UniSim : Learning Interactive Real-World Simulator

2025.06.10

송건학

요약 정리

- 생성 모델을 활용하여 현실 세계의 다양한 상호작용(조작, 이동 등)을 자연어 기반 동작이 가능한 범용 Simulator인 UniSim을 제안.

- UniSim은 **다양한 데이터셋**에서 수집한 동작과 비디오를 공통형식으로 **통합**하고, observation prediction 기반으로 장기적이고 복잡한 상호작용을 생성

- 현실 세계에서의 Zero-shot Simulation이 가능하며, 실제 에이전트 학습 및 계획에 활용될 수 있는 **실용적 시뮬레이터 플랫폼**으로 작동 가능

Goal & Motivation

- **Ultimate Goal**:
  - 인터넷 데이터로 훈련된 생성 모델의 최종 목표는 인간, 로봇, 기타 에이전트의 행동에 대한 현실적인 경험을 시뮬레이션하는 것.
  - 이는 게임/영화 콘텐츠 제작, 로봇 훈련, 다양한 AI 모델 학습에 활용 가능

- **Challenge**: Real-World Simulator 구축시, 가장 큰 어려움 중 하나는 각기 다른 정보를 보유한 **Dataset**
  - Text-Image Data: 풍부한 장면과 객체는 있지만, 움직임 정보 부족
  - Video-Captioning and Question answering data: high-level 설명은 풍부하지만, low-level 움직임 세부 정보는 부족
  - Human activity data: 풍부한 인간 행동은 있지만, 기계적인 움직임은 부족
  - Robot data: 풍부한 로봇 행동은 있지만, 데이터 양이 제한적

"Motivation: Why Build a Real-World Simulator?"

- Limitation of Existing Models
  - Language models excel at text tasks, not physical tasks
  - Internet-scale models synthesize images and videos
    but lack multi-turn interaction
  - Learning accurate **dynamics** models remains a challenge
  - Most systems limit knowledge sharing by learning one dynamics model per system
  - Existing video generation work doesn’t treat it as a dynamics modeling problem

Proposed Method(UniSim)

- Proposed Method
  - Universal simulator learns from diverse data.
  - Formulated as an observation prediction model
  - Supports consistent simulation across video generation boundaries

- **Detailed Methodology**
  - Simulator instantiated under a unified action-in-video-out interface.
  - Designed to support long-horizon repeated interactions.
  - Visually indistinguishable from the real world, bridging the sim-to-real gap

Proposed Method(UniSim)

