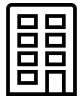
Human Intention-based Data Generation for Robotic Manipulation



Sogang University
Vision & Display Systems Lab, Dept. of Electronic Engineering

Presented By
김태우

Contents

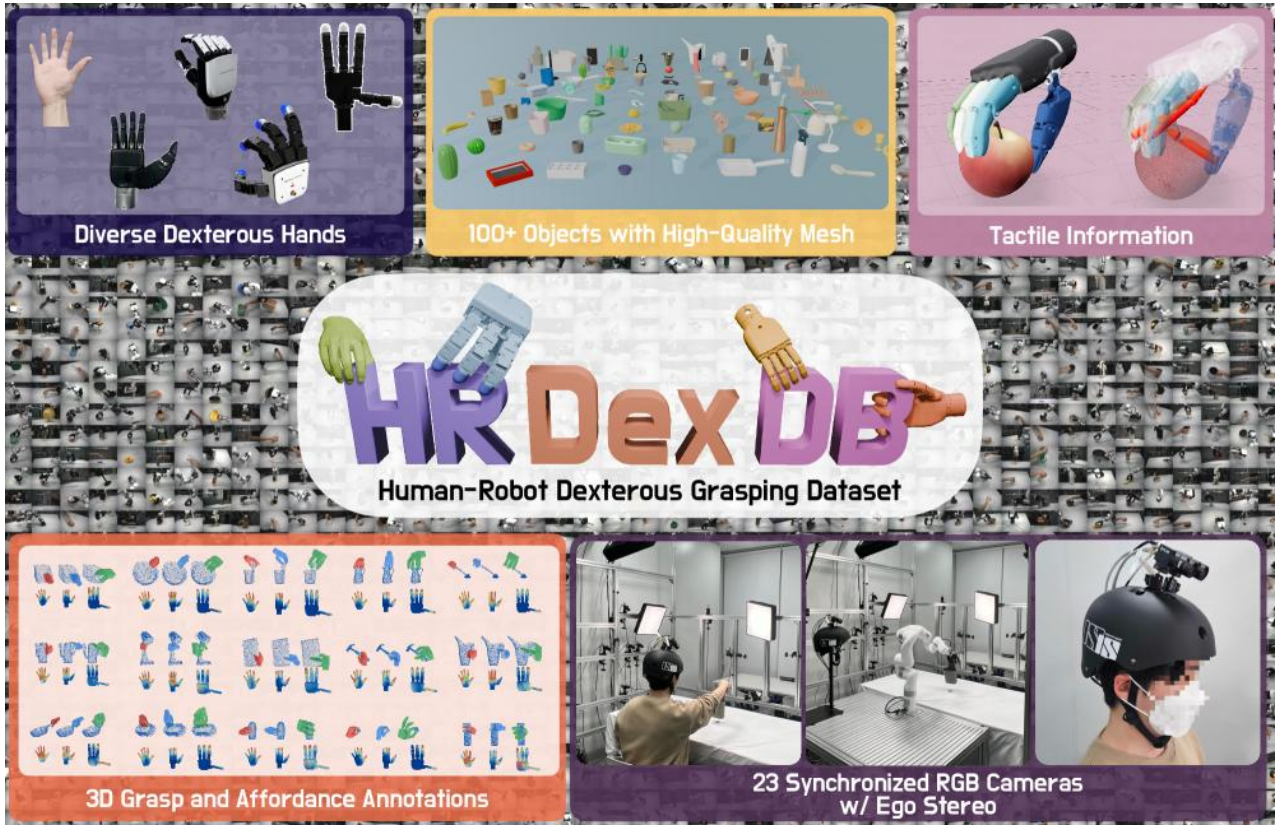
- 주제: Human Intention-based Data Generation for Robotic Manipulation

- 논문 선정 배경
  - Human intention 기반 정책 전이는 연구된 바 있으나, scalable data generator 관점에서는 아직 공백이 있어 향후 유망한 연구 방향으로 판단

- 관련 논문 소개
  - HRDexDB: A Paired Human-Robot Dataset for Cross-Embodiment Dexterous Grasping [arXiv 2026]
  - DynScene: Scalable Generation of Dynamic Robotic Manipulation Scenes for Embodied AI [CVPR 2025]

HRDexDB: A Paired Human-Robot Dataset for Cross-Embodiment Dexterous Grasping

Jongbin Lim, Taeyun Ha, et al. | Seoul National University | arXiv:2604.14944 (2026)



1) 1) O'Neill, Abby, et al. "Open X-Embodiment: Robotic Learning Datasets and RT-X Models." *ICRA*. 2024.

Background

- 기존 데이터셋의 한계: Paired Human-Robot 데이터의 부재
  - 인간 손-물체 상호작용(HOI) 데이터셋: DexYCB, ARCTIC, OakInk2, GigaHands 등
    - 3D 손 포즈, 접촉 주석은 풍부하나 로봇 embodiment와의 대응 관계 없음
  - 로봇 조작 데이터셋: Open X-Embodiment[1)], DROID, AgiBot World 등
    - 규모는 크지만 저자유도 그리퍼 중심, 물체 6D 추적 불완전, 인간 시연과의 페어 없음
  - Paired 수집 시도: RH20T, DexWild, H&R 등
    - 다중 dexterous 로봇 손에 대한 페어 부재, markerless RGB,촉각 신호 부족
  - 시사점: 인간의 손재주가 다양한 로봇 손으로 어떻게 전이되는지 연구할 수 있는
    - "Paired + Multi-Embodiment + 동일 물체" 데이터셋이 필요함

