Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware

저비용 하드웨어를 활용한 정밀한 양손 조작 학습

Tony Z. Zhao[1] Vikash Kumar[3] Sergey Levine[2] Chelsea Finn[1]
[1] Stanford University [2] UC Berkeley [3] Meta

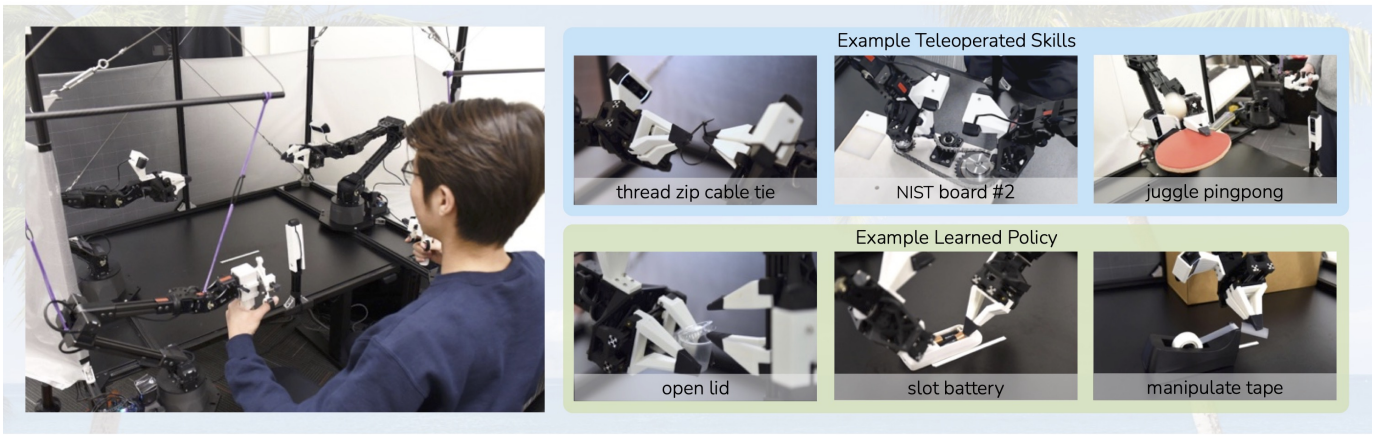


Fig. 1: *ALOHA* : *A Low-cost Open-source Hardware System for Bimanual Teleoperation*. The whole system costs <$20k with off-the-shelf
그림 1: ALOHA: 양손 원격조작을 위한 저비용 오픈소스 하드웨어 시스템. 전체 시스템 비용은 상용 기성품을 사용해도 <$20k이다
robots and 3D printed components. *Left:* The user teleoperates by backdriving the leader robots, with the follower robots mirroring the motion. *Right: ALOHA* is capable of precise, contact-rich, and dynamic tasks. We show examples of both teleoperated and learned skills.

***Abstract*—Fine manipulation tasks, such as threading cable ties or slotting a battery, are notoriously difficult for robots because**
초록—케이블 타이를 끼우거나 배터리를 슬롯에 삽입하는 것과 같은 정밀 조작 과제는 다음과 같은 이유로 로봇에게 매우 어렵다
they require precision, careful coordination of contact forces,
정밀성과 접촉력의 세심한 조정이 필요하고
and closed-loop visual feedback. Performing these tasks typically
폐루프 시각 피드백이 필요하다. 이러한 과제를 수행하려면 일반적으로
requires high-end robots, accurate sensors, or careful calibration,
고급 로봇, 정확한 센서 또는 세심한 보정이 필요하며
which can be expensive and difficult to set up. *Can learning enable low-cost and imprecise hardware to perform these fine manipulation tasks?* We present a low-cost system that performs
이는 비용이 많이 들고 설정하기도 어렵다. 학습을 통해 저비용의 부정확한 하드웨어로 이러한 정밀 조작 과제를 수행할 수 있을까? 본 연구에서는 다음을 수행하는 저비용 시스템을 제시한다
end-to-end imitation learning directly from real demonstrations,
실제 시연으로부터 직접 종단 간 모방학습을 수행하며
collected with a custom teleoperation interface. Imitation learning, however, presents its own challenges, particularly in high-
맞춤형 원격조작 인터페이스를 사용해 수집한다. 그러나 모방학습은 특히 높은
precision domains: errors in the policy can compound over time,
정밀도가 요구되는 영역에서 그 자체의 문제를 제기한다.
정책의 오류가 시간에 따라 누적될 수 있고
and human demonstrations can be non-stationary. To address these challenges, we develop a simple yet novel algorithm, *Action Chunking with Transformers (ACT)*, which learns a generative model over action sequences. ACT allows the robot to learn 6 difficult tasks in the real world, such as opening a translucent
인간의 시연은 비정상적일 수 있다. 이러한 문제를 해결하기 위해 행동 시퀀스에 대한 생성 모델을 학습하는 간단하면서도 새로운 알고리즘인 Action Chunking with Transformers (ACT)를 개발한다. ACT를 사용하면 로봇이 반투명한 것을 여는 등 실제 환경의 6가지 어려운 과제를 학습할 수 있다.
condiment cup and slotting a battery with 80-90% success,
양념 컵을 열고 배터리를 슬롯에 삽입하는 과제를 80-90%의 성공률로 수행하며
with only 10 minutes worth of demonstrations. Project website: tonyzhaozh.github.io/aloha

시연 시간이 10분에 불과해도 가능하다. 프로젝트 웹사이트: tonyzhaozh.github.io/aloha

I. Introduction

I. 서론

Fine manipulation tasks involve precise, closed-loop feedback and require high degrees of hand-eye coordination to adjust and re-plan in response to changes in the environment.
정밀 조작 과제는 정밀한 폐루프 피드백을 포함하며, 환경 변화에 대응해 조정하고 다시 계획하기 위해 높은 수준의 손과 눈의 협응을 요구한다.

Examples of such manipulation tasks include opening the lid of a condiment cup or slotting a battery, which involve delicate operations such as pinching, prying, and tearing rather than broad-stroke motions such as picking and placing. Take opening
이러한 조작 과제의 예로는 양념 컵의 뚜껑을 열거나 배터리를 슬롯에 삽입하는 작업이 있으며, 여기에는 집어 올려 놓기와 같은 큰 동작이 아니라 집기, 들어 올려 벌리기, 찢기와 같은 섬세한 조작이 포함된다.
the lid of a condiment cup in Figure 1 as an example, where the cup is initialized upright on the table: the right gripper needs to first tip it over, and nudge it into the opened left gripper. Then the left gripper closes gently and lifts the cup
그림 1의 양념 컵 뚜껑 열기를 예로 들어 보자. 컵은 테이블 위에 똑바로 세워진 상태로 초기화되어 있으며, 오른쪽 그리퍼가 먼저 컵을 기울인 다음 열린 왼쪽 그리퍼 쪽으로 살짝 밀어야 한다.
off the table. Next, one of the right fingers approaches the
그런 다음 왼쪽 그리퍼가 부드럽게 닫히면서 컵을 테이블에서 들어 올린다.
cup from below and pries the lid open. Each of these steps
다음으로 오른쪽 손가락 중 하나가 컵 아래쪽으로 접근해 뚜껑을 들어 올려 연다.
requires high precision, delicate hand-eye coordination, and rich contact. Millimeters of error would lead to task failure.
이 각각의 단계에는 높은 정밀도와 섬세한 손과 눈의 협응, 풍부한 접촉이 필요하다. 수 밀리미터의 오차만으로도 과제가 실패한다.

Existing systems for fine manipulation use expensive robots and high-end sensors for precise state estimation [29, 60, 32, 41]. In this work, we seek to develop a low-cost system for fine
기존의 정밀 조작 시스템은 정확한 상태 추정을 위해 고가의 로봇과 고급 센서를 사용한다 [29, 60, 32, 41].
manipulation that is, in contrast, accessible and reproducible.
본 연구에서는 이와 대조적으로 접근성과 재현성을 갖춘 정밀 조작용 저비용 시스템을 개발하고자 한다.

However, low-cost hardware is inevitably less precise than high-end platforms, making the sensing and planning challenge more pronounced. One promising direction to resolve this is
그러나 저비용 하드웨어는 필연적으로 고급 플랫폼보다 정밀도가 낮으므로 감지와 계획의 문제가 더욱 두드러진다.
to incorporate learning into the system. Humans also do not
이를 해결할 유망한 방향 중 하나는 시스템에 학습을 도입하는 것이다.
have industrial-grade proprioception [71], and yet we are able to perform delicate tasks by learning from closed-loop visual feedback and actively compensating for errors. In our system,
인간 역시 산업용 수준의 고유수용감각을 갖고 있지는 않지만 [71], 폐루프 시각 피드백을 통해 학습하고 오류를 능동적으로 보상함으로써 섬세한 과제를 수행할 수 있다.
we therefore train an end-to-end policy that directly maps RGB images from commodity web cameras to the actions.
따라서 우리의 시스템에서는 상용 웹 카메라에서 얻은 RGB 이미지를 행동으로 직접 매핑하는 종단 간 정책을 학습한다.

This pixel-to-action formulation is particularly suitable for fine manipulation, because fine manipulation often involves objects with complex physical properties, such that learning the manipulation policy is much simpler than modeling the whole environment. Take the condiment cup example: modeling the
이러한 픽셀-행동 정식화는 정밀 조작에 특히 적합하다. 정밀 조작에는 복잡한 물리적 특성을 지닌 물체가 자주 포함되므로 전체 환경을 모델링하는 것보다 조작 정책을 학습하는 것이 훨씬 단순하기 때문이다.
contact when nudging the cup, and also the deformation when prying open the lid involves complex physics on a large number of degrees of freedom. Designing a model accurate enough for
양념 컵을 예로 들면, 컵을 살짝 밀 때의 접촉과 뚜껑을 들어 올려 열 때의 변형을 모델링하려면 많은 자유도에 걸친 복잡한 물리 현상을 다뤄야 한다.
planning would require significant research and task specific engineering efforts. In contrast, the policy of nudging and
계획에 충분할 정도로 정확한 모델을 설계하려면 상당한 연구와 과제별 엔지니어링 작업이 필요하다.
opening the cup is much simpler, since a closed-loop policy can react to different positions of the cup and lid rather than precisely anticipating how it will move in advance.

이와 대조적으로 컵을 살짝 밀고 여는 정책은 훨씬 단순하다. 폐루프 정책은 컵과 뚜껑이 앞으로 어떻게 움직일지를 정확히 예측하는 대신 서로 다른 위치에 반응할 수 있기 때문이다.

Training an end-to-end policy, however, presents its own challenges. The performance of the policy depends heavily
그러나 종단 간 정책을 학습하는 데에도 고유한 어려움이 따른다.
on the training data distribution, and in the case of fine manipulation, high-quality human demonstrations can provide tremendous value by allowing the system to learn from human dexterity. We thus build a low-cost yet dexterous teleoperation
정책의 성능은 학습 데이터 분포에 크게 좌우되며, 정밀 조작의 경우 고품질 인간 시연은 시스템이 인간의 손재주로부터 학습할 수 있게 하므로 매우 큰 가치를 제공한다.
system for data collection, and a novel imitation learning algorithm that learns effectively from the demonstrations. We
따라서 데이터 수집을 위해 저비용이면서도 손재주가 뛰어난 원격조작 시스템을 구축하고, 시연으로부터 효과적으로 학습하는 새로운 모방학습 알고리즘을 개발한다.
overview each component in the following two paragraphs.
다음 2개 문단에서 각 구성 요소를 개괄한다.