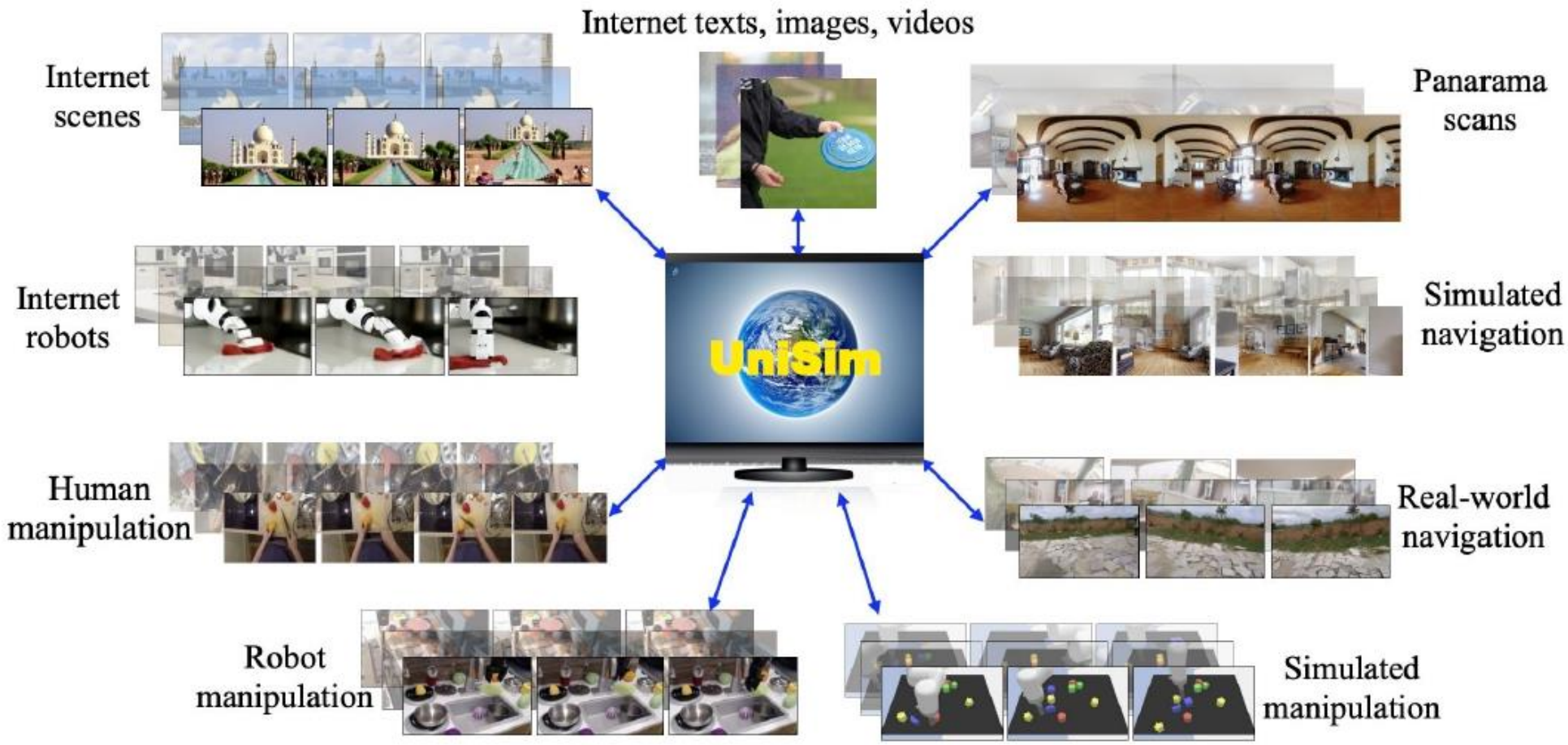


Figure 1: **A universal simulator (UniSim).** The simulator of the real-world learns from broad data with diverse information including objects, scenes, human activities, motions in navigation and manipulation, panorama scans, and simulations and renderings.

2. Learning an Interactive Real-world simulator

- **Real-World Simulator 정의:**
  - 현실 세계 시뮬레이터는 세상의 특정 상태(e.g. image frame)를 input으로 받아,
    어떤 행동(action)이 주어졌을 때 그 행동의 시각적인 결과(e.g. video)를 output으로 제공하는 모델

- **학습의 어려움:**
  - **다양한 행동 형식:** 사용되는 행동이 언어 명령, 로봇 제어, 카메라 움직임 등 여러 형식을 가지고 있음
  - **다양한 비디오 프레임 속도:** 비디오들이 서로 다른 프레임 속도를 가짐

- **제안 방안:**
  - **데이터 처리 전략(orchestrating diverse datasets):**
    각 데이터 유형을 처리하여 행동 공간을 통일(unify the action space)하고,
    가변적인 길이의 비디오를 행동에 맞춰 정렬(align videos of variable lengths to actions).
  - **행동 조건부 비디오 생성 모델 학습(**Simulating Long-Horizon Interactions through observation prediction **):**
    통일된 행동 공간을 사용하여, 다양한 데이터셋에 걸쳐 정보를 융합하는 행동 조건부 비디오 생성 모델
    (action-conditioned video generation model)을 학습.

2.1 orchestrating diverse datasets

데이터 소스 유형	대표 데이터셋	특징	UniSim 처리 방식
Simulated data	Habitat, Language Table	시뮬레이터 기반, 저비용, 행동 라벨 포함	continuous control action을 encoding language embeddings하고 이산화된 제어 값과 concatenate
Real robot data	Bridge Data, RT-1, RT-2	실제 로봇 영상+작업 설명, 제어 형식 다양	로봇별 제어 차이 극복을 위한 task descriptions은 high level action으로 사용& 이산화
Human activity videos	Ego4D, EPIC-KITCHENS, Something-Something V2	풍부한 일상 행동 포함, 비디오 분류 및 행동 인식 레이블 제공	행동 레이블 → 텍스트로 변환, 프레임 샘플링 후 observation chunks 구성
Panorama scans	Matterport3D	정적인 3D 실내 환경 스캔, 직접적 행동(action) 없음	static scan data에 action 정보를 부여하기 위해 scan image를 잘라내어 특정 시점 이미지 생성. 이때, 두 이미지 사이 camera poses 정보를 활용하여 카메라 포즈 변화를 action으로 간주
Internet text-image data	LAION	Action이 없는 정적 이미지, 간접적 행동 묘사	개별 이미지들을 단일 프레임 비디오로 활용, Image caption을 행동(action)으로 간주

For each of these datasets, we process **text tokens** into continuous representations
using **T5** language model embeddings concatenated with low-level actions such as robot control

2.2 Simulating Long-Horizon Interactions through observation prediction

- Simulating Real-World Interactions
  - Observation space O, Action space A
  - Interactive step t, history frame h, extended action a, next set of video frame o
  - 핵심 아이디어는 Simulation를 observation prediction model로 학습

$$p(o_t|h_{t-1}, a_{t-1})$$

- Parametrizing and Training the Simulator

$$\epsilon_\theta(o_t^{(k)}, k|h_{t-1}, a_{t-1}) = (1+\eta)\epsilon_\theta(o_t^{(k)}, k|h_{t-1}, a_{t-1}) - \eta\epsilon_\theta(o_t, k|h_{t-1}), \quad (1)$$

where η controls action conditioning strength. With this parametrization, we train ϵ_θ by minimizing

$$\mathcal{L}_{\text{MSE}} = \left\| \epsilon - \epsilon_\theta\left(\sqrt{1-\beta^{(k)}}o_t + \sqrt{\beta^{(k)}}\epsilon,\ k \middle| h_{t-1}, a_{t-1}\right)\right\|^2,$$

