

- 경로 계획 (c_t)에 기반하여 실제 수행해야 할 행동 예측
- 실제 학습 데이터에서 수행된 행동과의 차이를 줄이는 방향으로 학습 (diffusion-based)



Advanced Methods

ThinkAct

- ❖ 추론 과정에 대한 보상은 주지 않아도 괜찮을까?
 - 결과에 대한 보상을 통한 강화학습이 정말 LLM의 추론 능력을 향상시킬까?
 - 결과에 대한 보상만 있어도 점점 더 논리적인 CoT를 학습하게 된다는 것을 증명

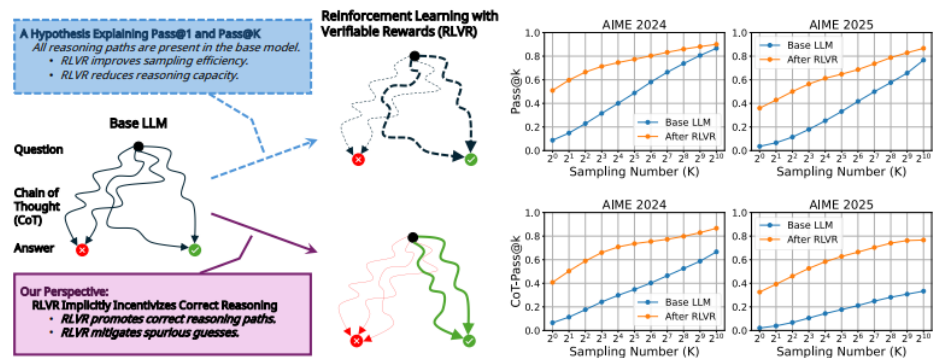


Figure 1: An illustration of our perspective: **RLVR implicitly incentivizes correct reasoning in base LLMs**. We visualize how different explanation frameworks lead to varying reasoning paths being activated, with our perspective shown in the lower left and a recent popular hypothesis explaining Pass@K observations (Yue et al., 2025) summarized in the upper left. In this diagram, the line width represents the sampling probability of a reasoning path, while the color distinguishes correct paths (green) from incorrect ones (red). If all reasoning paths after applying RLVR are already present in the base model, the reasoning model merely adjusts the sampling probabilities of these existing paths (visualized in dashed lines). This hypothesis effectively accounts for the key observation shown in the upper-right part, where, for a moderately large K , a base LLM can catch up to the reasoning model after RLVR using the Pass@K metric. In this study, we unveil the **extended reasoning capability boundary** in math tasks using a refined metric, **CoT-Pass@K**, which emphasizes both the correctness of answers and the validity of reasoning CoTs. Verification results supporting this figure are publicly available at https://huggingface.co/datasets/XumengWen/AIME24-25_CoT_Verification.

Advanced Methods

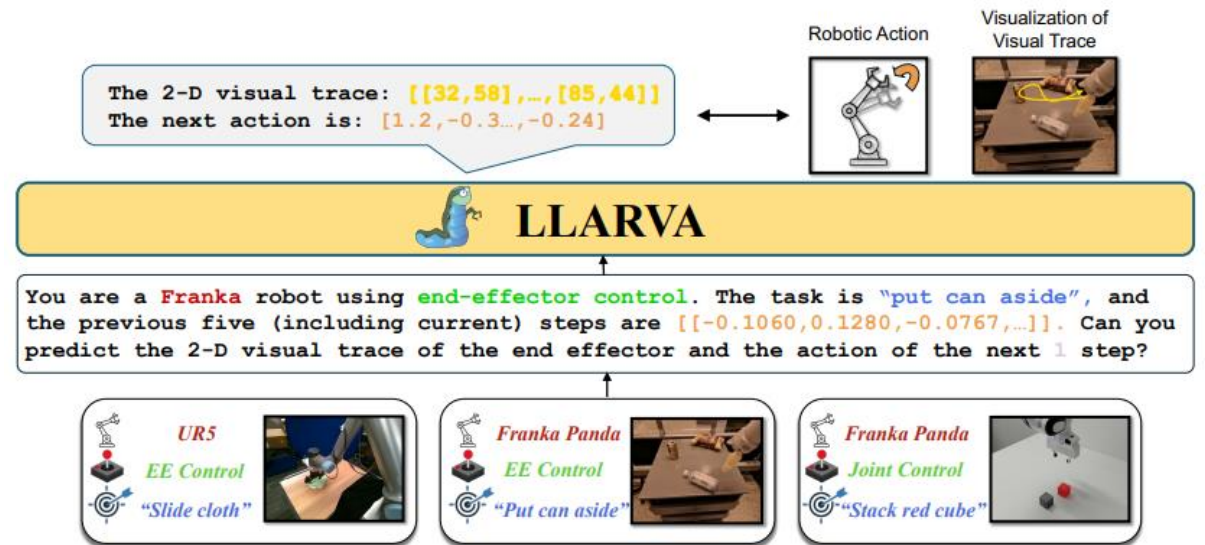
ThinkAct

❖ How to Extract Intermediate CoT & Ground-Truth Latent Plan?