Teleoperation system. We devise a teleoperation setup with
원격조작 시스템
two sets of low-cost, off-the-shelf robot arms. They are
저비용 상용 로봇 팔 2세트를 사용하는 원격조작 구성을 고안한다.
approximately scaled versions of each other, and we use joint-space mapping for teleoperation. We augment this setup with
두 로봇 팔은 서로 대략적으로 크기를 조정한 형태이며, 원격조작을 위해 관절 공간 매핑을 사용한다.
3D printed components for easier backdriving, leading to a highly capable teleoperation system within a $20k budget. We
이 구성에 3D 프린팅 부품을 추가하여 역구동을 더 쉽게 만들었으며, 그 결과 20k달러 예산으로 높은 성능의 원격조작 시스템을 구현했다.
showcase its capabilities in Figure 1, including teleoperation of precise tasks such as threading a zip tie, dynamic tasks such as juggling a ping pong ball, and contact-rich tasks such as assembling the chain in the NIST board #2 [4].
그림 1에서 이 시스템의 성능을 선보인다. 여기에는 지퍼 타이를 끼우는 정밀한 작업, 탁구공을 저글링하는 동적인 작업, NIST 보드 #2에서 체인을 조립하는 접촉이 풍부한 작업 등이 포함된다 [4].

Imitation learning algorithm. Tasks that require precision and
모방학습 알고리즘
visual feedback present a significant challenge for imitation learning, even with high-quality demonstrations. Small errors
정밀도와 시각 피드백을 요구하는 작업은 고품질 시연이 있더라도 모방학습에 상당한 난제를 제기한다.
in the predicted action can incur large differences in the state, exacerbating the "compounding error" problem of imitation learning [47, 64, 29]. To tackle this, we take inspiration from
예측된 행동의 작은 오차도 상태에 큰 차이를 유발할 수 있으며, 이는 모방학습의 '누적 오차' 문제를 악화시킨다 [47, 64, 29].
action chunking, a concept in psychology that describes how sequences of actions are grouped together as a chunk, and executed as one unit [35]. In our case, the policy predicts the
이를 해결하기 위해, 행동 시퀀스를 하나의 청크로 묶어 하나의 단위로 실행하는 과정을 설명하는 심리학 개념인 행동 청킹에서 착안한다 [35].
target joint positions for the next k timesteps, rather than just one step at a time. This reduces the effective horizon of the task
이 경우 정책은 한 번에 1개 단계만 예측하는 대신 다음 k개 타임스텝에 대한 목표 관절 위치를 예측한다.
by k-fold, mitigating compounding errors. Predicting action
이를 통해 작업의 유효 시간 범위를 k배 줄여 누적 오차를 완화한다.
sequences also helps tackle temporally correlated confounders [61], such as pauses in demonstrations that are hard to model with Markovian single-step policies. To further improve the
행동 시퀀스를 예측하면 시간적으로 상관된 교란 요인도 해결하는 데 도움이 된다 [61]. 예를 들어 시연 중의 일시 정지는 마르코프 단일 단계 정책으로 모델링하기 어렵다.
smoothness of the policy, we propose *temporal ensembling*, which queries the policy more frequently and averages across the overlapping action chunks. We implement action chunking
정책을 더 부드럽게 만들기 위해 시간적 앙상블을 제안한다. 이는 정책을 더 자주 질의하고 서로 겹치는 행동 청크들에 대해 평균을 계산하는 방법이다.
policy with Transformers [65], an architecture designed for sequence modeling, and train it as a conditional VAE (CVAE) [55, 33] to capture the variability in human data. We name our
시퀀스 모델링을 위해 설계된 아키텍처인 Transformers [65]를 사용하여 행동 청킹 정책을 구현하고, 인간 데이터의 변동성을 포착하도록 조건부 VAE(CVAE) [55, 33]로 학습한다.
method *Action Chunking with Transformers* (ACT), and find that it significantly outperforms previous imitation learning algorithms on a range of simulated and real-world fine manipulation tasks.
우리의 행동 청킹 방법을 Action Chunking with Transformers (ACT)라고 명명하며, 다양한 시뮬레이션 및 실제 정밀 조작 작업에서 기존 모방학습 알고리즘보다 훨씬 뛰어난 성능을 보임을 확인했다.

The key contribution of this paper is a low-cost system for learning fine manipulation, comprising a teleoperation system and a novel imitation learning algorithm. The teleoperation
이 논문의 핵심 기여는 원격조작 시스템과 새로운 모방학습 알고리즘으로 구성된, 정밀 조작 학습을 위한 저비용 시스템이다.
system, despite its low cost, enables tasks with high precision and rich contacts. The imitation learning algorithm, Action
원격조작 시스템은 저비용임에도 높은 정밀도와 풍부한 접촉이 필요한 작업을 수행할 수 있게 한다.
Chunking with Transformers (ACT), is capable of learning precise, close-loop behavior and drastically outperforms previous methods. The synergy between these two parts allows learning
모방학습 알고리즘인 행동 청킹 기반 Action Chunking with Transformers (ACT)는 정밀한 폐루프 동작을 학습할 수 있으며 기존 방법보다 훨씬 뛰어난 성능을 보인다.
of 6 fine manipulation skills directly in the real-world, such as opening a translucent condiment cup and slotting a battery with 80-90% success, from only 10 minutes or 50 demonstration trajectories.
이 2개 구성 요소의 결합을 통해 실제 환경에서 6개의 정밀 조작 기술을 직접 학습할 수 있다. 예를 들어 반투명 양념 컵 열기와 배터리 삽입을 10분 또는 50개의 시연 궤적만으로 80~90%의 성공률로 학습할 수 있다.

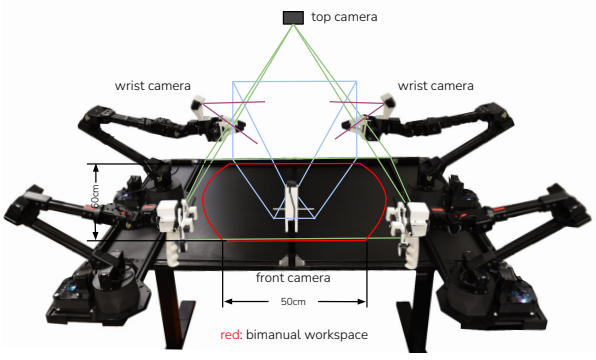
II. Related Work

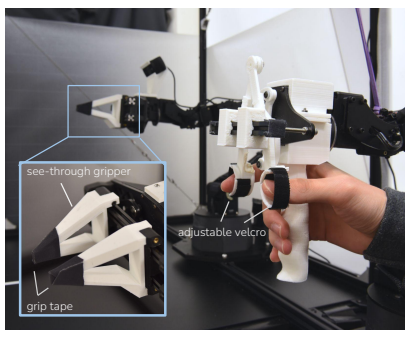
II. 관련 연구

Imitation learning for robotic manipulation. Imitation learn-
로봇 조작을 위한 모방학습
ing allows a robot to directly learn from experts. Behavioral
모방학습을 통해 로봇은 전문가로부터 직접 학습할 수 있다.
cloning (BC) [44] is one of the simplest imitation learning algorithms, casting imitation as supervised learning from observations to actions. Many works have then sought to
행동 복제(BC) [44]는 가장 단순한 모방학습 알고리즘 중 하나로, 관찰에서 행동으로의 매핑을 지도학습으로 처리한다.
improve BC, for example by incorporating history with various architectures [39, 49, 26, 7], using a different training objective [17, 42], and including regularization [46]. Other works
이후 많은 연구에서 다양한 아키텍처를 통해 이력을 통합하거나 [39, 49, 26, 7], 다른 학습 목적 함수를 사용하거나 [17, 42], 정규화를 포함하는 방식으로 [46] BC를 개선하고자 했다.
emphasize the multi-task or few-shot aspect of imitation learning [14, 25, 11], leveraging language [51, 52, 26, 7], or exploiting the specific task structure [43, 68, 28, 52]. Scaling
다른 연구에서는 모방학습의 다중 작업 또는 퓨샷 측면을 강조하며 [14, 25, 11], 언어를 활용하거나 [51, 52, 26, 7], 특정 작업 구조를 이용한다 [43, 68, 28, 52].
these imitation learning algorithms with more data has led to impressive systems that can generalize to new objects, instructions, or scenes [15, 26, 7, 32]. In this work, we focus
이러한 모방학습 알고리즘을 더 많은 데이터로 확장함으로써 새로운 객체, 지시 또는 장면으로 일반화할 수 있는 인상적인 시스템이 등장했다 [15, 26, 7, 32].
on building an imitation learning system that is low-cost yet capable of performing delicate, fine manipulation tasks. We
본 연구에서는 저비용이면서도 섬세한 정밀 조작 작업을 수행할 수 있는 모방학습 시스템 구축에 초점을 맞춘다.
tackle this from both hardware and software, by building a high-performance teleoperation system, and a novel imitation learning algorithm that drastically improves previous methods on fine manipulation tasks.
이를 하드웨어와 소프트웨어 양쪽에서 해결하기 위해 고성능 원격조작 시스템과 정밀 조작 작업에서 기존 방법을 크게 향상하는 새로운 모방학습 알고리즘을 구축한다.

Addressing compounding errors. A major shortcoming of BC
누적 오차 해결
is compounding errors, where errors from previous timesteps accumulate and cause the robot to drift off of its training distribution, leading to hard-to-recover states [47, 64]. This
BC의 주요 단점은 누적 오차이다. 이는 이전 타임스텝의 오차가 누적되어 로봇이 학습 분포에서 벗어나게 되고, 그 결과 복구하기 어려운 상태에 이르는 현상이다 [47, 64].
problem is particularly prominent in the fine manipulation setting [29]. One way to mitigate compounding errors is to
이 문제는 정밀 조작 환경에서 특히 두드러진다 [29].
allow additional on-policy interactions and expert corrections, such as DAgger [47] and its variants [30, 40, 24]. However,
누적 오차를 완화하는 한 가지 방법은 DAgger [47] 및 그 변형 [30, 40, 24]과 같이 추가적인 온폴리시 상호작용과 전문가 교정을 허용하는 것이다.
expert annotation can be time-consuming and unnatural with a teleoperation interface [29]. One could also inject noise at
그러나 원격조작 인터페이스를 사용한 전문가 주석 작업은 시간이 오래 걸리고 부자연스러울 수 있다 [29].
demonstration collection time to obtain datasets with corrective behavior [36], but for fine manipulation, such noise injection can directly lead to task failure, reducing the dexterity of teleoperation system. To circumvent these issues, previous
시연 수집 시점에 노이즈를 주입하여 교정 행동이 포함된 데이터셋을 얻을 수도 있지만 [36], 정밀 조작에서는 이러한 노이즈 주입이 직접적으로 작업 실패를 초래하여 원격조작 시스템의 손재주를 저하시킬 수 있다.
works generate synthetic correction data in an offline manner [16, 29, 70]. While they are limited to settings where low-
이러한 문제를 피하기 위해 기존 연구에서는 오프라인 방식으로 합성 교정 데이터를 생성한다 [16, 29, 70].
dimensional states are available, or a specific type of task like grasping. Due to these limitations, we need to address the
저차원 상태를 사용할 수 있는 환경이나 파악과 같은 특정 유형의 과제로 제한된다.
compounding error problem from a different angle, compatible with high-dimensional visual observations. We propose to
이러한 한계로 인해 고차원 시각 관측과 호환되는 다른 관점에서 누적 오차 문제를 해결해야 한다.
reduce the effective horizon of tasks through action chunking, i.e., predicting an action sequence instead of a single action, and then ensemble across overlapping action chunks to produce trajectories that are both accurate and smooth.
행동 청킹을 통해 과제의 유효 시간 범위를 줄이는 방법을 제안한다. 즉, 단일 행동 대신 행동 시퀀스를 예측한 다음, 서로 겹치는 행동 청크에 걸쳐 앙상블하여 정확하면서도 매끄러운 궤적을 생성한다.