Background

- 연구 분야: Cross-Embodiment Transfer (Dexterous Grasping)
  - 핵심 목표 (Objective)
    - 인간 시연의 파지(grasp) 전략을 형태가 다른 로봇 손으로 전이
  - 핵심 난제: Embodiment Gap
    - 손마다 형태,기구학,실현 가능한 접촉 패턴이 달라 단순 모방(kinematic retargeting)은 한계
  - 데이터 요구사항: 동일 물체에 대한 인간,다중 로봇 손의 paired 파지 데이터 필요

Introduction

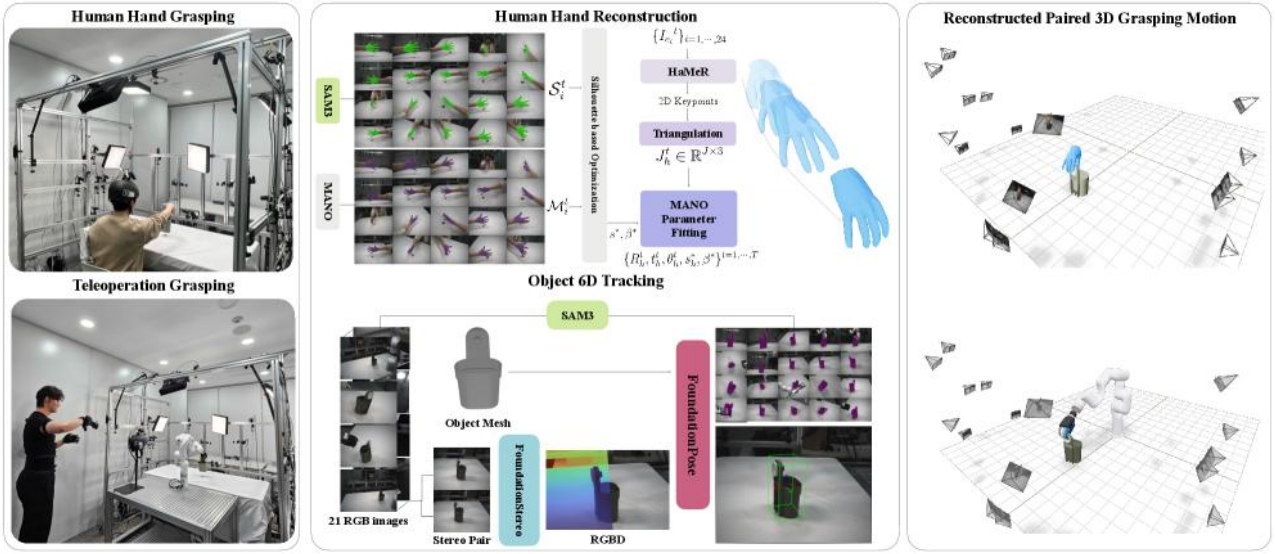
- Contribution
  - 최초의 Paired Multi-Embodiment Dexterous Grasping 데이터셋
    - 인간 손 + 4종 로봇 손(Allegro V4/V5 Plus, Inspire)이 동일한 100개 물체를 파지
    - 2.1K 시퀀스, 24M 프레임 - 마커 없는(markerless) 멀티뷰 RGB 기반
  - 고정밀 멀티모달 주석
    - 21 exocentric + 2 egocentric 동기화 RGB, 3D 손 모션(MANO), 물체 6D pose, 로봇 상태
    - 촉각(tactile) 신호와 파지 성공/실패 라벨 제공
  - 활용성 검증 (Benchmarks)
    - Human-to-Robot 전이: Contact Map Transfer, Cross-Embodiment Grasp Retrieval
    - Perception: 가림이 심한 조작 환경에서 3D 손 포즈,물체 6D pose 벤치마크

Method

- 총 3개의 파트로 설명
  - Part 1: Constructing HRDexDB
    - 멀티모달 캡처 시스템, Paired 수집 프로토콜, 멀티모달 상태 복원(Reconstruction)
  - Part 2: Human-to-Robot Transfer Applications
    - 인간 접촉 정보를 로봇에 맞게 변환하는 Contact Map Transfer
    - 공유 잠재공간에서 로봇 파지를 검색하는 Latent-Space Grasp Retrieval
  - Part 3: Perception Benchmarks
    - 심한 가림 환경에서의 3D Hand Pose Estimation , Object 6D Pose Estimation

Method

- Part 1: Constructing HRDexDB - 캡처 시스템 및 수집 프로토콜
  - 캡처 시스템: 21대 동기화 exocentric RGB 카메라 + 2대 egocentric 스테레오
    - 심한 손-물체 가림(occlusion) 대응 위한 밀집 멀티뷰, 인간용 스테레오 헬멧 착용
  - 로봇 텔레오퍼레이션: Xsens 모션캡처 수트 + MANUS 글러브
  - Paired 수집 프로토콜: 인간 시연 → 동일 물체,작업공간에서 의미적으로 대응되는 로봇 파지 수행