where $\epsilon \sim \mathcal{N}(0, I)$, and $\beta^{(k)} \in \mathbb{R}$ are a set of K different noise levels for each $k \in [1, K]$. Given the learned ϵ_θ, an observation o_t can be generated by sampling from the initial distribution $o_t^{(K)} \sim \mathcal{N}(0, I)$ and iteratively denoising according to the following process for k from K to 0

$$o_t^{(k-1)} = \alpha^{(k)}(o_t^{(k)} - \gamma^{(k)}\epsilon_\theta(o_t^{(k)}, k|h_{t-1}, a_{t-1})) + \xi, \quad \xi \sim \mathcal{N}(0, \sigma_k^2 I), \quad (2)$$

where $\gamma^{(k)}$ is the denoising step size, $\alpha^{(k)}$ is a linear decay on the current denoised sample, and σ_k is a time varying noise level that depends on $\alpha^{(k)}$ and $\beta^{(k)}$.

2.2 Simulating Long-Horizon Interactions through observation prediction

Hyperparameter	Value
Base channels	1024
Optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.99$)
Channel multipliers	1, 2, 4
Learning rate	0.0001
Blocks per resolution	3
Batch size	256
Attention resolutions	6, 12, 24
Num attention heads	16, 16, 8
Conditioning embedding dimension	4096
Conditioning embedding MLP layers: 4	
Conditioning token length	64
EMA	0.9999
Dropout	0.1
Training hardware	512 TPU-v3 chips
Training steps	1000000
Diffusion noise schedule	cosine
Noise schedule log SNR range	[-20, 20]
Sampling timesteps	256
Sampling log-variance interpolation	$\gamma = 0.1$
Weight decay	0.0
Prediction target	ϵ

Table 6: Hyperparameters for training UniSim diffusion model.

- Video U-Net (2022) 사용
- 5.6B parameter,
- 512 TPU-v3 (2018)
  - 현 TPU v6까지 나옴 (25.06.10)
- 20 days training

2.2 Simulating Long-Horizon Interactions through observation prediction

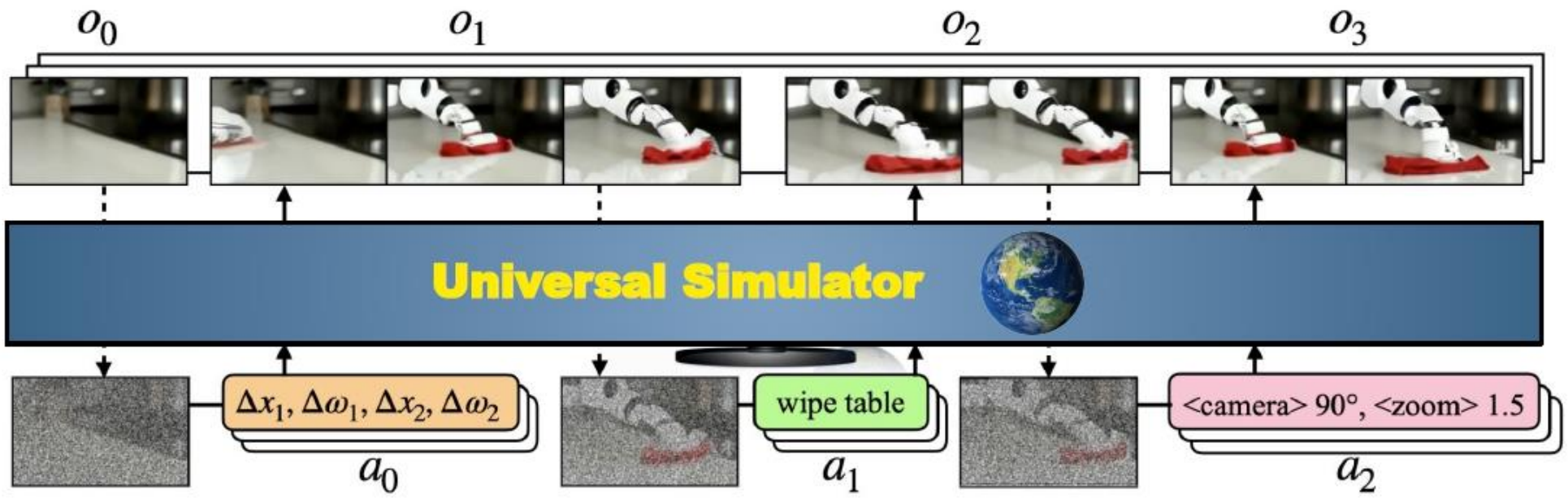


Figure 2: **Training and inference of UniSim.** UniSim is a video diffusion model trained to predict the next (variable length) set of observation frames (o_t) given observations from the past (e.g., o_{t-1}) and action input a_{t-1}. UniSim can handle temporally extended actions in various modalities such as motor controls ($\Delta x_1, \Delta\omega_1, \Delta x_2, ...$), language descriptions ("wipe table"), and actions extracted from camera motions and other sources. Each dotted arrow indicates concatenating the initial noise sample for the next video segment with the previous frame.