Bimanual manipulation. Bimanual manipulation has a long
양손 조작
history in robotics, and has gained popularity with the lowering of hardware costs. Early works tackle bimanual manipulation
양손 조작은 로봇공학에서 오랜 역사를 지니며, 하드웨어 비용이 낮아지면서 인기를 얻었다.
from a classical control perspective, with known environment dynamics [54, 48], but designing such models can be time-consuming, and they may not be accurate for objects with complex physical properties. More recently, learning has been
초기 연구에서는 환경 동역학이 알려져 있다는 가정하에 고전적 제어 관점에서 양손 조작을 다루었지만 [54, 48], 이러한 모델을 설계하는 데는 시간이 많이 걸릴 수 있으며 복잡한 물리적 특성을 지닌 물체에는 정확하지 않을 수 있다.
incorporated into bimanual systems, such as reinforcement learning [9, 10], imitating human demonstrations [34, 37, 59, 67, 32], or learning to predict key points that chain together





ViperX 6dof Arm (follower)

#Dofs	6+gripper
Reach	750mm
Span	1500mm
Repeatability	1mm
Accuracy	5-8mm
Working Payload	750g

Fig. 3: *Left:* Camera viewpoints of the front, top, and two wrist cameras, together with an illustration of the bimanual workspace of *ALOHA*. *Middle:* Detailed view of the "handle and scissor" mechanism and custom grippers. *Right:* Technical spec of the ViperX 6dof robot [1].

그림 3: 왼쪽: 전면, 상단 및 2개의 손목 카메라에서 본 시점과 ALOHA의 양손 작업 공간을 보여주는 그림. 가운데: "handle and scissor" 메커니즘과 맞춤형 그리퍼의 상세 모습. 오른쪽: ViperX 6dof 로봇의 기술 사양 [1].

motor primitives [20, 19, 50]. Some of the works also focus
최근에는 강화학습 [9, 10], 인간 시연 모방 [34, 37, 59, 67, 32], 또는 운동 프리미티브를 연결하는 핵심 지점을 예측하는 학습 [20, 19, 50]과 같이 학습을 양손 시스템에 통합하는 연구가 이루어졌다.
on fine-grained manipulation tasks such as knot untying, cloth flattening, or even threading a needle [19, 18, 31], while using robots that are considerably more expensive, e.g. the da Vinci surgical robot or ABB YuMi. Our work turns to low-cost
일부 연구는 매듭 풀기, 천 펴기, 심지어 바늘에 실 꿰기 [19, 18, 31]와 같은 정밀 조작 과제에도 초점을 맞추지만, da Vinci 수술 로봇이나 ABB YuMi와 같이 훨씬 고가인 로봇을 사용한다.
hardware, e.g. arms that cost around $5k each, and seeks to enable them to perform high-precision, closed-loop tasks. Our
본 연구는 저비용 하드웨어, 예를 들어 팔 하나당 약 $5k인 로봇으로 방향을 전환하고, 이를 통해 높은 정밀도의 폐루프 과제를 수행할 수 있도록 하는 것을 목표로 한다.
teleoperation setup is most similar to Kim et al. [32], which
본 연구의 원격조작 설정은 Kim et al.의 연구와 가장 유사하다.
also uses joint-space mapping between the leader and follower robots. Unlike this previous system, we do not make use of
[32] 이 연구 역시 리더 로봇과 팔로워 로봇 사이에 관절 공간 매핑을 사용한다.
special encoders, sensors, or machined components. We build
이전 시스템과 달리 본 연구에서는 특수 인코더, 센서 또는 기계 가공 부품을 사용하지 않는다.
our system with only off-the-shelf robots and a handful of 3D printed parts, allowing non-experts to assemble it in less than 2 hours.
본 시스템은 기성품 로봇과 소수의 3D 프린팅 부품만으로 제작하므로, 비전문가도 2시간 이내에 조립할 수 있다.

III. ALOHA: A Low-cost Open-source Hardware System for Bimanual Teleoperation

III. ALOHA: 저비용 오픈소스 하드웨어 양손 원격조작을 위한 시스템

We seek to develop an accessible and high-performance teleoperation system for fine manipulation. We summarize our
본 연구는 접근성이 높고 성능이 우수한 정밀 조작용 원격조작 시스템을 개발하고자 한다.
design considerations into the following 5 principles.
설계 고려 사항을 다음 5가지 원칙으로 요약한다.

1) **Low-cost**: The entire system should be within budget for most robotic labs, comparable to a single industrial arm.
1) 저비용: 전체 시스템은 대부분의 로봇 연구실에서 단일 산업용 로봇 팔과 비슷한 비용으로 구매할 수 있는 예산 범위 내에 있어야 한다.
2) **Versatile**: It can be applied to a wide range of fine manipulation tasks with real-world objects.
2) 다목적성: 실제 물체를 대상으로 하는 다양한 정밀 조작 과제에 적용할 수 있어야 한다.
3) **User-friendly**: The system should be intuitive, reliable, and easy to use.
3) 사용자 친화성: 시스템은 직관적이고 신뢰할 수 있으며 사용하기 쉬워야 한다.
4) **Repairable**: The setup can be easily repaired by researchers, when it inevitably breaks.
4) 수리 용이성: 시스템은 필연적으로 고장 나더라도 연구자가 쉽게 수리할 수 있어야 한다.
5) **Easy-to-build**: It can be quickly assembled by researchers, with easy-to-source materials.
5) 제작 용이성: 연구자가 쉽게 구할 수 있는 재료를 사용하여 신속하게 조립할 수 있어야 한다. 사용할 로봇을 선택할 때 원칙 1, 4, 5에 따라

When choosing the robot to use, principles 1, 4, and 5 lead us to build a bimanual parallel-jaw grippers setup with two ViperX 6-DoF robot arms [1, 66]. We do not employ dexterous
ViperX 6-DoF 로봇 팔 2개를 사용하는 양손 평행 죠 그리퍼 시스템을 구축했다 [1, 66].
hands due to price and maintenance considerations. The ViperX
가격과 유지 관리 문제로 인해 다관절 손은 사용하지 않는다.
arm used has a working payload of 750g and 1.5m span, with an accuracy of 5-8mm. The robot is modular and simple to
사용한 ViperX 팔은 작업 하중 750g, 작업 범위 1.5m이며, 정확도는 5-8mm이다.
repair: in the case of motor failure, the low-cost Dynamixel motors can be easily replaced. The robot can be purchased
이 로봇은 모듈식이고 수리가 간단하다. 모터가 고장 나는 경우 저비용 Dynamixel 모터를 쉽게 교체할 수 있다.
off-the-shelf for around $5600. The OEM fingers, however,
이 로봇은 약 $5600에 기성품으로 구매할 수 있다.
are not versatile enough to handle fine manipulation tasks. We
그러나 OEM 손가락은 정밀 조작 과제를 수행하기에 충분히 다목적이지 않다.
thus design our own 3D printed "see-through" fingers and fit it with gripping tape (Fig 3). This allows for good visibility
따라서 자체적으로 3D 프린팅한 "see-through" 손가락을 설계하고 그 위에 그립 테이프를 부착했다(그림 3).
when performing delicate operations, and robust grip even with thin plastic films.
이를 통해 섬세한 조작을 수행할 때 시야를 잘 확보할 수 있으며, 얇은 플라스틱 필름을 사용하더라도 견고하게 잡을 수 있다.

We then seek to design a teleoperation system that is maximally user-friendly around the ViperX robot. Instead
그다음 ViperX 로봇을 중심으로 사용자 친화성이 최대한 높은 원격조작 시스템을 설계하고자 했다.
of mapping the hand pose captured by a VR controller or camera to the end-effector pose of the robot, i.e. task-space mapping, we use direct joint-space mapping from a smaller robot, WidowX, manufactured by the same company and costs $3300 [2]. The user teleoperates by backdriving the smaller
VR 컨트롤러나 카메라로 포착한 손 자세를 로봇 말단 효과기의 자세로 매핑하는 방식, 즉 작업 공간 매핑 대신, 같은 회사에서 제조한 더 작은 로봇 WidowX를 이용한 직접적인 관절 공간 매핑을 사용하며 비용은 $3300이다 [2].
WidowX ("the leader"), whose joints are synchronized with the larger ViperX ("the follower"). When developing the setup, we
사용자는 더 작은 WidowX("리더")를 역구동하여 원격조작하며, 이 로봇의 관절은 더 큰 ViperX("팔로워")의 관절과 동기화된다.
noticed a few benefits of using joint-space mapping compared to task-space. (1) Fine manipulation often requires operating
시스템을 개발하면서 작업 공간 매핑과 비교해 관절 공간 매핑을 사용할 때 몇 가지 이점이 있음을 확인했다.
near singularities of the robot, which in our case has 6 degrees of freedom and no redundancy. Off-the-shelf inverse kinematics
(1) 정밀 조작은 로봇의 특이점 근처에서 작업해야 하는 경우가 많으며, 본 연구의 로봇은 6 자유도를 갖고 중복성이 없다.
(IK) fails frequently in this setting. Joint space mapping, on
이러한 환경에서는 기성품 역기구학(IK)이 자주 실패한다.
the other hand, guarantees high-bandwidth control within the joint limits, while also requiring less computation and reducing latency. (2) The weight of the leader robot prevents the user
반면 관절 공간 매핑은 관절 한계 내에서 높은 대역폭의 제어를 보장하며, 계산량도 적고 지연 시간도 줄인다.
from moving too fast, and also dampens small vibrations.
(2) 리더 로봇의 무게는 사용자가 너무 빠르게 움직이지 못하게 하며 작은 진동도 감쇠한다.
We notice better performance on precise tasks with joint-space mapping rather than holding a VR controller. To further
VR 컨트롤러를 들고 조작하는 것보다 관절 공간 매핑을 사용할 때 정밀한 과제에서 더 나은 성능을 확인했다.
improve the teleoperation experience, we design a 3D-printed "handle and scissor" mechanism that can be retrofitted to the leader robot (Fig 3). It reduces the force required from the
원격조작 경험을 더욱 향상하기 위해 리더 로봇에 추가 장착할 수 있는 3D 프린팅 "handle and scissor" 메커니즘을 설계했다(그림 3).
operator to backdrive the motor, and allows for continuous control of the gripper, instead of binary opening or closing.
이 메커니즘은 작업자가 모터를 역구동하는 데 필요한 힘을 줄이고, 그리퍼를 이진적으로 열거나 닫는 대신 연속적으로 제어할 수 있게 한다.
We also design a rubber band load balancing mechanism that partially counteracts the gravity on the leader side. It reduces the
또한 리더 측에서 중력의 영향을 부분적으로 상쇄하는 고무 밴드 하중 균형 메커니즘을 설계했다.
effort needed from the operator and makes longer teleoperation sessions (e.g. >30 minutes) possible. We include more details
이를 통해 작업자에게 필요한 수고를 줄이고 더 긴 원격조작 세션(예: >30분)이 가능해진다.
about the setup in the project website.
설정에 관한 자세한 내용은 프로젝트 웹사이트에 제시한다.