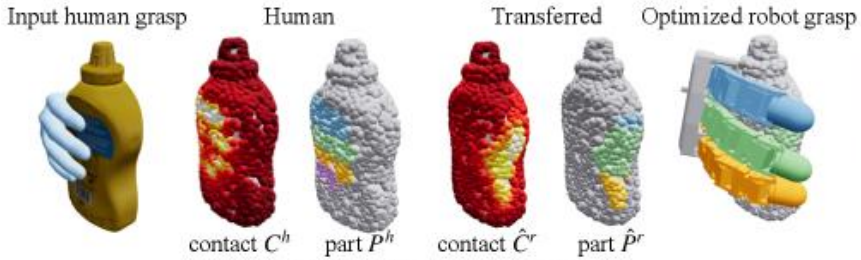
<Capture and Reconstruction Pipeline>

Method

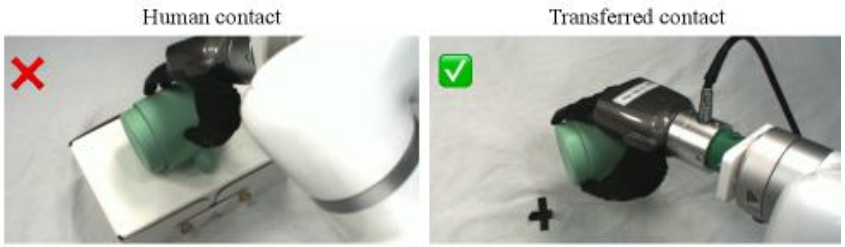
- Part 1: Constructing HRDexDB - 멀티모달 상태 복원(Reconstruction)
  - Human Hand Reconstruction (MANO 기반)
    - HaMeR로 각 뷰의 2D 키포인트 검출 → 삼각측량(Triangulation) → MANO 파라미터 최적화
    - SAM3 마스크 기반 실루엣 정렬로 피험자별 손 형태 보정, 시간적 필터링으로 떨림 제거
  - Object 6D Tracking
    - 스테레오 쌍에서 FoundationStereo로 깊이 추정 + SAM3 물체 마스크
    - RGB-D + CAD 모델로 FoundationPose 6D 추정 → 시간적 추적으로 일관성 유지
    - 모든 캘리브레이션 뷰에 메시를 렌더링, 실루엣 불일치 최소화로 드리프트(drift) 억제
  - 결과: 통일된 월드 좌표계에서 인간,로봇,물체의 정렬된 3D 궤적 확보

Experiments

- Part 2: Human-to-Robot Contact Map Transfer
  - 과제 정의: 인간 접촉 정보를 로봇 손에 맞는 접촉 표현으로 변환
    - 형태,기구학 차이로 인간 접촉 패턴의 직접 모방은 최적이 아님
  - 방법
    - 입력: 인간 접촉 맵(C^h),파트 맵(P^h) + PointNet++ 물체 특징 → 로봇별 접촉,파트 맵 예측
    - 예측된 접촉 맵을 CEDex 물리 기반 최적화의 목적함수로 사용해 파지 합성



(a) Human-to-robot contact transfer



(b) Real-world execution

Experiments

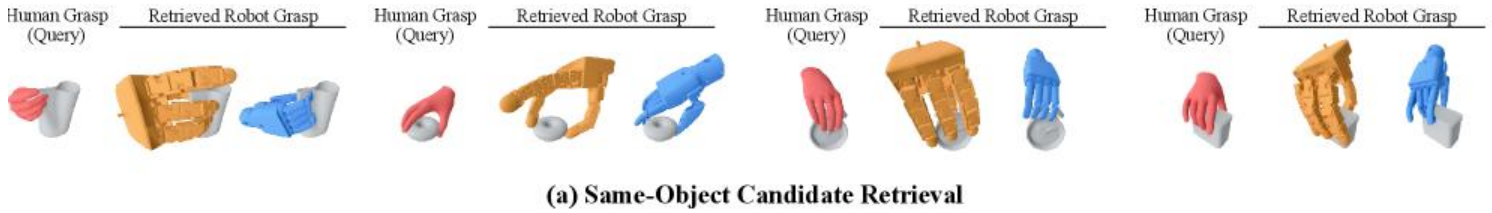
- Part 2: Contact Map Transfer - 결과
  - 실험 설정
    - Sim: Isaac Gym, 6개 직교 축 힘 인가 (1000 trials) / Real: 물체를 들어 10초 유지 시 성공 (60/30 trials)
    - 동일한 CEDex 최적화기 사용 - 접촉 목적함수(Human vs Transferred)만 상이
  - 결과: Transferred-Contact가 시뮬레이션,실제 모두 성공률 향상 (Allegro Real 63.3 → 80.0%)
    - 인간 파지의 기능적 의도는 유지하면서 접촉 분포를 로봇 형태에 맞게 적응

Method	Inspire Sim ↑	Inspire Real ↑	Allegro Sim ↑	Allegro Real ↑
Human-Contact	54.6	66.7	60.2	63.3
Transferred (Ours)	**55.6**	**73.3**	**65.8**	**80.0**