3. Simulating Real-World Interactions

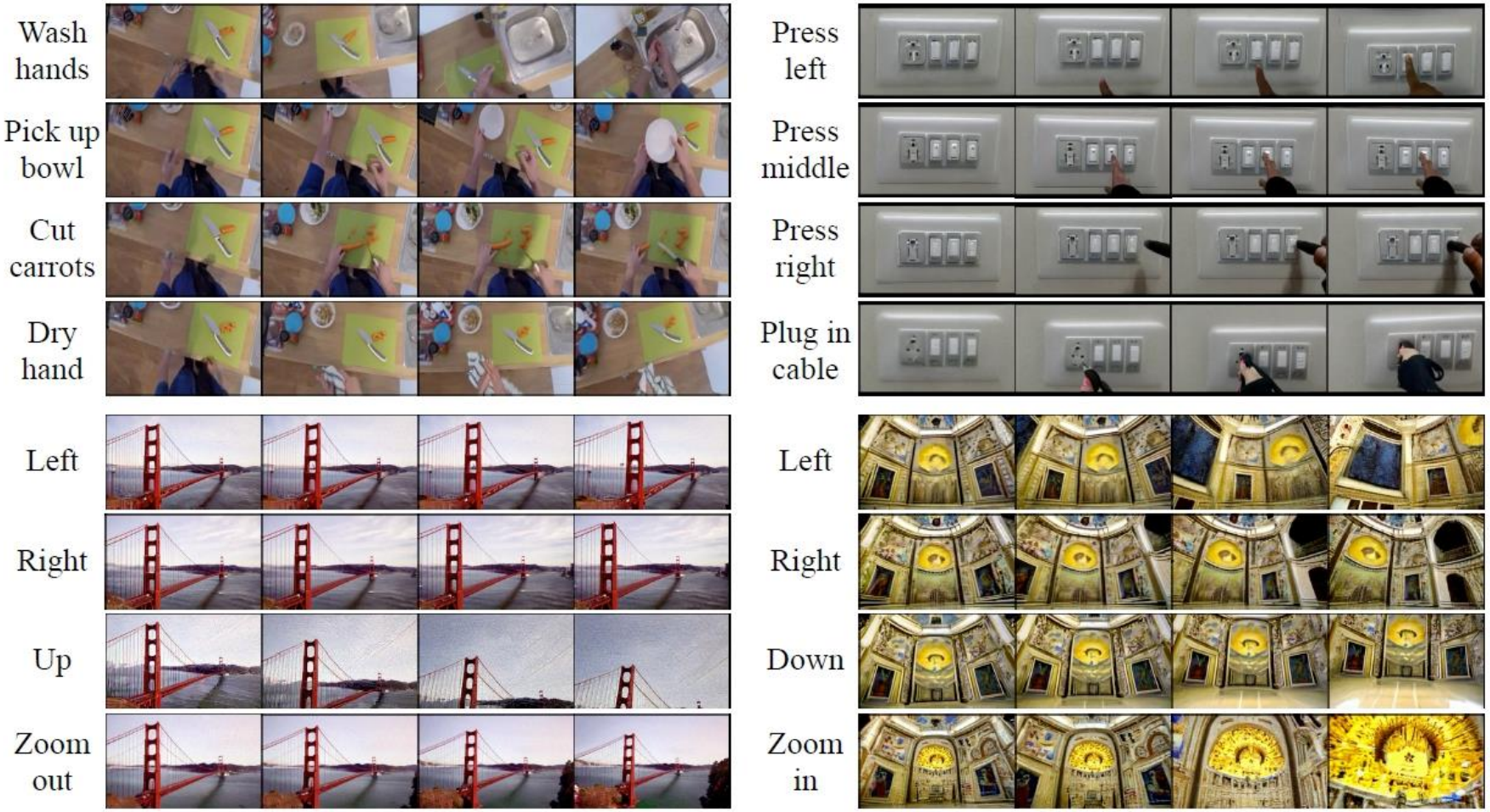


Figure 3: **Action-rich simulations.** UniSim can support manipulation actions such as "cut carrots", "wash hands", and "pickup bowl" from the same initial frame (top left) and other navigation actions.