The rest of the setup includes a robot cage with 20×20mm aluminum extrusions, reinforced by crossing steel cables. There
나머지 장비는 20×20mm 알루미늄 압출재로 구성하고 교차하는 강철 케이블로 보강한 로봇 케이지를 포함한다.
is a total of four Logitech C922x webcams, each streaming 480×640 RGB images. Two of the webcams are mounted on
총 4대의 Logitech C922x 웹캠이 있으며, 각 웹캠은 480×640 RGB 이미지를 스트리밍한다.
the wrist of the follower robots, allowing for a close-up view of the grippers. The remaining two cameras are mounted on the
웹캠 중 2대는 추종 로봇의 손목에 장착하여 그리퍼를 근접 촬영할 수 있도록 한다.
front and at the top respectively (Fig 3). Both the teleoperation
나머지 2대의 카메라는 각각 전면과 상단에 장착한다(Fig 3).
and data recording happen at 50Hz.
원격조작과 데이터 기록은 모두 50Hz로 수행한다.

With the design considerations above, we build the bimanual teleoperation setup *ALOHA* within a 20k USD budget, compa-

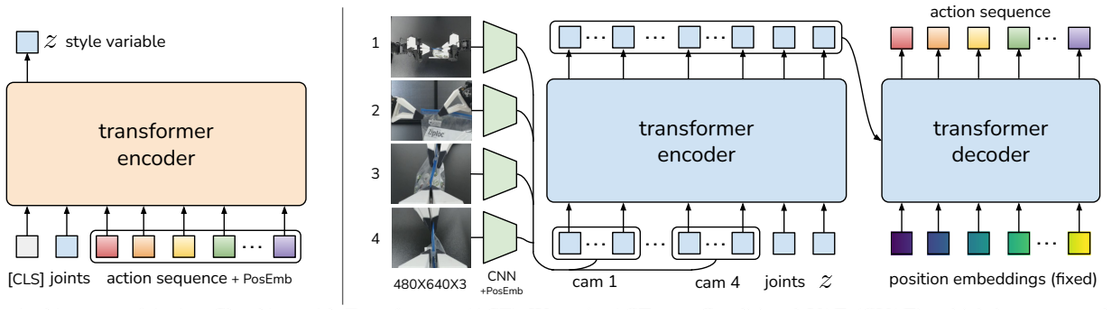


Fig. 4: *Architecture of Action Chunking with Transformers (ACT)*. We train ACT as a Conditional VAE (CVAE), which has an encoder and a decoder. *Left:* The encoder of the CVAE compresses action sequence and joint observation into z, the style variable. The encoder is discarded at test time. *Right:* The decoder or policy of ACT synthesizes images from multiple viewpoints, joint positions, and z with a transformer encoder, and predicts a sequence of actions with a transformer decoder. z is simply set to the mean of the prior (i.e. zero) at test time.

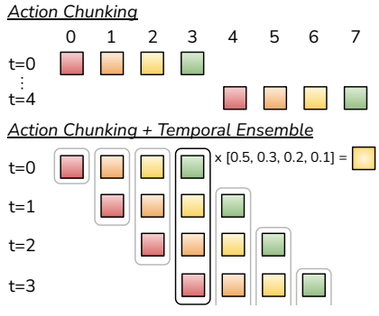


Fig. 5: We employ both Action Chunking and Temporal Ensembling when applying actions, instead of interleaving observing and executing.

rable to a single research arm such as Franka Emika Panda. *ALOHA* enables the teleoperation of:

- **Precise tasks** such as threading zip cable ties, picking credit cards out of wallets, and opening or closing ziploc bags.
- **Contact-rich tasks** such as inserting 288-pin RAM into a computer motherboard, turning pages of a book, and assembling the chains and belts in the NIST board #2 [4]

그림 4: Transformers를 활용한 행동 청킹(ACT)의 아키텍처. ACT는 조건부 VAE(CVAE)로 학습하며, 인코더와 그림 5: 다음 시간 단계에 행동 청킹과 시간적 앙상블을 모두 적용한다. 직관적으로 ACT는 인간이 위의 설계 고려사항을 바탕으로 20k USD 예산 내에서 양손 원격조작 설정 ALOHA를 구축한다. compafrom 이전 행동은 학습 범위를 벗어난 상태로 이어져 컴퓨터 마더보드를 조립하고, 책장을 넘기며, 분포. NIST 보드 #2의 체인과 벨트를 조립하는 작업 [4]

- **Dynamic tasks** such as juggling a ping pong ball with a real ping pong paddle, balancing the ball without it falling off, and swinging open plastic bags in the air.

Skills such as threading a zip tie, inserting RAM, and juggling ping pong ball, to our knowledge, are not available for existing teleoperation systems with 5-10x the budget [21, 5]. We include a more detailed price & capability comparison in Appendix A, as well as more skills that *ALOHA* is capable of in Figure 9. To make *ALOHA* more accessible, we open-source all software and hardware with a detailed tutorial covering 3D printing, assembling the frame to software installations. You can find the tutorial on the project website.

실제 탁구 라켓으로 탁구공을 저글링하고, 공이 떨어지지 않도록 균형을 잡으며, 공중에서 비닐봉지를 흔들어 여는 작업과 같은 동적인 과제를 수행한다. 케이블 타이를 끼우고, RAM을 삽입하며, 탁구공을 저글링하는 기술은 우리가 아는 한 5-10배의 예산이 드는 기존 원격조작 시스템에서도 제공되지 않는다 [21, 5]. 부록 A에는 가격 및 성능 비교를 더 자세히 제시하며, ALOHA가 수행할 수 있는 추가 기술은 그림 9에도 제시한다. ALOHA를 더 쉽게 이용할 수 있도록 3D 프린팅부터 프레임 조립과 소프트웨어 설치까지 다루는 상세한 튜토리얼과 함께 모든 소프트웨어 및 하드웨어를 오픈 소스로 공개한다. 튜토리얼은 프로젝트 웹사이트에서 확인할 수 있다.

IV. Action Chunking with Transformers

IV. TRANSFORMERS를 활용한 행동 청킹

As we will see in Section V, existing imitation learning algorithms perform poorly on fine-grained tasks that require high-frequency control and closed-loop feedback. We therefore develop a novel algorithm, *Action Chunking with Transformers (ACT)*, to leverage the data collected by *ALOHA*. We first summarize the pipeline of training ACT, then dive into each of the design choices.

5절에서 살펴보겠지만, 기존 모방학습 알고리즘은 고주파 제어와 폐루프 피드백이 필요한 정밀한 과제에서 성능이 좋지 않다. 따라서 ALOHA가 수집한 데이터를 활용하기 위해 새로운 알고리즘인 Transformers를 활용한 행동 청킹(ACT)을 개발한다. 먼저 ACT 학습 파이프라인을 요약한 다음, 각 설계 선택을 자세히 살펴본다.

To train ACT on a new task, we first collect human demonstrations using *ALOHA*. We record the joint positions of the leader robots (i.e. input from the human operator) and use them as actions. It is important to use the leader joint positions instead of the follower's, because the amount of force applied is implicitly defined by the difference between them, through the low-level PID controller. The observations are composed of the current joint positions of follower robots and the image feed from 4 cameras. Next, we train ACT to predict the *sequence of future actions* given the current observations. An action here corresponds to the target joint positions for both arms in the next time step. Intuitively, ACT tries to imitate what a human operator would do in the following time steps given current observations. These target joint positions are then tracked by the low-level, high-frequency PID controller inside Dynamixel motors. At test time, we load the policy that achieves the lowest validation loss and roll it out in the environment. The main challenge that arises is compounding errors, where errors from previous actions lead to states that are outside of training distribution.

A. Action Chunking and Temporal Ensemble

A. 행동 청킹 및 시간적 앙상블

To combat the compounding errors of imitation learning in a way that is compatible with pixel-to-action policies (Figure II), we seek to reduce the effective horizon of long trajectories collected at high frequency. We are inspired by *action chunking*, a neuroscience concept where individual actions are grouped together and executed as one unit, making them more efficient to store and execute [35]. Intuitively, a chunk of actions could correspond to grasping a corner of the candy wrapper or inserting a battery into the slot. In our implementation, we fix the chunk size to be k: every k steps, the agent receives an observation, generates the next k actions, and executes the actions in sequence (Figure 5). This implies a k-fold reduction in the effective horizon of the task. Concretely, the policy models $\pi_\theta(a_{t:t+k}|s_t)$ instead of $\pi_\theta(a_t|s_t)$. Chunking can also help model non-Markovian behavior in human demonstrations. Specifically, a single-step policy would struggle with temporally correlated confounders, such as pauses in the middle of a demonstration [61], since the behavior not only depends on the state, but also the timestep. Action chunking can mitigate

고주파로 수집된 긴 궤적의 유효 시간 범위를 줄여, 픽셀-행동 정책(그림 II)과 호환되는 방식으로 모방학습의 누적 오차에 대응하고자 한다. 개별 행동을 함께 묶어 하나의 단위로 실행함으로써 저장과 실행을 더 효율적으로 만드는 신경과학 개념인 행동 청킹에서 영감을 얻었다 [35]. 직관적으로 행동 청크는 사탕 포장지의 모서리를 잡거나 슬롯에 배터리를 삽입하는 동작에 해당할 수 있다. 구현에서는 청크 크기를 k로 고정한다. k스텝마다 에이전트는 관측을 받고, 다음 k개의 행동을 생성한 뒤 행동을 순서대로 실행한다(그림 5). 이는 과제의 유효 시간 범위를 k배 줄이는 효과가 있다. 구체적으로 정책은 πθ(at:t+k|st)가 아니라 πθ(at|st)를 모델링한다. 청킹은 인간 시연에서 나타나는 비마르코프적 행동을 모델링하는 데에도 도움이 될 수 있다. 구체적으로 단일 스텝 정책은 시연 중간의 멈춤과 같이 시간적으로 상관된 교란 요인을 처리하는 데 어려움을 겪을 수 있다 [61]. 이는 행동이 상태뿐만 아니라 시간 단계에도 의존하기 때문이다. 행동 청킹은 완화할 수 있다

Algorithm 1 ACT Training

알고리즘 1 ACT 학습

Given: Demo dataset $\mathcal{D}$, chunk size k, weight β.

1: 주어짐: 시연 데이터셋 D, 청크 크기 k, 가중치 β.

Let a_t, o_t represent action and observation at timestep t, $\bar{o}_t$ represent o_t without image observations.

2: at, ot는 시점 t의 행동과 관측을 나타내며, ¯ot는 이미지 관측을 제외한 ot를 나타낸다.