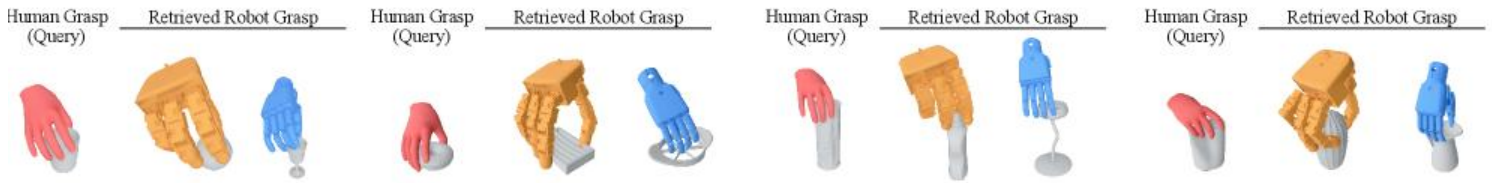
<Table 2. Grasp success rate (%)>

Experiments

- Part 2: Latent-Space Robot Grasp Retrieval
  - 과제 정의: 인간 파지 쿼리로부터 실행 가능한 로봇 파지 후보를 검색하는 공유 잠재 공간 학습
  - 방법: CLIP-style multi-branch 검색 모델
    - 인간 손,Inspire-F1,Allegro-V5 별도 포인트클라우드 인코더 + 공유 물체 인코더
    - Symmetric contrastive loss로 페어링된 파지가 잠재 공간에서 가깝도록 학습
    - 추론 시 유사도 순으로 후보 랭킹 - Same-Object / Cross-Object 두 설정 평가



(a) Same-Object Candidate Retrieval



(b) Cross-Object Candidate Retrieval

Experiments

- Latent-Space Grasp Retrieval - 결과
  - Retrieval 정확도 (33개 후보): 무작위 대비 크게 상회 → embodiment-aware 표현 학습 확인
  - 다운스트림 검증: 검색된 파지로 BODex 최적화 초기화 (7 unseen objects, 33 episodes)
    - Retrieval 초기화가 Vanilla,Kinematic Retargeting 대비 성공률 향상 / Top1 vs Top5 - precision–coverage trade-off

Direction	R@1	R@3	R@5
Human → Inspire	36.4%	81.8%	100.0%
Human → Allegro	24.2%	63.6%	72.7%
Inspire → Allegro	8.2%	57.6%	72.7%

Init Method	Seed (I/A)	Episode (I/A)
Vanilla	3.4 / 16.2	69.7 / 84.9
Kin. Retargeting	3.5 / 1.2	42.4 / 30.3
Retrieval-top5	10.8 / 17.1	75.8 / 93.9
Retrieval-top1	12.2 / 21.3	57.6 / 75.8

<Table 3. Retrieval accuracy / Table 4. BODex initialization success (%)>

Experiments

- Part 3: Benchmarking 3D Hand Pose Estimation
  - 심한 손–물체 가림 상황에서 SOTA 손 복원 모델 평가 (vs FreiHAND)
    - 대부분의 모델에서 오차 증가(Δ>0) → HRDexDB는 더 도전적인 벤치마크
  - HRDexDB를 학습 데이터에 혼합 (2.7M samples) 시 FreiHAND 성능도 향상
    - 중복이 아닌 상보적(complementary) 정보를 제공함을 시사

Model	Ours (J/V)	FreiHAND (J/V)
WiLoR	5.94 / 6.09	5.71 / 5.27
HaMeR	6.15 / 6.16	6.11 / 5.72
Hamba	6.11 / 6.10	6.14 / 5.84
MeshGraphormer	8.31 / 8.10	6.64 / 6.78
FrankMocap	10.61 / 12.48	9.52 / 11.64

<Table 5. PA-MPJPE / PA-MPVPE (mm)>

Method	Baseline (J/V)	+ Ours (J/V)
HaMeR	6.108 / 5.718	6.027 / 5.679
WiLoR	5.711 / 5.273	5.677 / 5.260

<Table 6. Finetuning w/ HRDexDB>

Experiments

- Part 3: Benchmarking Object 6D Pose Estimation
  - 인간,로봇 파지 프레임에서 RGB 기반 6D pose 추정 평가 (FoundPose, GigaPose, PicoPose ± MegaPose)
    - 모든 방법이 로봇 파지에서 성능 저하 → 로봇 링크,핑거팁이 물체 경계와 겹쳐 모호성 유발
  - MegaPose refiner를 HRDexDB robot-grasp 데이터로 미세조정
    - Held-out 환경(OmniRobotHome)에서 ADD-S 10.2% 개선

Method	Human (ADD/ARMSSD)	Robot (ADD/ARMSSD)
FoundPose	6.91 / 44.1	8.74 / 33.3
FoundPose + MegaPose	3.35 / 70.0	4.40 / 64.1
GigaPose	13.10 / 19.7	13.80 / 17.3
GigaPose + MegaPose	5.99 / 54.1	8.02 / 49.2
PicoPose	6.31 / 48.4	8.39 / 38.8

<Table 7. ADD (cm) ↓ / ARMSSD (%) ↑>

Refiner	ADD (cm) ↓	ADD-S (cm) ↓
Original	8.23	4.40
Fine-tuned	7.90	3.95

<Table 8. MegaPose refiner fine-tuning>

Conclusion

- Conclusion & Limitations

  - 결론

    - 최초의 paired 인간-로봇 dexterous 조작 데이터셋 - 다중 embodiment, 100개 물체, 2.1K 시퀀스, 3D 손,6D 물체 주석 및 tactile 신호 제공

    - Contact map transfer, grasp retrieval, 3D hand pose, object 6D pose 벤치마크로 활용성 입증

    - 향후 계획: 1,000개 물체 및 복잡한 functional manipulation 태스크로 확장

  - Limitations

    - Tactile Heterogeneity: 촉각 신호는 로봇 손에만 존재하고 센서 사양이 플랫폼마다 달라 통합 분석이 어려움

    - Trajectory Correspondence: 의미 수준의 페어링만 제공 - 서로 다른 손 형태 간 '기능적으로 동일한 모션'의 정의는 여전히 open problem