- Intermediate CoT : 입력 프롬프트에 추론 템플릿 제공
- Latent Plan : off-the-shelf detector인 LLARVA 활용하여 비디오에서 실제 경로 추출

Table A1: Reasoning prompt template for reinforced fine-tuning.

Data Type	Prompt Template
2D Manipulation Trajectory	Given an image of a robot manipulation scene and the task instruction "{Instruction}", please generate a sequence of 8 keypoints, representing the gripper's 2D trajectory on the image from its current position to the task-completion position. Please think about this planning process as if you were a human carefully reasoning through the manipulation task. Engage in an internal dialogue while considering the scene, the goal, possible subtasks, the motion path, and any obstacles. It's encouraged to include reflections on the environment, analysis of the goal state, decomposition into subtasks, and any adjustments to the planned trajectory as you think through the process. Provide your detailed reasoning between the <think> </think> tags, and then give your final prediction between the <answer> </answer> tags based on the reasoning. Please provide the trajectory [(x1, y1), (x2, y2), ..., (x8, y8)] with coordinates normalized to [0,1] within <answer> </answer> tags.
QA Tasks	{Question} Please think about this question as if you were a human pondering deeply. Engage in an internal dialogue using expressions such as 'let me think', 'wait', 'Hmm', 'oh, I see', 'let's break it down', etc, or other natural language thought expressions. It's encouraged to include self-reflection or verification in the reasoning process. Provide your detailed reasoning between the <think> </think> tags, and then give your final answer between the <answer> </answer> tags based on the reasoning. (MCQ) Please provide only the single option letter (e.g., A, B, C, D, etc.) within the <answer> </answer> tags. OR (Free-form) Please provide your text answer within the <answer> </answer> tags.



Advanced Methods

ThinkAct

❖ Experiments

- **Pre-training** : Qwen2.5-VL-7B를 VLM 모델로 사용하였으며, 로봇 시연 데이터 뿐만 아니라 QA 데이터셋 (RoboVQA)도 활용하여 추론 능력 강화
- **Fine-tuning** : 시뮬레이션 환경인 SimpleEnv와 LIBERO로만 전이 학습 및 평가

Dataset	Split	Octo-Base [45]	RT1-X [5]	OpenVLA [16]	DiT-Policy [9]	TraceVLA [55]	CoT-VLA [54]	Magma [48]	ThinkAct (Ours)
Simpler-Google (Visual Matching)	Open/Close Drawer	1.0	22.5	49.5	44.9	57.0	–	56.0	50.0
	Move Near	3.0	55.0	47.1	58.9	53.7	–	65.4	72.4
	Pick Coke Can	1.3	52.8	15.3	64.3	28.0	–	83.7	92.0
	Overall	1.8	43.4	37.3	56.0	46.2	–	68.4	71.5
Simpler-Google (Variant Aggregation)	Open/Close Drawer	22.0	56.0	22.5	35.5	31.0	–	53.4	47.6
	Move Near	4.2	34.2	54.0	52.8	56.4	–	65.7	63.8
	Pick Coke Can	17.0	54.0	52.8	56.4	60.0	–	68.8	84.0
	Overall	14.4	48.1	43.1	48.2	49.1	–	62.6	65.1
Simpler-Bridge (Visual Matching)	Put Carrot on Plate	8.3	4.2	4.2	29.4	–	–	31.0	37.5
	Stack Blocks	0.0	0.0	0.0	0.0	–	–	12.7	8.7
	Put Spoon on Towel	12.5	0.0	8.3	34.5	–	–	37.5	58.3
	Put Eggplant in Basket	43.1	0.0	45.8	65.5	–	–	60.5	70.8
	Overall	16.0	1.1	14.6	32.4	–	–	35.4	43.8
LIBERO	Spatial	78.9	–	84.7	82.6	84.6	87.5	–	88.3
	Object	85.7	–	88.4	84.7	85.2	91.6	–	91.4
	Goal	84.6	–	79.2	82.1	75.1	87.6	–	87.1
	Long	51.1	–	53.7	57.6	54.1	69.0	–	70.9
	Overall	75.1	–	76.5	76.8	74.8	83.9	–	84.4

Advanced Methods

ThinkAct

❖ Visualization of Reasoning and Latent Planning

- VLM의 디코딩을 통해 실제 추론 결과 및 행동 계획을 시각화할 수 있어 사후 분석에 용이

Simpler-Google

"Move the 7up can near the apple"