Initialize encoder $q_\phi(z|a_{t:t+k}, \bar{o}_t)$

Initialize decoder $\pi_\theta(\hat{a}_{t:t+k}|o_t, z)$

for iteration $n = 1, 2, ...$ **do**

  Sample o_t, $a_{t:t+k}$ from $\mathcal{D}$

  Sample z from $q_\phi(z|a_{t:t+k}, \bar{o}_t)$

  Predict $\hat{a}_{t:t+k}$ from $\pi_\theta(\hat{a}_{t:t+k}|o_t, z)$

3: 인코더 qφ(z|at:t+k, ¯ot)를 초기화한다. 4: 디코더 πθ(ˆat:t+k|ot, z)를 초기화한다. 5: 반복 횟수 n = 1, 2, ...에 대해 반복한다. 6: D에서 ot, at:t+k를 샘플링한다. 7: qφ(z|at:t+k, ¯ot)에서 z를 샘플링한다. 8: πθ(ˆat:t+k|ot, z)에서 ˆat:t+k를 예측한다.

  $\mathcal{L}_{reconst} = MSE(\hat{a}_{t:t+k}, a_{t:t+k})$

  $\mathcal{L}_{reg} = D_{KL}(q_\phi(z|a_{t:t+k}, \bar{o}_t) \parallel \mathcal{N}(0, I))$

  Update θ, ϕ with ADAM and $\mathcal{L} = \mathcal{L}_{reconst} + \beta\mathcal{L}_{reg}$

11: ADAM을 사용하여 θ, φ을 업데이트하고, L = Lreconst + βLreg로 설정한다.

Algorithm 2 ACT Inference

알고리즘 2 ACT 추론

Given: trained π_θ, episode length T, weight m.

1: 주어짐: 학습된 πθ, 에피소드 길이 T, 가중치 m.

Initialize FIFO buffers $\mathcal{B}[0 : T]$, where $\mathcal{B}[t]$ stores actions predicted *for* timestep t.

2: FIFO 버퍼 B[0 : T]를 초기화하며, B[t]에는 행동을 저장한다. 시점 t에 대해 예측한다.

for timestep $t = 1, 2, ...T$ **do**

  Predict $\hat{a}_{t:t+k}$ with $\pi_\theta(\hat{a}_{t:t+k}|o_t, z)$ where $z = 0$

  Add $\hat{a}_{t:t+k}$ to buffers $\mathcal{B}[t : t + k]$ respectively

  Obtain current step actions $A_t = \mathcal{B}[t]$

  Apply $a_t = \sum_i w_i A_t[i] / \sum_i w_i$, with $w_i = \exp(-m * i)$

3: 시점 t = 1, 2, ...T에 대해 반복한다. 4: πθ(ˆat:t+k|ot, z)를 사용하여 ˆat:t+k를 예측하며, 이때 z = 0이다. 5: ˆat:t+k를 각각 버퍼 B[t : t + k]에 추가한다. 6: 현재 단계의 행동 At = B[t]를 얻는다. 7: at = P i wi를 적용하며, wi = exp(−m ∗i)이다.

this issue when the confounder is within a chunk, without introducing the causal confusion issue for history-conditioned policies [12].

혼란 변수가 청크 내에 있을 때 이 문제를 해결하면서도, 이력 조건부 정책에서 발생하는 인과적 혼동 문제는 도입하지 않는다 [12].

A naïve implementation of action chunking can be sub-optimal: a new environment observation is incorporated abruptly every k steps and can result in jerky robot motion.

행동 청킹을 순진하게 구현하면 최적이 아닐 수 있다. 새로운 환경 관측이 k 스텝마다 갑작스럽게 반영되어 로봇 동작이 급격해질 수 있다.

To improve smoothness and avoid discrete switching between executing and observing, we query the policy at every timestep.

부드러움을 향상하고 실행과 관측 사이의 불연속적인 전환을 방지하기 위해 매 시점마다 정책에 질의한다.

This makes different action chunks overlap with each other, and at a given timestep there will be more than one predicted action. We illustrate this in Figure 5 and propose a *temporal ensemble* to combine these predictions. Our *temporal ensemble* performs a weighted average over these predictions with an exponential weighting scheme $w_i = \exp(-m * i)$, where w_0 is the weight for the oldest action. The speed for incorporating new observation is governed by m, where a smaller m means faster incorporation. We note that unlike typical smoothing, where the current action is aggregated with actions in adjacent timesteps, which leads to bias, we aggregate actions predicted for the *same* timestep. This procedure also incurs no additional training cost, only extra inference-time computation. In practice, we find both action chunking and temporal ensembling to be important for the success of ACT, which produces precise and smooth motion. We discuss these components in more detail in the ablation studies in Subsection VI-A.

이로 인해 서로 다른 행동 청크가 서로 겹치며, 특정 시점에는 예측된 행동이 1개보다 많아진다. 이를 Figure 5에 설명하고, 이러한 예측을 결합하기 위해 시간적 앙상블을 제안한다. 우리의 시간적 앙상블은 지수 가중 방식 wi = exp(−m ∗i)을 사용하여 이러한 예측의 가중 평균을 계산하며, 여기서 w0는 가장 오래된 행동의 가중치이다. 새로운 관측을 반영하는 속도는 m에 의해 결정되며, m이 작을수록 더 빠르게 반영된다. 이 절차에는 추가적인 학습 비용이 들지 않으며, 추론 시간의 계산만 추가로 필요하다. 실제로 정밀하고 부드러운 동작을 생성하는 ACT의 성공에는 행동 청킹과 시간적 앙상블이 모두 중요하다는 사실을 확인했다. Subsection VI-A의 절제 연구에서 이러한 구성 요소를 더 자세히 논의한다.

B. Modeling human data

B. 인간 데이터 모델링

Another challenge that arises is learning from noisy human demonstrations. Given the same observation, a human can use different trajectories to solve the task. Humans will also be more stochastic in regions where precision matters less [38]. Thus, it is important for the policy to focus on regions where high precision matters. We tackle this problem by training our action chunking policy as a generative model. Specifically, we train the policy as a conditional variational autoencoder (CVAE) [55], to generate an action sequence conditioned on current observations. The CVAE has two components: a CVAE encoder and a CVAE decoder, illustrated on the left and right side of Figure 4 respectively. The CVAE encoder only serves to train the CVAE decoder (the policy) and is discarded at test time. Specifically, the CVAE encoder predicts the mean and variance of the style variable z's distribution, which is parameterized as a diagonal Gaussian, given the current observation and action sequence as inputs. For faster training in practice, we leave out the image observations and only condition on the proprioceptive observation and the action sequence. The CVAE decoder, i.e. the policy, conditions on both z and the current observations (images + joint positions) to predict the action sequence. At test time, we set z to be the mean of the prior distribution i.e. zero to deterministically decode. The whole model is trained to maximize the log-likelihood of demonstration action chunks, i.e. $\min_\theta -\sum_{s_t, a_{t:t+k} \in D} \log \pi_\theta(a_{t:t+k}|s_t)$, with the standard VAE objective which has two terms: a reconstruction loss and a term that regularizes the encoder to a Gaussian prior. Following [23], we weight the second term with a hyperparameter β. Intuitively, higher β will result in less information transmitted in z [62]. Overall, we found the CVAE objective to be essential in learning precise tasks from human demonstrations. We include a more detailed discussion in Subsection VI-B.

action chunking policy as a generative model. Specifically, we 이 문제를 해결하기 위해 행동 청킹 정책을 생성 모델로 학습한다. train the policy as a conditional variational autoencoder (CVAE) [55], to generate an action sequence conditioned on current observations. The CVAE has two components: a CVAE encoder 구체적으로 정책을 조건부 변분 오토인코더(CVAE) [55]로 학습하여 현재 관측을 조건으로 하는 행동 시퀀스를 생성하도록 한다. and a CVAE decoder, illustrated on the left and right side of Figure 4 respectively. The CVAE encoder only serves to train CVAE는 CVAE 인코더와 CVAE 디코더라는 2개의 구성 요소로 이루어지며, 각각 Figure 4의 왼쪽과 오른쪽에 나타나 있다. the CVAE decoder (the policy) and is discarded at test time. CVAE 인코더는 CVAE 디코더(정책)를 학습하는 역할만 수행하며 테스트 시에는 제거된다. Specifically, the CVAE encoder predicts the mean and variance of the style variable z's distribution, which is parameterized as a diagonal Gaussian, given the current observation and action sequence as inputs. For faster training in practice, we leave out 구체적으로 CVAE 인코더는 현재 관측과 행동 시퀀스를 입력으로 받아 대각 가우시안으로 매개변수화된 스타일 변수 z의 분포에 대한 평균과 분산을 예측한다. the image observations and only condition on the proprioceptive observation and the action sequence. The CVAE decoder, i.e. 실제로 더 빠른 학습을 위해 이미지 관측은 제외하고 고유수용성 관측과 행동 시퀀스만 조건으로 사용한다. the policy, conditions on both z and the current observations (images + joint positions) to predict the action sequence. At CVAE 디코더, 즉 정책은 z와 현재 관측(이미지 + 관절 위치)을 모두 조건으로 사용하여 행동 시퀀스를 예측한다. test time, we set z to be the mean of the prior distribution i.e. zero to deterministically decode. The whole model is trained to 테스트 시에는 z를 사전 분포의 평균, 즉 0으로 설정하여 결정론적으로 디코딩한다. maximize the log-likelihood of demonstration action chunks, i.e. $\min_\theta -\sum_{s_t, a_{t:t+k} \in D} \log \pi_\theta(a_{t:t+k}|s_t)$, with the standard VAE objective which has two terms: a reconstruction loss and a term that regularizes the encoder to a Gaussian prior. Following [23], 전체 모델은 시연 행동 청크의 로그 가능도를 최대화하도록, 즉 minθ −P st,at:t+k∈D log πθ(at:t+k|st)로 학습하며, 표준 VAE 목적 함수는 2개의 항으로 구성된다. a 재구성 손실과 인코더를 a 가우시안 사전 분포로 정규화하는 a 항이다. we weight the second term with a hyperparameter β. Intuitively, [23]에 따라 두 번째 항에 하이퍼파라미터 β의 가중치를 부여한다. higher β will result in less information transmitted in z [62]. 직관적으로 β이 높을수록 z로 전달되는 정보가 줄어든다 [62]. Overall, we found the CVAE objective to be essential in learning precise tasks from human demonstrations. We include 전반적으로 CVAE 목적 함수가 인간 시연으로부터 정밀한 과제를 학습하는 데 필수적이라는 사실을 확인했다. a more detailed discussion in Subsection VI-B.

Subsection VI-B에서 더 자세히 논의한다.