- 자연어 명령을 통해 복잡하고 다양한 행동을 시뮬레이션

3. Simulating Real-World Interactions

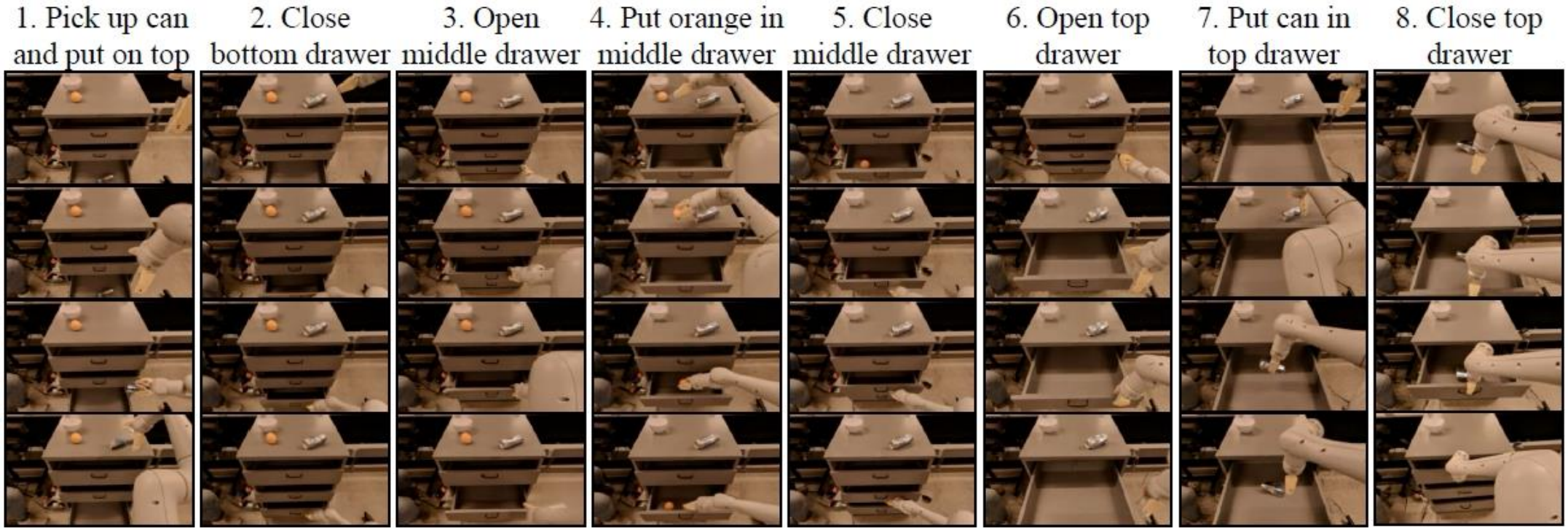


Figure 4: **Long-horizon simulations.** UniSim sequentially simulates 8 interactions autoregressively. The simulated interactions maintain temporal consistency across long-horizon interactions, correctly preserving objects and locations (can on counter in column 2-7, orange in drawer in column 4-5).

- 여러 단계에 걸쳐 순차적인 상호 작용을 시뮬레이션 가능.
- 각 상호 작용 단계는 이전 관찰 결과와 새로운 언어 행동에 기반하여 시뮬레이션 가능.
- 이 과정에서 시뮬레이터는 객체의 위치나 상태를 일관성 있게 유지.

3. Simulating Real-World Interactions

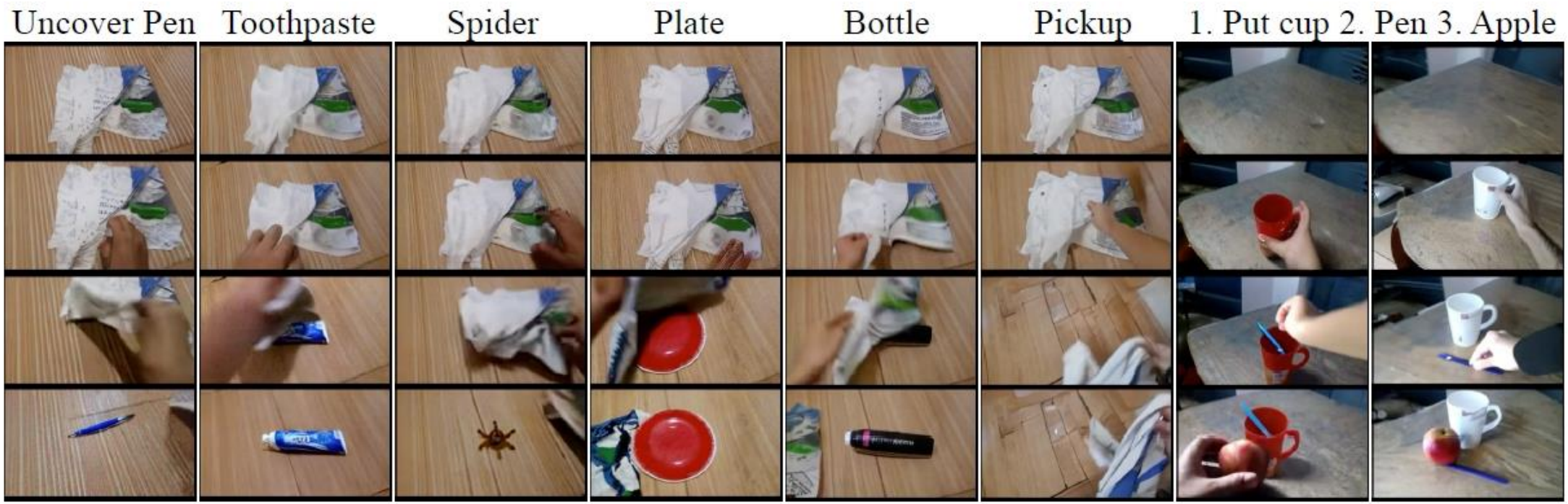


Figure 5: **Diverse and stochastic simulations.** On the left, we use text to specify the object being revealed by suffixing "uncovering" with the object name. On the right, we only specify "put cup" or "put pen", and cups and pens of different colors are sampled as a result of the stochastic sampling process during video generation.