DynScene: Scalable Generation of Dynamic Robotic Manipulation Scenes for Embodied AI

CVPR 2025 | Sangmin Lee, Sungyong Park, Heewon Kim (Soongsil University)

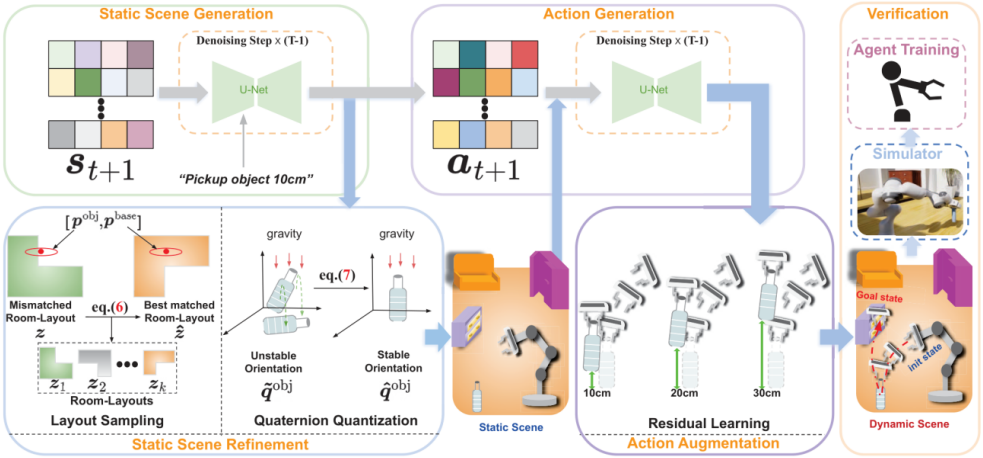
Background

- 기존 텍스트 기반 데이터 생성의 한계
  - Static Scene Generation: 태스크 수행 가능성 미고려 → 부적절한 물체 위치로 실패
  - Action Generation: 환경과의 상호작용 미반영 → 데이터 다양성 제한
  - 두 분야가 독립적으로 발전 → 물리적으로 일관된 (장면+행동) 데이터 부재

<Fig 1. Types of text-guided data generation>

Background

- 연구 분야: Dynamic Scene Generation
  - 핵심 목표 (Objective)
    - 텍스트 지시어 하나로 정적 장면(Static Scene)과 로봇 행동(Action)을 통합 생성
      - Dynamic Scene = Static Scene(초기 배치) + Action Sequence(행동 궤적)
  - 기대 효과
    - 물리적으로 타당하고 다양한 상호작용 데이터를 대규모 자동 생성 → 수작업 수집 병목 해소
  - DynScene [CVPR 2025]: 정적 장면과 행동을 함께 생성하는 최초의 프레임워크



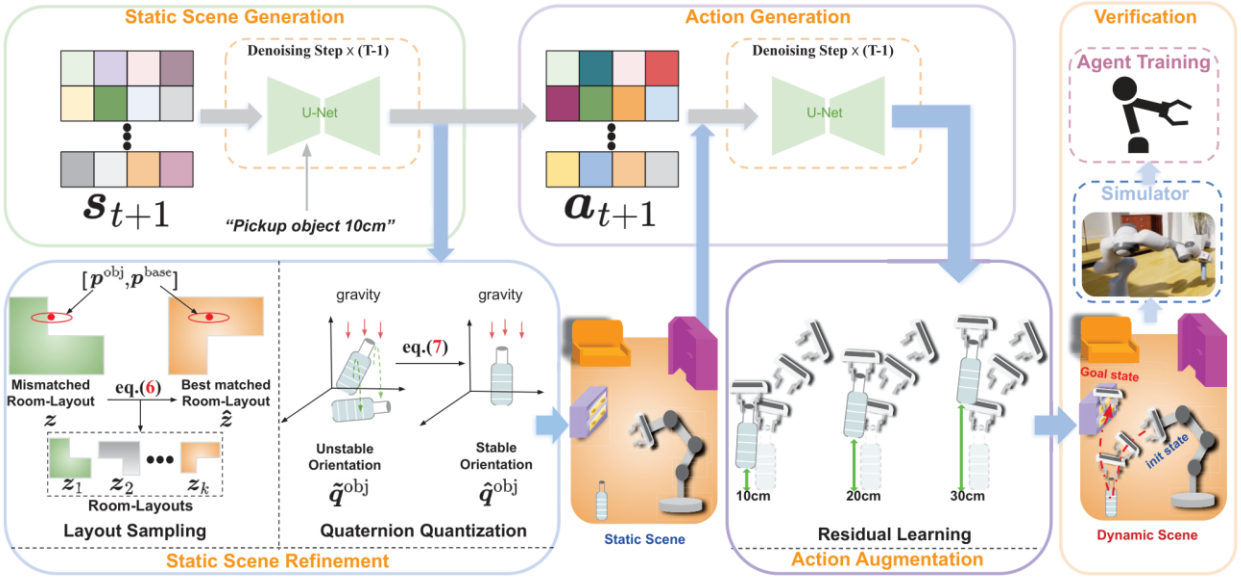
<DynScene: 텍스트 지시어 기반 통합 생성 파이프라인>

Introduction

- Contribution
  - Dynamic Scene 통합 생성 프레임워크
    - 텍스트 지시어로부터 정적 장면과 로봇 행동을 Diffusion 기반으로 공동 생성
  - Residual Action 표현
    - 행동을 End-effector의 변화량(Δp, Δq)으로 표현 → 장면에 독립적, 다양한 배치에 일반화
  - 물리적 타당성 보장 기법
    - Layout Sampling(충돌 없는 배치 선택) + Quaternion Quantization(물체 자세 안정화)
  - 주요 성능
    - 인간 수집 대비 생성 속도 26.8×, 성공률 1.84×, 행동 다양성 +28%
    - 학습된 에이전트 성공률 최대 +19.4% 향상 (ARNOLD benchmark)