C. Implementing ACT

C. ACT 구현

We implement the CVAE encoder and decoder with transformers, as transformers are designed for both synthesizing information across a sequence and generating new sequences. The CVAE encoder is implemented with a BERT-like transformer encoder [13]. The inputs to the encoder are the current joint positions and the target action sequence of length k from the demonstration dataset, prepended by a learned "[CLS]" token similar to BERT. This forms a $k+2$ length input (Figure 4 left). After passing through the transformer, the feature corresponding to "[CLS]" is used to predict the mean and variance of the "style variable" z, which is then used as input to the decoder. The CVAE decoder (i.e. the policy) takes the current observations and z as the input, and predicts the next k actions (Figure 4 right). We use ResNet image encoders, a transformer encoder, and a transformer decoder to implement the CVAE decoder. Intuitively, the transformer encoder synthesizes information from different camera viewpoints, the joint positions, and the style variable, and the transformer decoder generates a coherent action sequence. The observation includes 4 RGB images, each at 480×640 resolution, and joint positions for two robot arms (7+7=14 DoF in total). The action space is the absolute joint positions for two robots, a 14-dimensional vector. Thus with action chunking, the policy outputs a $k \times 14$ tensor given the current observation. The policy first process the images with ResNet18 backbones [22], which convert $480 \times 640 \times 3$ RGB images into $15 \times 20 \times 512$ feature maps. We then flatten along the spatial dimension to obtain a sequence of 300×512. To preserve the spatial information, we add a 2D sinusoidal

We implement the CVAE encoder and decoder with transformers, as transformers are designed for both synthesizing information across a sequence and generating new sequences. The 트랜스포머는 시퀀스 전반의 정보를 통합하고 새로운 시퀀스를 생성하도록 설계되었으므로, 트랜스포머를 사용하여 CVAE 인코더와 디코더를 구현한다. CVAE encoder is implemented with a BERT-like transformer encoder [13]. The inputs to the encoder are the current joint CVAE 인코더는 BERT와 유사한 트랜스포머 인코더 [13]로 구현한다. positions and the target action sequence of length k from the demonstration dataset, prepended by a learned "[CLS]" token similar to BERT. This forms a $k+2$ length input (Figure 4 left). 인코더의 입력은 현재 관절 위치와 시연 데이터셋에서 얻은 길이 k의 목표 행동 시퀀스이며, BERT와 유사하게 학습된 "[CLS]" 토큰을 앞에 추가한다. 이로써 길이가 k+2인 입력이 구성된다(그림 4 왼쪽). After passing through the transformer, the feature corresponding to "[CLS]" is used to predict the mean and variance of the "style variable" z, which is then used as input to the decoder. The 트랜스포머를 통과한 후 "[CLS]"에 해당하는 특징을 사용하여 스타일 변수 z의 평균과 분산을 예측하고, 이를 디코더의 입력으로 사용한다. CVAE decoder (i.e. the policy) takes the current observations and z as the input, and predicts the next k actions (Figure 4 right). We use ResNet image encoders, a transformer encoder, CVAE 디코더(즉, 정책)는 현재 관측과 z를 입력으로 받아 다음 k개의 행동을 예측한다(그림 4 오른쪽). and a transformer decoder to implement the CVAE decoder. CVAE 디코더를 구현하기 위해 ResNet 이미지 인코더, 트랜스포머 인코더, 트랜스포머 디코더를 사용한다. Intuitively, the transformer encoder synthesizes information from different camera viewpoints, the joint positions, and the style variable, and the transformer decoder generates a coherent action sequence. The observation includes 4 RGB images, each 직관적으로 트랜스포머 인코더는 서로 다른 카메라 시점의 정보, 관절 위치, 스타일 변수의 정보를 통합하고, 트랜스포머 디코더는 일관된 행동 시퀀스를 생성한다. at 480×640 resolution, and joint positions for two robot arms (7+7=14 DoF in total). The action space is the absolute joint 관측에는 각각 480 × 640 해상도인 4개의 RGB 이미지와 2개 로봇 팔의 관절 위치가 포함된다(총 7+7=14 DoF). positions for two robots, a 14-dimensional vector. Thus with 행동 공간은 2개 로봇의 절대 관절 위치이며, 14차원 벡터이다. action chunking, the policy outputs a $k \times 14$ tensor given the current observation. The policy first process the images 따라서 행동 청킹을 사용하면 정책은 현재 관측을 바탕으로 k × 14 텐서를 출력한다. with ResNet18 backbones [22], which convert $480 \times 640 \times 3$ RGB images into $15 \times 20 \times 512$ feature maps. We then flatten 정책은 먼저 ResNet18 백본 [22]을 사용하여 이미지를 처리하며, 이 백본은 480 × 640 × 3 RGB 이미지를 15 × 20 × 512 특징 맵으로 변환한다. along the spatial dimension to obtain a sequence of 300×512. 그런 다음 공간 차원을 따라 평탄화하여 300 × 512의 시퀀스를 얻는다. To preserve the spatial information, we add a 2D sinusoidal

position embedding to the feature sequence [8]. Repeating this
공간 정보를 보존하기 위해 특징 시퀀스에 2D 사인파 위치 임베딩을 추가한다 [8].
for all 4 images gives a feature sequence of 1200×512 in dimension. We then append two more features: the current
이를 4개 이미지 모두에 반복 적용하면 차원이 1200 × 512인 특징 시퀀스를 얻는다.
joint positions and the "style variable" z. They are projected
그런 다음 현재 관절 위치와 "스타일 변수" z라는 2개의 특징을 추가한다.
from their original dimensions to 512 through linear layers respectively. Thus, the input to the transformer encoder is
각각을 선형 레이어를 통해 원래 차원에서 512차원으로 투영한다.
1202×512. The transformer decoder conditions on the encoder
따라서 트랜스포머 인코더의 입력은 1202×512이다.
output through cross-attention, where the input sequence is a fixed position embedding, with dimensions $k \times 512$, and the keys and values are coming from the encoder. This gives the
트랜스포머 디코더는 교차 어텐션을 통해 인코더 출력에 조건화된다. 이때 입력 시퀀스는 차원이 k × 512인 고정 위치 임베딩이며, 키와 값은 인코더에서 가져온다.
transformer decoder an output dimension of $k \times 512$, which is then down-projected with an MLP into $k \times 14$, corresponding to the predicted target joint positions for the next k steps. We
이에 따라 트랜스포머 디코더의 출력 차원은 k × 512가 되며, 이를 MLP를 통해 k × 14로 차원 축소한다. 이는 다음 k개 단계에 대해 예측된 목표 관절 위치에 해당한다.
use L1 loss for reconstruction instead of the more common L2 loss: we noted that L1 loss leads to more precise modeling of the action sequence. We also noted degraded performance
재구성에는 더 일반적인 L2 손실 대신 L1 손실을 사용한다. L1 손실이 행동 시퀀스를 더 정밀하게 모델링한다는 점을 확인했기 때문이다.
when using delta joint positions as actions instead of target joint positions. We include a detailed architecture diagram in
또한 목표 관절 위치 대신 델타 관절 위치를 행동으로 사용할 경우 성능이 저하되는 것을 확인했다.
Appendix C.
부록 C에 상세한 아키텍처 다이어그램을 제시한다.

We summarize the training and inference of ACT in Algorithms 1 and 2. The model has around 80M parameters,
알고리즘 1과 2에서 ACT의 학습 및 추론을 요약한다.
and we train from scratch for each task. The training takes
모델은 약 80M개의 파라미터를 가지며, 각 과제마다 처음부터 학습한다.
around 5 hours on a single 11G RTX 2080 Ti GPU, and the inference time is around 0.01 seconds on the same machine.
학습에는 단일 11G RTX 2080 Ti GPU에서 약 5시간이 걸리며, 동일한 장치에서 추론 시간은 약 0.01초이다.

V. Experiments

V. 실험

We present experiments to evaluate ACT's performance on fine manipulation tasks. For ease of reproducibility, we build
정밀 조작 과제에서 ACT의 성능을 평가하기 위한 실험을 제시한다.
two simulated fine manipulation tasks in MuJoCo [63], in addition to 6 real-world tasks with *ALOHA*. We provide videos
재현을 용이하게 하기 위해 MuJoCo [63]에서 2개의 시뮬레이션 정밀 조작 과제를 구축하고, ALOHA를 사용하여 6개의 실제 환경 과제를 추가로 구성한다.
for each task on the project website.
각 과제의 동영상은 프로젝트 웹사이트에서 제공한다.