3. Simulating Real-World Interactions

Condition	FID ↓	FVD ↓	IS ↑	CLIP ↑
1 frame	59.47	315.69	3.03	22.55
4 distant	34.89	237	3.43	22.62
4 recent	**34.63**	**211.3**	**3.52**	**22.63**

Table 1: **Ablations of history conditioning** using FVD, FID, and Inception score, and CLIP score on Ego4D. Conditioning on multiple frames is better than on a single frame, and recent history has an edge over distant history.

Figure 6: **Simulations of low-data domains** using the Habitat object navigation using HM3D dataset (Ramakrishnan et al., 2021) with only 700 training examples. Prefixing language actions with dataset identifier leads to video samples that complete the action (top).

Table 1
- 단일 프레임보다 여러 프레임에 조건화하는 것이 더 좋은 성능
- 너무 오래된 과거 프레임보다는 최근 프레임에 조건화하는 것이 더 효과적

Figure 6
- 데이터 양이 적은 도메인에서의 시뮬레이션 성능을 분석
- 훈련 데이터가 적은 경우, 행동(action)에 데이터셋 이름을 함께 입력(prefixing)하면 해당 도메인에서의 시뮬레이션 품질이 향상된다는 것을 발견
- 하지만 다른 도메인으로의 일반화 성능을 저해됨.

4.1 Training Long-horizon Vision-Language policies through hindsight labeling

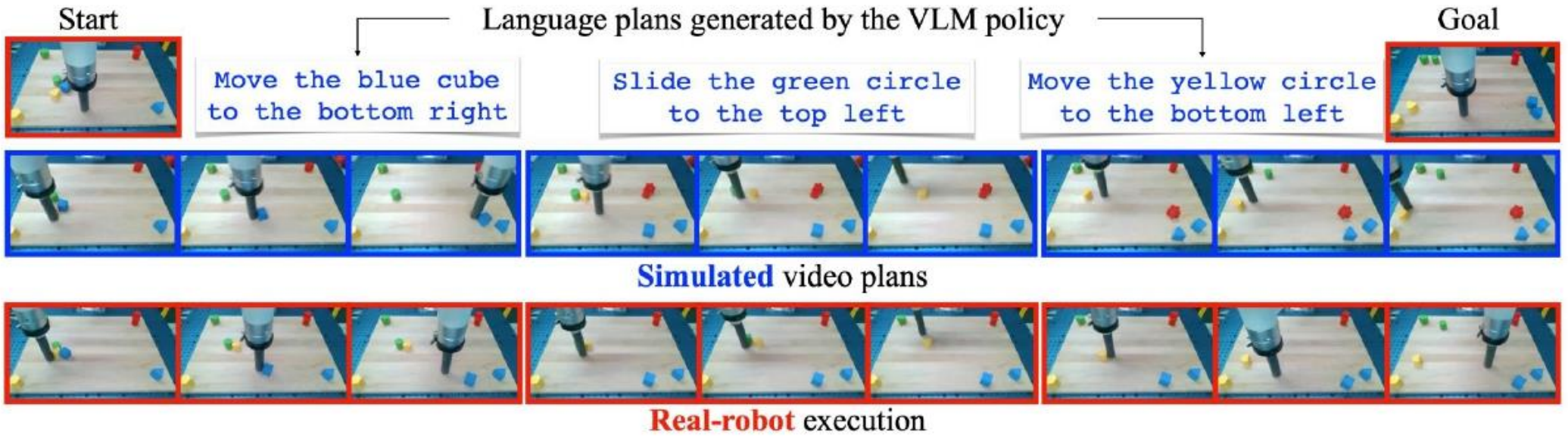


Figure 7: **Long-horizon simulation.** A VLM poliy generates high-level language actions (first row) which are executed in the simulator (middle row) similar to how they are executed in the real world (bottom row) using the Language Table robot. The VLM trained on data from the simulator complete long-horizon tasks by successfully moving three blocks (blue, green, yellow) to match their target location in the goal image.

- 이미지/텍스트 기반으로 작동하는 VLM을 policy로 사용하는 연구가 활발하지만, 길고 복잡한 작업에 필요한 대량의 language action label 데이터 수집이 매우 어려움이 존재.
- UniSim은 VLM 정책 훈련에 필요한 대량의 훈련 데이터를 생성 가능하기에 이를 통해 데이터 부족 문제를 해결하고 학습 효율 향상 가능.