Method

- 전체 파이프라인: 총 4단계로 구성
  - ① Static Scene Generation & Refinement: 텍스트 조건 Diffusion(U-Net)으로 장면 생성 후 정제
  - ② Action Generation: 정적 장면을 조건으로 행동 시퀀스 생성
  - ③ Action Augmentation: Residual 표현으로 하나의 장면에서 다수의 행동 증강
  - ④ Verification: Isaac Sim 물리 검증 통과 데이터만 에이전트 학습에 사용



<Fig 2. DynScene framework overview>

Method

- 데이터 표현: Dynamic Scene = (Static Scene *s*, Action *a*)
  - Static Scene - 절대 좌표 표현
    - 물체(o): 위치, 자세(quaternion), 클래스, 크기, 형상 코드(point cloud embedding), 초기/목표 상태
    - 로봇(r): 베이스, end-effector 위치와 자세, 그리퍼 상태 / 룸 레이아웃(z)
  - Action - 잔차(Residual) 좌표 표현
    - $a_k = [\Delta p_k, \Delta q_k, g_k]$: 이전 스텝 대비 end-effector 변화량으로 K개 스텝 표현
    - 절대 위치와 무관 → 동일 태스크면 어떤 장면에도 적용 가능 (Action Augmentation의 기반)

<Fig 6. 형상 코드 학습용 물체 point cloud>

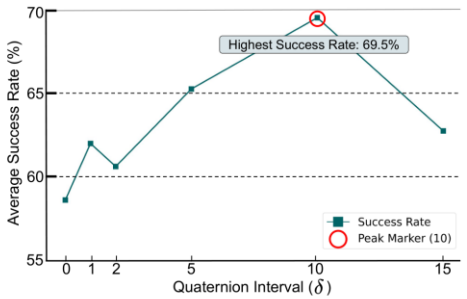
Method

- ① Static Scene 생성 및 정제
  - Conditional Diffusion: 텍스트 지시어를 조건으로 U-Net 기반 노이즈 제거로 정적 장면 생성
  - Layout Sampling: 잘못된 배치 → 로봇-환경 충돌, 접근 불가 발생
    - 예측된 물체, 로봇 위치와 가장 가까운 레이아웃을 데이터셋에서 선택하여 대체
  - Quaternion Quantization: 미세한 자세 오차 → 중력에 의해 물체가 기울거나 쓰러짐
    - 예측 자세를 고정 간격 δ로 양자화(round) → 안정적 초기 배치 확보, δ=10에서 성공률 최고(69.5%)

(a) Improper room layout causing inaccessibility or collision

(b) Improper object orientation leading to instability

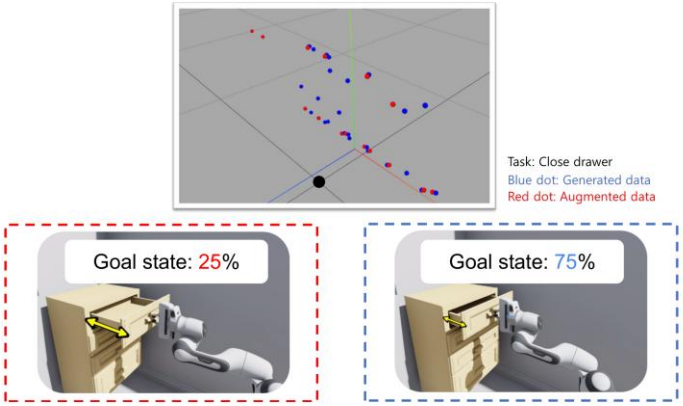
<Fig 3. 부적절한 장면 구성의 문제>



<Fig 10. δ별 성공률>

Method

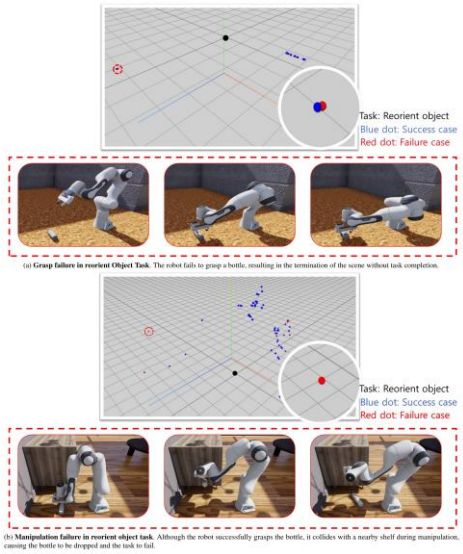
- Part 3: Action 생성 & Augmentation
  - Static Scene 조건 Diffusion
    - 생성된 정적 장면 s를 조건으로 K단계 액션 시퀀스를 동일한 diffusion 구조로 생성
  - Action Augmentation: Residual 표현의 효과
    - 액션이 장면과 분리되어 있어 하나의 정적 장면에 여러 액션 결합 가능
    - 정적 장면 10개 × 액션 10개 = **100**개의 다양한 학습 데이터로 증강
    - 다른 상태에서 학습한 액션을 병합해 원래 장면에 없던 목표 상태(20cm, 30cm 등)도 생성 가능