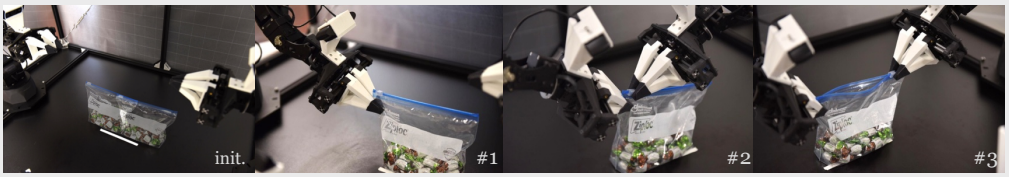
A. Tasks

A. 과제

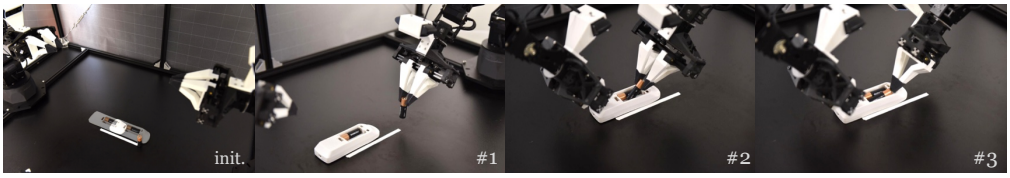
All 8 tasks require fine-grained, bimanual manipulation, and are illustrated in Figure 6. For ***Slide Ziploc***, the right gripper
8개 과제 모두 정밀한 양손 조작을 필요로 하며, 그림
needs to accurately grasp the slider of the ziploc bag and open it, with the left gripper securing the body of the bag. For ***Slot***
6에 제시되어 있다. Slide Ziploc에서는 오른쪽 그리퍼가 지퍼백의 슬라이더를 정확하게 잡아 열고, 왼쪽 그리퍼는 봉투 본체를 고정해야 한다.
Battery, the right gripper needs to first place the battery into the slot of the remote controller, then using the tip of fingers to delicately push in the edge of the battery, until it is fully inserted. Because the spring inside the battery slot causes the
Slot Battery에서는 오른쪽 그리퍼가 먼저 배터리를 리모컨의 슬롯에 넣은 다음, 손가락 끝으로 배터리 가장자리를 조심스럽게 밀어 완전히 삽입해야 한다.
remote controller to move in the opposite direction during insertion, the left gripper pushes down on the remote to keep it in place. For ***Open Cup***, the goal is to open the lid of a
배터리 슬롯 내부의 스프링 때문에 삽입 중 리모컨이 반대 방향으로 움직이므로, 왼쪽 그리퍼가 리모컨을 아래로 눌러 제자리에 고정한다.
small condiment cup. Because of the cup's small size, the
Open Cup의 목표는 작은 양념 컵의 뚜껑을 여는 것이다.
grippers cannot grasp the body of the cup by just approaching it from the side. Therefore we leverage both grippers: the right
컵의 크기가 작기 때문에 그리퍼는 옆에서 접근하는 것만으로 컵 본체를 잡을 수 없다.
fingers first lightly tap near the edge of the cup to tip it over, and then nudge it into the open left gripper. This nudging
따라서 두 그리퍼를 모두 활용한다. 오른쪽 손가락으로 먼저 컵 가장자리 근처를 가볍게 두드려 컵을 기울인 다음, 열린 왼쪽 그리퍼 쪽으로 살짝 민다.
step requires high precision and closing the loop on visual perception. The left gripper then closes gently and lifts the cup
이러한 미세 이동 단계에는 높은 정밀도와 시각 인식에 대한 폐루프 제어가 필요하다.
off the table, followed by the right finger prying open the lid, which also requires precision to not miss the lid or damage the cup. The goal of ***Thread Velcro*** is to insert one end of
그런 다음 왼쪽 그리퍼가 부드럽게 닫히면서 컵을 테이블에서 들어 올리고, 오른쪽 손가락이 뚜껑을 벌려 연다. 이 과정에서도 뚜껑을 놓치거나 컵을 손상하지 않으려면 정밀성이 필요하다.
a velcro cable tie into the small loop attached to other end.
Thread Velcro의 목표는 벨크로 케이블 타이의 한쪽 끝을 다른 쪽 끝에 부착된 작은 고리에 삽입하는 것이다.
The left gripper needs to first pick up the velcro tie from the table, followed by the right gripper pinching the tail of the tie in mid-air. Then, both arms coordinate to insert one end
왼쪽 그리퍼가 먼저 테이블에서 벨크로 타이를 집어 들고, 이어서 오른쪽 그리퍼가 공중에서 타이의 꼬리 부분을 집는다.
of the velcro tie into the other in mid-air. The loop measures
그런 다음 두 팔이 협응하여 공중에서 벨크로 타이의 한쪽 끝을 다른 쪽 끝에 삽입한다.
3mm x 25mm, while the velcro tie measures 2mm x 10-25mm depending on the position. For this task to be successful, the
고리의 크기는 3mm x 25mm이고, 벨크로 타이의 크기는 위치에 따라 2mm x 10-25mm이다.
robot must use visual feedback to correct for perturbations with each grasp, as even a few millimeters of error during the first grasp will compound in the second grasp mid-air, giving more than a 10mm deviation in the insertion phase. For ***Prep***
이 과제를 성공적으로 수행하려면 로봇이 각 그립에서 발생하는 섭동을 보정하기 위해 시각 피드백을 사용해야 한다. 첫 번째 그립에서 몇 밀리미터만 오차가 발생해도 공중에서 수행하는 두 번째 그립에서 오차가 누적되어 삽입 단계에서 10mm를 초과하는 편차가 발생하기 때문이다.
Tape, the goal is to hang a small segment of the tape on the edge of a cardboard box. The right gripper first grasps the tape
Prep Tape의 목표는 작은 테이프 조각을 판지 상자의 가장자리에 걸치는 것이다.
and cuts it with the tape dispenser's blade, and then hands the tape segment to the left gripper mid-air. Next, both arms
오른쪽 그리퍼가 먼저 테이프를 잡고 테이프 디스펜서의 날로 자른 다음, 공중에서 테이프 조각을 왼쪽 그리퍼에 건넨다.
approach the box, the left arm gently lays the tape segment on the box surface, and the right fingers push down on the tape to prevent slipping, followed by the left arm opening its gripper to release the tape. Similar to *Thread Velcro*, this task
다음으로 두 팔이 상자에 접근한다. 왼쪽 팔이 테이프 조각을 상자 표면에 부드럽게 올려놓고, 오른쪽 손가락이 테이프가 미끄러지지 않도록 누른 다음, 왼쪽 팔이 그리퍼를 열어 테이프를 놓는다.
requires multiple steps of delicate coordination between the two arms. For ***Put On Shoe***, the goal is to put the shoe on a
Thread Velcro와 마찬가지로 이 과제에서도 두 팔 사이의 섬세한 다단계 협응이 필요하다.
fixed manniquin foot, and secure it with the shoe's velcro strap.
Put On Shoe의 목표는 고정된 마네킹 발에 신발을 신기고 신발의 벨크로 스트랩으로 고정하는 것이다.
The arms would first need to grasp the tongue and collar of the shoe respectively, lift it up and approach the foot. Putting
두 팔은 먼저 각각 신발의 설포와 칼라를 잡고 들어 올린 다음 발에 접근해야 한다.
the shoe on is challenging because of the tight fitting: the arms would need to coordinate carefully to nudge the foot in, and both grasps need to be robust enough to counteract the friction between the sock and shoe. Then, the left arm goes around to
신발을 신기는 일은 꽉 맞는 구조 때문에 어렵다. 두 팔이 발을 안쪽으로 밀어 넣기 위해 세심하게 협응해야 하며, 양쪽 그립 모두 양말과 신발 사이의 마찰에 대응할 수 있을 만큼 견고해야 한다.
the bottom of the shoe to support it from dropping, followed by the right arm flipping the velcro strap and pressing it against the shoe to secure. The task is only considered successful if
그런 다음 왼쪽 팔이 신발 아래쪽으로 돌아가 신발이 떨어지지 않도록 지지하고, 오른쪽 팔이 벨크로 스트랩을 뒤집어 신발에 눌러 고정한다.
the shoe clings to the foot after both arms releases. For the
두 팔을 모두 놓은 후에도 신발이 발에 붙어 있어야만 해당 과제가 성공한 것으로 간주한다.
simulated task ***Transfer Cube***, the right arm needs to first pick up the red cube lying on the table, then place it inside the gripper of the other arm. Due to the small clearance between
시뮬레이션 과제인 Transfer Cube에서는 먼저 오른팔이 테이블 위에 놓인 빨간색 큐브를 집은 다음, 다른 팔의 그리퍼 안에 넣어야 한다.
the cube and the left gripper (around 1cm), small errors could result in collisions and task failure. For the simulated task
큐브와 왼쪽 그리퍼 사이의 간격이 약 1cm로 좁기 때문에 작은 오차만으로도 충돌이 발생하여 과제가 실패할 수 있다.
Bimanual Insertion, the left and right arms need to pick up the socket and peg respectively, and then insert in mid-air so the peg touches the "pins" inside the socket. The clearance is
시뮬레이션 과제인 Bimanual Insertion에서는 왼팔과 오른팔이 각각 소켓과 핀을 집은 다음, 공중에서 삽입하여 핀이 소켓 내부의 "핀"에 닿게 해야 한다.
around 5mm in the insertion phase. For all 8 tasks, the initial
삽입 단계에서의 간격은 약 5mm이다.
placement of the objects is either varied randomly along the 15cm white reference line (real-world tasks), or uniformly in 2D regions (simulated tasks). We provide illustrations of both
8개 과제 모두에서 물체의 초기 배치는 15cm 길이의 흰색 기준선을 따라 무작위로 변화하거나(실세계 과제), 시뮬레이션 과제의 경우 2D 영역 내에서 균일하게 배치된다.
the initial positions and the subtasks in Figure 6 and 7. Our
그림 6과
evaluation will additionally report the performance for each of these subtasks.
7에 초기 위치와 하위 과제를 모두 도시한다. 또한 평가에서는 각 하위 과제의 성능도 보고한다.

In addition to the delicate bimanual control required to solve these tasks, the objects we use also present a significant perception challenge. For example, the ziploc bag is largely
이러한 과제를 해결하려면 정교한 양손 제어가 필요할 뿐 아니라, 사용되는 물체도 상당한 인식 문제를 제기한다.
transparent, with a thin blue sealing line. Both the wrinkles
예를 들어 지퍼백은 대부분 투명하고 얇은 파란색 밀봉선이 있다.
on the bag and the reflective candy wrappers inside can vary during the randomization, and distract the perception system.
무작위화 과정에서 봉지의 주름과 내부에 있는 반사성 사탕 포장지가 모두 변할 수 있으며, 인식 시스템을 방해한다.
Other transparent or translucent objects include the tape and both the lid and body of the condiment cup, making them hard to perceive precisely and ill-suited for depth cameras. The
테이프와 양념 컵의 뚜껑 및 본체도 투명하거나 반투명한 물체이므로 정밀하게 인식하기 어렵고 깊이 카메라에 적합하지 않다.
black table top also creates a low-contrast against many objects of interest, such as the black velcro cable tie and the black

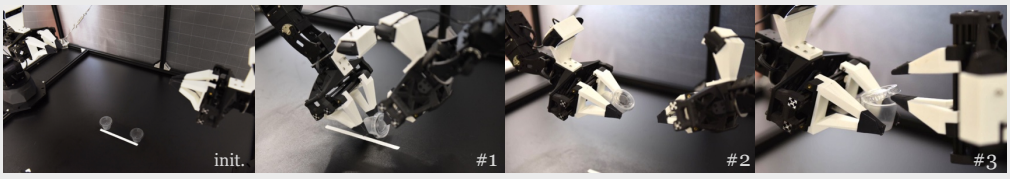
Real-World Task Definitions



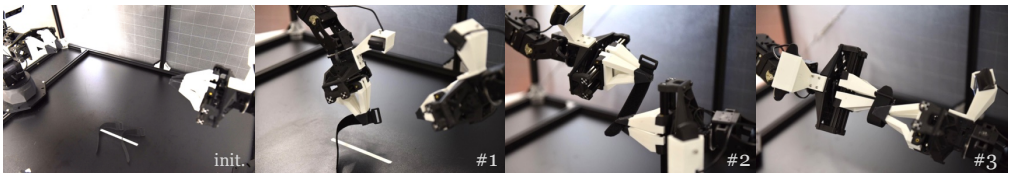
Slide Ziploc: Open the ziploc bag that is standing upright on the table. The bag is randomized along the 15cm white line. It is dropped from ~5cm above the table to randomize the deformation, which affects the height and appearance of the bag. The left arm first grasps the bag body *(Subtask#1 Grasp)* followed by the right arm pinching the slider *(Subtask #2 Pinch)*. Then the right arm moves right to unzip the bag *(Subtask #3 Open)*.



Slot Battery: Insert the battery into the remote controller. The controller is randomized along the 15cm white line. The battery is initialized in roughly the same position with different rotations. The right arm first grasps the battery *(Subtask#1 Grasp)* then places it into the slot *(Subtask#2 Place)*. The left arm presses onto the remote to prevent it from sliding, while the right arm pushes in the battery *(Subtask#3 Insert)*.



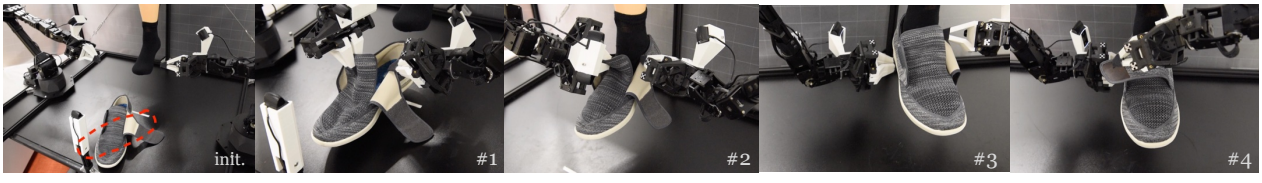
Open Cup: Pick up and open the lid of a translucent condiment cup. The cup is randomized along the 15cm white line. Both arms approach the cup, and the right gripper gently tips over the cup *(Subtask#1 Tip Over)* and pushes it into the gripper of the left arm. The left arm then gently closes its gripper and lifts the cup off the table *(Subtask#2 Grasp)*. Next, the right gripper approaches the cup lid from below and prys open the lid.



Thread Velcro: Pick up the velcro cable tie and insert one end into the small loop on the other end. The velcro tie is randomized along the 15cm white line. The left arm first picks up the velcro tie by pinching near the plastic loop *(Subtask#1 Lift)*. The right arm grasps the tail of the velcro tie mid-air *(Subtask#2 Grasp)*. Next, both arms coordinate to deform the velcro tie and insert one end of it into the plastic loop on the other end.