4.2 Real-World simulator for RL

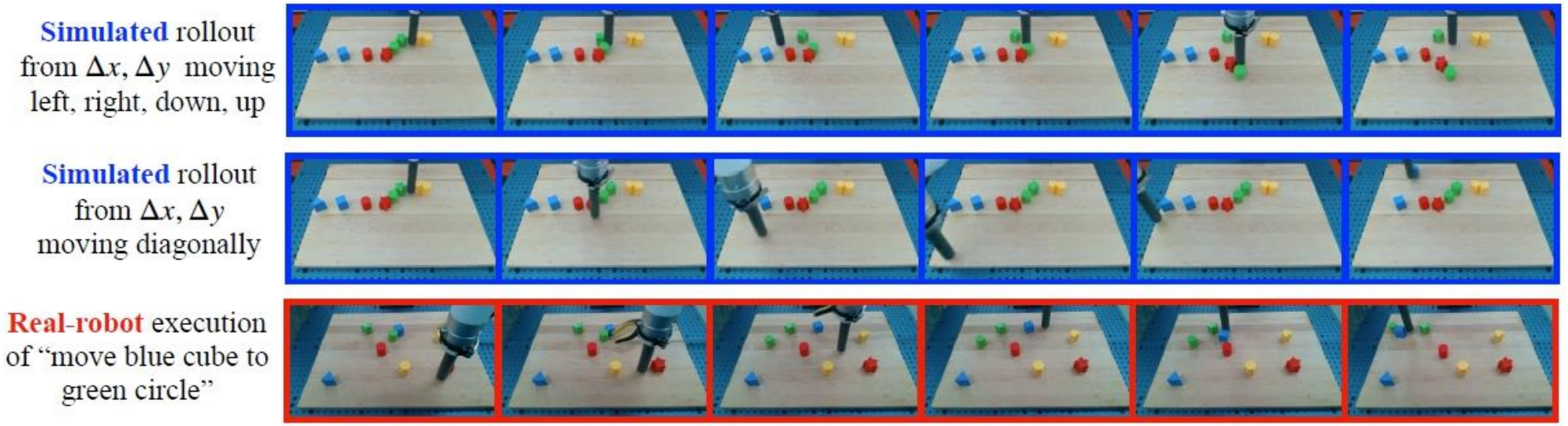


Figure 8: **[Top] Simulation from low-level controls**. UniSim supports low-level control actions as inputs to move endpoint horizontally, vertically, and diagonally. **[Bottom] Real-robot execution of an RL policy** trained in simulation and zero-shot onto the real Language Table task. The RL policy can successfully complete the task of "moving blue cube to green circle".

- 강화 학습(RL)은 가상 게임 환경(e.g. Go, Atari)에서 우수한 성능을 보이지만, 실제 환경에 적용시 물리적 환경에 대한 이해가 부족하다는 큰 제약 존재. (Sim-to-Real Gap 문제)
- UniSim을 통한 현실적인 환경 제공을 통해 RL Agent가 효과적인 훈련을 가능하게 함.

4.3 Realistic Simulator for broader vision-language tasks

	Activity	MSR-VTT	VATEX	SMIT
No finetune	15.2	21.91	13.31	9.22
Activity	54.90	24.88	36.01	16.91
Simulator	46.23	**27.63**	**40.03**	**20.58**

Table 4: **VLM trained in UniSim** to perform video captioning tasks. CIDEr scores for PaLI-X finetuned only on simulated data from UniSim compared to no finetuning and finetuning on true video data from ActivityNet Captions. Finetuning only on simulated data has a large advantage over no finetuning and transfers better to other tasks than finetuning on true data.

- 데이터 수집이 어려운 경우, UniSim을 통해 training data 생성에 활용 할 수 있으며, 이에 기반한 VLM 모델 학습은 video captioning 에서 상당한 성능 향상을 보임.

- UniSim으로 생성된 데이터로만 훈련된 PaLI-X 모델은 Fine-tuning을 하지 않은 경우에 비해 Video Captioning 성능이 크게 향상됨 (ActivityNet 기준 CIDEr 점수 15.2점 → 46.23점).

Limitation

- 환각 (Hallucination)
  - When an action is unrealistic given the scene(예: 테이블 로봇에게 ”손 씻기” 명령), 환각 발생
- 제한된 기억 (Limited memory)
  - 시뮬레이터가 최근 몇 프레임의 기록에만 조건을 부여하기 때문에 장기적인 기억을 포착하지 못함.
  - e.g. 사과를 서랍에 넣는 행동이 조건부 기록에 포함되지 않으면, 서랍을 열 때 사과가 사라질 수 있음.
- 제한된 도메인 외 일반화 (Limited out-of-domain generalization)
  - 훈련 데이터에서 사용 및 확장되지 않은 도메인 영역 내 일반화 능력이 제한적
  - 훈련 데이터를 더욱 확장하는 것이 도움이 될 수 있으나, 부족한 상황
- 시각 시뮬레이션만 (Visual simulation only)
  - UniSim은 action이 시각적 관찰 변화를 유발하지 않는 환경에는 적합하지 않음
  - 진정한 범용 시뮬레이터는 시각적 경험 외의 세계의 모든 측면(예: 소리, 감각 등)을 포착해야 함.