<Fig 9. Action augmentation (close drawer)>

Method

- Part 4: 검증 및 에이전트 학습
  - Filtering Invalid Dynamic Scenes
    - 생성된 장면을 Isaac Sim에서 실행, 성공한 장면만 학습에 사용
  - Phase-Aware Agent Training (PerAct-PSA)
    - Grasping / Manipulation 단계를 별도 네트워크로 분리 학습
    - 단계별 특화 학습으로 복잡한 조작 태스크에서 더 정밀한 행동 학습



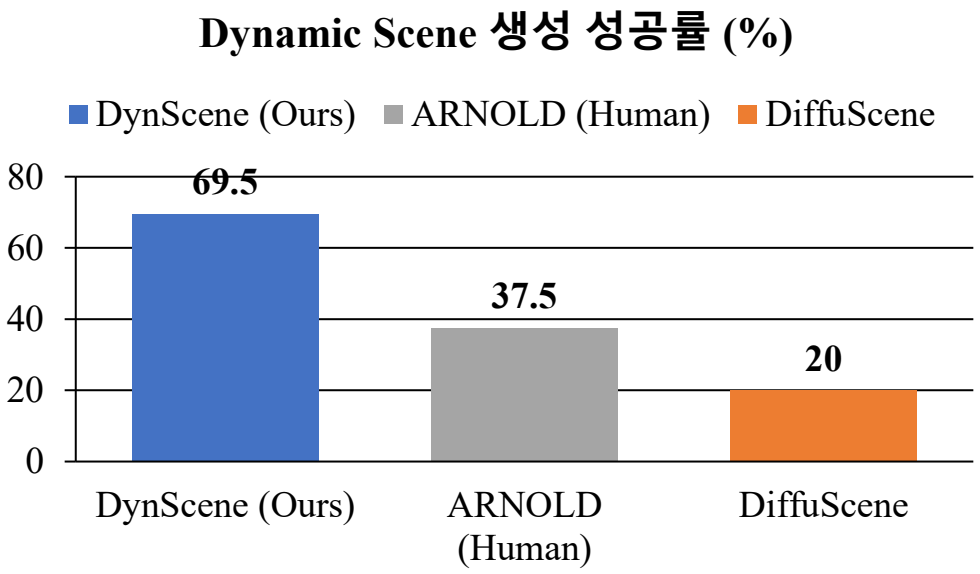
<Fig 7. 시뮬레이션 검증에서 걸러지는 실패 사례>

Experiments

- Setup
  - Dataset: ARNOLD Benchmark
    - 언어 지시 기반 8개 조작 태스크, 연속 상태 지원(예: “서랍을 75% 열어라”)
    - 3,571 train / 773 test 샘플
  - Metrics
    - 행동 다양성: Fréchet Distance(FD), Dynamic Time Warping(DTW), Spatial Coverage(SC)
    - 성공률(SR): 태스크당 100개 생성 장면에 대한 평균 점수
  - Training
    - DynScene: 태스크당 100,000 epochs, Adam(lr 2e-4), RTX 4090
    - 에이전트 학습에는 Isaac Sim 검증을 통과한 장면만 사용

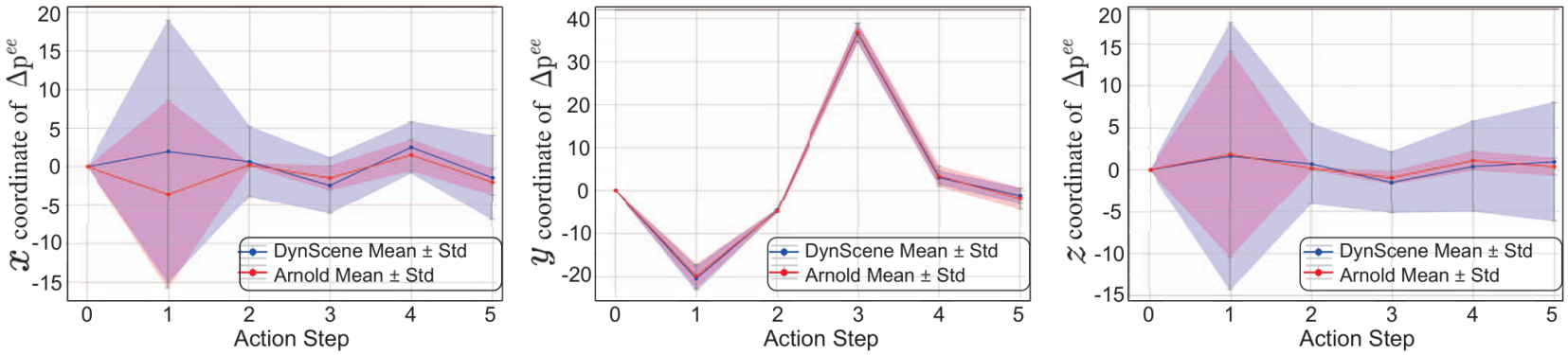
Experiments

- 생성 효율성: 인간 전문가 및 기존 생성 모델과 비교
  - vs 인간 전문가(ARNOLD): 장면당 **2.52**초 vs 67.5초 → **26.8× 빠른 생성**, 성공률 **69.5%** vs 37.5% (**1.84×**)
  - vs DiffuScene(정적 장면 생성 특화 모델): 10× 빠른 생성(2.95s vs 30.24s), 성공률 3.5× (69.5% vs 20.0%)
    - 정적 장면과 외부 액션을 짝짓는 방식의 한계 → 공동 생성(joint generation)의 효과 입증