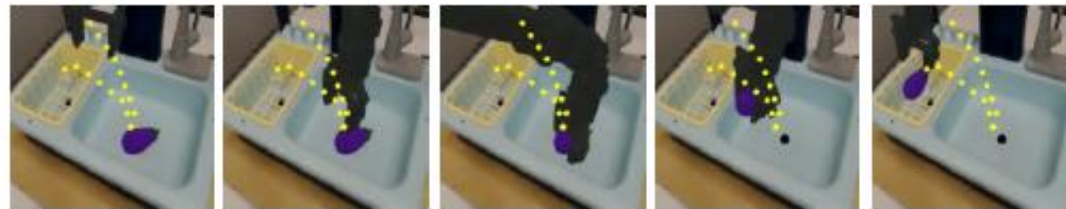
<think>
The image shows a 7up can, a sponge, and an apple on a table. ...
First, I need to locate the 7up can and the apple. ... Now, I need to move the 7up can. ...
I can now place it near the apple. ... The can is now right next to the apple. ...
</think>



Simpler-Bridge

"Put eggplant into yellow basket"

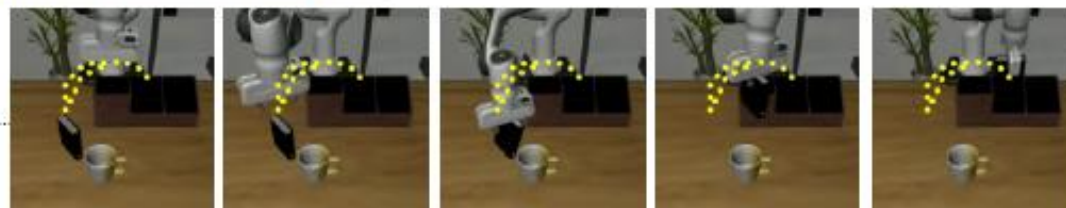
<think>
First, the eggplant is in the sink, and the basket is in the top left corner. ...
The gripper will move toward the eggplant, lift it, and carry it to the basket. ...
Okay, I'm ready to give the final trajectory: move to eggplant, lift it, and place it in basket.
</think>



LIBERO-Long

"Pick up the book and place it in the back compartm."

<think>
Let's break down the task step by step:
First, identify the book on the left side. The back compartment is on the right.
Use the robot's arm near the book to pick it up ...
Move the book smoothly from left to right ...
Place it in the compartment. ...
</think>



Advanced Methods

ThinkAct

❖ Qualitative Analysis for RL-Enhanced Reasoning

- 강화학습 기반 전이 학습 유무에 따른 추론 결과 비교
- 로봇 시연 뿐만 아니라 QA 태스크에서도, 강화학습을 통해 추론 능력이 향상됨을 확인



Q: Is it possible to move the paper to left right now?

ThinkAct w/o RL

<think>Let me think about this for a moment. In the video, I can see that the desk has a few items on it. On the right side, there are already some pieces of paper stacked neatly, and near the front of the desk, I notice some everyday objects—maybe a tissue box, a pen holder, and what looks like a cup.

Wait, the scene feels a bit cluttered, but not overly messy. The lighting also makes it easy to see the shadows of these objects. It seems like a typical workspace with multiple tools laid out for use.

Given the current setup, I'd say everything looks organized and placed intentionally.</think>
<answer>no</answer>

ThinkAct

<think>Let me think about this for a moment. The image shows a desk with various items on it, and a robotic arm is holding a piece of paper above the desk.

The question is whether the paper can be moved to the left right now. Since the paper is already lifted off the surface, the arm has full control of its movement.

Given that, there's nothing preventing it from being moved left. So yes, it's possible.</think>
<answer>yes</answer>

(a) RoboVQA



Q: What shape is the carpet on the floor?

ThinkAct w/o RL

<think>Let me think about this for a moment. The carpet on the floor stands out in the image. It's quite large and has a noticeable floral pattern.

Looking closely, the design seems centered, with blue tones dominating the color scheme. The overall look gives it a decorative and elegant appearance. Yes, it's a large blue carpet with a floral design.</think>

<answer>The carpet on the floor is large and blue with a floral pattern.</answer>

ThinkAct

<think>Let me think about this for a moment. The image shows a room with a carpet on the floor. I'm trying to figure out its shape.

At first glance, the floral pattern in the center draws attention, which might make the shape seem more complex. But looking closely at the edges, they are straight and meet at right angles. The corners are clearly defined, and the sides are parallel.

So, despite the decorative design, the overall outline of the carpet is rectangular. That seems clear now.</think>
<answer>The carpet on the floor is rectangular.</answer>

(b) OpenEQA

Conclusion

Summary

- ❖ TraceVLA (ICLR 2025)
 - 로봇과 주요 객체의 움직임 궤적 (visual trace)를 제공하여 VLA의 시공간 인지 능력 강화
- ❖ CoT-VLA (CVPR 2025)
 - 행동을 예측하기 전에 미래에 달성해야 할 sub-goal 이미지를 생성함으로써, 시각적으로 계획하는 능력 강화
- ❖ ThinkAct (Neurips 2025)
 - 강화학습을 활용해 태스크 추론 및 행동 계획 수립 능력을 강화

고맙습니다