ICLR 2024 Outstanding paper

공식 언급 사유 / Reviewer 평가	상세 설명 및 중요성
중대한 도전 과제 해결 (Tackling a Grand Challenge)	"로봇 공학을 위한 파운데이션 모델... 중요한 단계", AI/로봇 공학의 주요 장기 목표 해결 시도
뛰어난 공학적 성과 (Exceptional Engineering)	데이터 집계 및 통합 인터페이스에서의 "공학적 위업", 다양한 데이터 및 행동 처리의 복잡한 실질적 문제 해결
혁신적인 방법론 (Innovative Methodology)	비전/언어 발전 활용, 융합을 위한 확산 모델 사용, 생성형 AI의 최첨단 응용 제시
높은 실질적 영향력 (High Practical Impact)	유망한 시뮬레이션-현실 전이, 로봇 훈련 혁신 잠재력, 기존 문제에 대한 실질적인 해결책 제시
강력한 동료 검토 지지 (Strong Peer Endorsement)	긍정적인 Reviewer 점수 (8,8,8,6), 해당 분야 전문가들의 높은 수준의 품질 및 수용도 시사

https://openreview.net/forum?id=sFyTZEqmUY

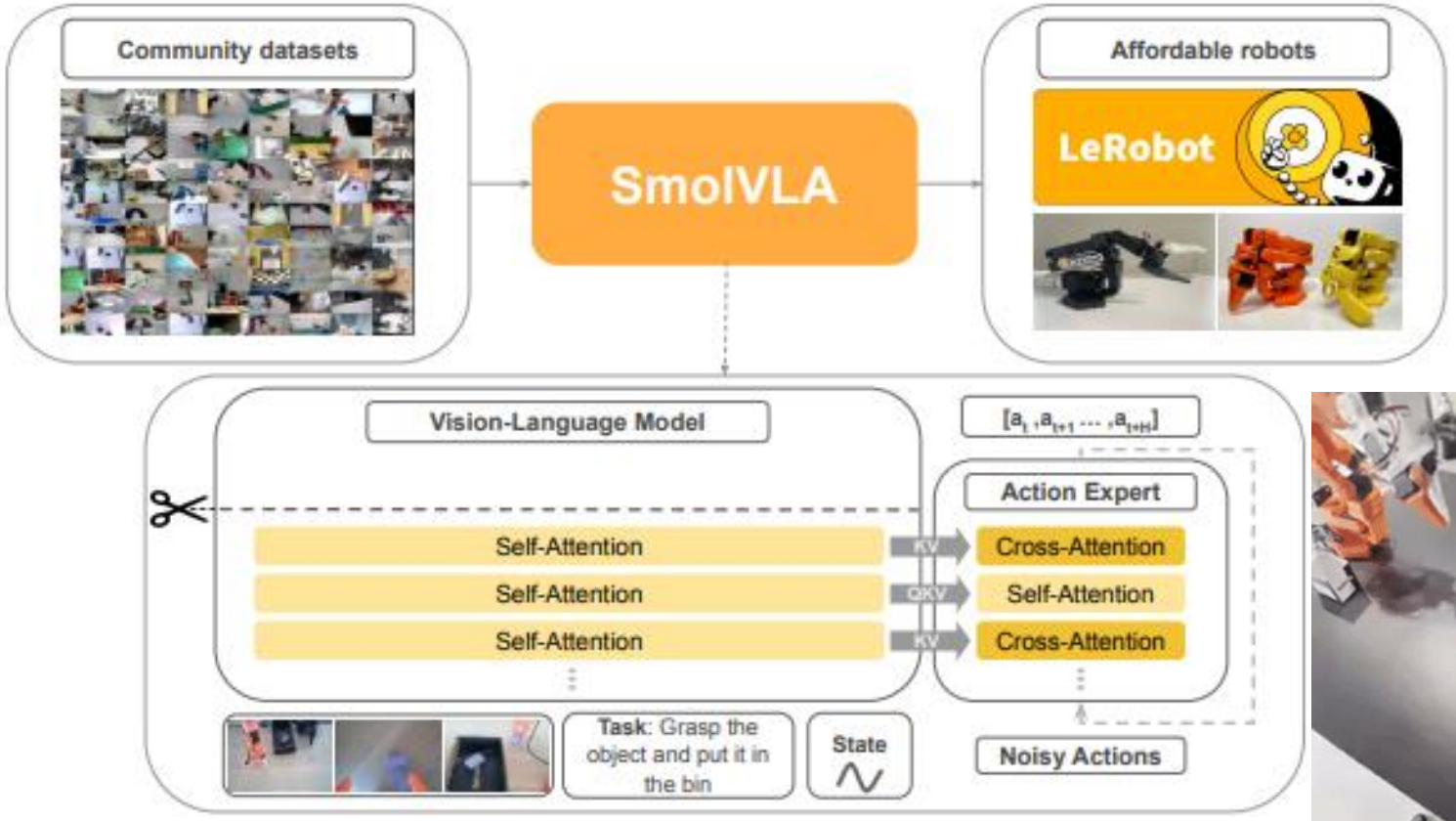
번외) NVIDIA Cosmos (25.02 CES)

- https://github.com/NVIDIA/Cosmos
- https://www.nvidia.com/ko-kr/ai/cosmos/

번외) Figure AI Helix (25.06)

Helix, a generalist Vision-Language-Action (VLA) model
that unifies perception, language understanding,
and learned control to overcome multiple longstanding challenges in robotics

번외) Huggingface SmolVLA (25.06)



SmolVLA: **Efficient Vision-Language-Action Model trained on Lerobot Community Data**
https://huggingface.co/lerobot/smolvla_base
https://arxiv.org/abs/2506.01844