Experiments

- Action Diversity: 학습 데이터보다 더 다양한 행동 생성
  - 8개 태스크 평균, 모든 다양성 지표에서 ARNOLD 학습 데이터 상회
    - FD **30.32** vs 28.39, DTW **55.25** vs 47.58, SC **3.62** vs 2.83 (약 **28% 높은 다양성**)
  - Pour Water 태스크 분석
    - Δx, Δz 분포는 ARNOLD보다 넓게 형성 → 공간적 다양성 증가
    - Δy 분포는 유사 → 물을 따르는 데 필요한 수직 정밀도 등 태스크 제약은 유지



<Fig 4. Pour Water 태스크의 축별 행동 분포>

Experiments

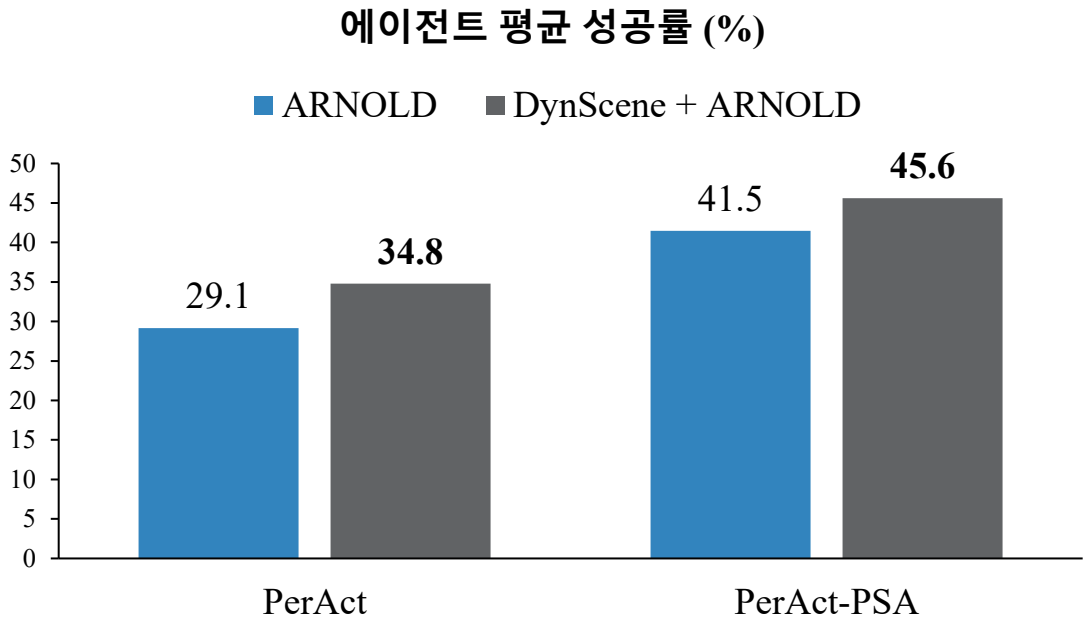
- Text-Conditioned Scene Generation
  - 동일 프롬프트에서 물체 형태, 초기 상태, 공간 배치가 다른 다양한 장면 생성
  - 수치 목표 상태를 정확히 반영
    - "10cm 들어올려라" → **9.99cm**, "180° 기울여라" → **179.94°**, "80% 부어라" → **80.09%**
    - 단, 태스크 난이도에 따라 성공률 편차 존재(캐비닛 25% 열기 40%)

(a) "dump fifty percent of the water from the glass" (b) "adjust the cabinet seventy-five percent closed" (c) "angle the bottle 45 degrees away from the high axis"

<Fig 5. 동일 프롬프트로 생성된 다양한 장면>

Experiments

- Agent 성능: DynScene 데이터 추가 학습 시 성공률 향상
  - ARNOLD test set 평균 성공률(8개 태스크) 일관된 향상
  - 특히 Reorient(3.9→20.0%), Open Cabinet(11.6→16.7%) 등 어려운 태스크에서 큰 폭 향상
  - 일반화: Novel Object(14.5→17.3), Novel Scene(19.0→22.9), Any State(11.0→13.5) 개선, 학습 범위 밖 Novel State는 한계



Conclusion

- 요약
  - DynScene: 텍스트 지시어로부터 정적 장면과 로봇 액션을 공동 생성하는 최초의 dynamic scene 생성 프레임워크
  - 물리 정제(Layout Sampling, Quaternion Quantization)와 residual 액션 표현 기반 증강으로 대규모 자동 데이터 생성 실현
  - 인간 수집 대비 **26.8×** 빠르고 **1.84×** 정확, 행동 다양성 28% 향상, 에이전트 성공률 일관 개선
- Limitation
  - ARNOLD 단일 벤치마크에서만 검증 → 타 도메인 일반화는 미확인
  - 학습 범위 밖(out-of-distribution) 상태 처리 한계 → 향후 더 다양한 상태 표현 연구 필요
- 시사점
  - 수작업 데이터 수집 병목을 우회하는 확장 가능한 시뮬레이션 데이터 생성 방향 제시