Prep Tape: Hang a short segment of tape on the edge of the box. The tape dispenser is randomized along the 15cm white line. First, the right gripper grasps the tape from the side *(Subtask#1 Grasp)*. It then lifts the tape and pulls to unroll it, followed by cutting it with the dispenser blade *(Subtask#2 Cut)*. Next, the right gripper hands the tape segment to the left gripper in mid-air *(Subtask#3 Handover)*, and both arms move toward the corner of the stationery cardboard box. The left arm then lays the tape segment flat on the surface of the box while the right gripper pushes down on the tape to prevent slipping. The left arm then opens its gripper to release the tape *(Subtask#4 Hang)*.

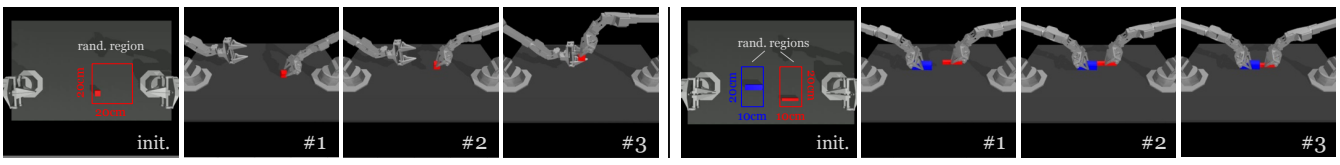
검은색 테이블 상판은 검은색 벨크로 케이블 타이와 미끄럼을 방지하기 위한 검은색 테이프처럼 관심 대상인 여러 물체와 낮은 대비를 형성한다. 그런 다음 왼쪽 팔이 그리퍼를 열어 테이프를 놓는다 (Subtask#4 Hang).



Put On Shoe: Put a velcro-strap shoe on a fixed manniquin foot. The shoe pose is randomized along the 15cm white line. First, both left and right grippers pick up the shoe *(Subtask#1 Lift)*. Then both arms coordinate to put it on, with the heel touching the heel counter *(Subtask#2 Insert)*. Next, the left arm moves to support the shoe *(Subtask#3 Support)*, followed by the right arm securing the velcro strap *(Subtask#4 Secure)*.

Fig. 6: *Real-World Task Definitions.* For each of the 6 real-world tasks, we illustrate the initializations and the subtasks.

Fig. 6: 실제 환경 과제 정의. 6개의 실제 환경 과제 각각에 대해 초기화와 하위 과제를 보여준다.



Left: Cube Transfer. Transfer the red cube to the other arm. The right arm touches *(#1)* and grasps *(#2)* the red cube, then hands it to the left arm.
Right: Bimanual Insertion. Insert the red peg into the blue socket. Both arms grasp *(#1)*, let socket and peg make contact *(#2)* and insertion.

Fig. 7: *Simulated Task Definitions.* For each of the 2 simulated tasks, we illustrate the initializations and the subtasks.

그림 7: 시뮬레이션 과제 정의. 2개의 시뮬레이션 과제 각각에 대해 초기화 상태와 하위 과제를 제시한다.

	Cube Transfer (sim)			**Bimanual Insertion** (sim)			**Slide Ziploc** (real)			**Slot Battery** (real)		
큐브 옮기기(시뮬레이션) Bimanual Insertion(시뮬레이션) Slide Ziploc(실제) Slot Battery(실제)												
	Touched	Lifted	Transfer	Grasp	Contact	Insert	Grasp	Pinch	Open	Grasp	Place	Insert
접촉 들어 올림 전달 잡기 접촉 삽입 잡기 집기 열기 잡기 배치 삽입												
BC-ConvMLP	34 | 3	17 | 1	1 | 0	5 | 0	1 | 0	1 | 0	0	0	0	0	0	0
BC-ConvMLP 34 | 3 17 | 1 1 | 0 5 | 0 1 | 0 1 | 0 0 0 0 0 0 0												
BeT	60 | 16	51 | 13	27 | 1	21 | 0	4 | 0	3 | 0	8	0	0	4	0	0
BeT 60 | 16 51 | 13 27 | 1 21 | 0 4 | 0 3 | 0 8 0 0 4 0 0												
RT-1	44 | 4	33 | 2	2 | 0	2 | 0	0 | 0	1 | 0	4	0	0	4	0	0
RT-1 44 | 4 33 | 2 2 | 0 2 | 0 0 | 0 1 | 0 4 0 0 4 0 0												
VINN	13 | 17	9 | 11	3 | 0	6 | 0	1 | 0	1 | 0	28	0	0	20	0	0
VINN 13 | 17 9 | 11 3 | 0 6 | 0 1 | 0 1 | 0 28 0 0 20 0 0												
ACT (Ours)	**97 | 82**	**90 | 60**	**86 | 50**	**93 | 76**	**90 | 66**	**32 | 20**	**92**	**96**	**88**	**100**	**100**	**96**
ACT (우리 방법) 97 | 82 90 | 60 86 | 50 93 | 76 90 | 66 32 | 20 92 96 88 100 100 96												

TABLE I: Success rate (%) for 2 simulated and 2 real-world tasks, comparing our method with 4 baselines. For the two simulated tasks, we report [training with scripted data | training with human data], with 3 seeds and 50 policy evaluations each. For the real-world tasks, we report training with human data, with 1 seed and 25 evaluations. Overall, ACT significantly outperforms previous methods.

표 I: 2개의 시뮬레이션 과제와 2개의 실제 환경 과제에 대한 성공률(%)로, 본 방법과 4개의 기준선을 비교한다. 2개의 시뮬레이션 과제에 대해서는 3개의 시드와 각각 50회의 정책 평가를 사용하여 [스크립트 데이터로 학습 | 인간 데이터로 학습]을 보고한다. 실제 환경 과제에 대해서는 1개의 시드와 25회의 평가를 사용하여 인간 데이터로 학습한 결과를 보고한다. 전반적으로 ACT는 이전 방법보다 유의하게 뛰어난 성능을 보인다.

	Open Cup (real)			**Thread Velcro** (real)			**Prep Tape** (real)				**Put On Shoe** (real)			
	Open Cup(실제) Thread Velcro(실제)						Prep Tape(실제) Put On Shoe(실제)							
	Tip Over	Grasp	Open Lid	Lift	Grasp	Insert	Grasp	Cut	Handover	Hang	Lift	Insert	Support	Secure
	기울여 쏟기	잡기	열기 뚜껑 들어 올리기		잡기	삽입	잡기	자르기	전달	걸기	들어 올리기	삽입	지지	고정
BeT	12	0	0	24	0	0	8	0	0	0	12	0	0	0
BeT 12 0 0 24 0 0 / 8 0 0 0 12 0 0 0														
ACT (Ours)	**100**	**96**	**84**	**92**	**40**	**20**	**96**	**92**	**72**	**64**	**100**	**92**	**92**	**92**
ACT (우리 방법) 100 96 84 92 40 20 / 96 92 72 64 100 92 92 92														

TABLE II: Success rate (%) for the remaining 3 real-world tasks. We only compare with the best performing baseline BeT.

표 II: 나머지 3개의 실제 환경 과제에 대한 성공률(%). 가장 성능이 뛰어난 기준선인 BeT와만 비교한다.

tape dispenser. Especially from the top view, it is challenging to localize the velcro tie because of the small projected area.

특히 위에서 내려다본 뷰에서는 투영 면적이 작기 때문에 벨크로 끈의 위치를 파악하기가 어렵다.

B. Data Collection

B. 데이터 수집

For all 6 real-world tasks, we collect demonstrations using *ALOHA* teleoperation. Each episode takes 8-14 seconds for the human operator to perform depending on the complexity of the task, which translates to 400-700 time steps given the control frequency of 50Hz. We record 50 demonstrations for each task, except for *Thread Velcro* which has 100. The total amount for demonstrations is thus around 10-20 minutes of data for each task, and 30-60 minutes in wall-clock time because of resets and teleoperator mistakes. For the two simulated tasks, we collect two types of demonstrations: one type with a scripted policy and one with human demonstrations. To teleoperate in simulation, we use the "leader robots" of *ALOHA* to control the simulated robot, with the operator looking at the real-time renderings of the environment on the monitor. In both cases, we record 50 successful demonstrations.

6개의 실제 환경 과제 모두에 대해 ALOHA 원격조작을 사용하여 시연을 수집한다. 각 에피소드는 과제의 복잡도에 따라 인간 조작자가 수행하는 데 8~14초가 걸리며, 제어 주파수가 50Hz이므로 이는 400~700개의 시간 단계에 해당한다. Thread Velcro의 경우 100개를 수집한 것을 제외하면, 각 과제에 대해 50개의 시연을 기록한다. 따라서 시연의 총량은 각 과제당 약 10~20분의 데이터이며, 초기화와 원격조작자의 실수로 인해 실제 소요 시간은 30~60분이다. 2개의 시뮬레이션 과제에 대해서는 2가지 유형의 시연을 수집한다. 하나는 스크립트 정책을 사용하고, 다른 하나는 인간 시연을 사용한다. 시뮬레이션에서 원격조작하기 위해 ALOHA의 "leader robots"를 사용하여 시뮬레이션된 로봇을 제어하고, 조작자는 모니터에서 환경의 실시간 렌더링을 보면서 조작한다. 두 경우 모두 성공적인 시연을 50개씩 기록한다.

We emphasize that all human demonstrations are inherently stochastic, even though a single person collects all of the demonstrations. Take the mid-air hand handover of the tape segment as an example: the exact position of the handover is different across each episode. The human has no visual or haptic reference to perform it in the same position. Thus to successfully perform the task, the policy will need to learn that the two grippers should never collide with each other during the handover, and the left gripper should always move to a position that can grasp the tape, instead of trying to memorize where exactly the handover happens, which can vary across demonstrations.

모든 인간 시연은 한 사람이 모든 시연을 수집하더라도 본질적으로 확률적이라는 점을 강조한다. 테이프 조각을 공중에서 손으로 전달하는 동작을 예로 들면, 각 에피소드에서 정확한 전달 위치가 다르다. 인간에게는 이를 동일한 위치에서 수행할 수 있는 시각적 또는 촉각적 기준이 없다. 따라서 과제를 성공적으로 수행하려면 정책은 전달 중에 2개의 그리퍼가 서로 절대 충돌해서는 안 되며, 시연마다 달라질 수 있는 정확한 전달 위치를 기억하려 하기보다 왼쪽 그리퍼가 테이프를 잡을 수 있는 위치로 항상 이동해야 한다는 점을 학습해야 한다.

C. Experiment Results

C. 실험 결과

We compare ACT with four prior imitation learning methods.

ACT를 이전의 4가지 모방학습 방법과 비교한다.

BC-ConvMLP is the simplest yet most widely used baseline [69, 26], which processes the current image observations with a convolutional network, whose output features are concatenated with the joint positions to predict the action.

BC-ConvMLP는 가장 단순하면서도 널리 사용되는 기준선 [69, 26]으로, 현재 이미지 관측을 합성곱 네트워크로 처리하고 그 출력 특징을 관절 위치와 연결하여 행동을 예측한다.

BeT [49] also leverages Transformers as the architecture, but with key differences: (1) no action chunking: the model predicts one action given the history of observations; and (2) the image observations are pre-processed by a separately trained frozen visual encoder. That is, the perception and control networks are not jointly optimized. ***RT-1*** [7] is another Transformer-based architecture that predicts one action from a fixed-length history of past observations. Both *BeT* and *RT-1* discretize the action space: the output is a categorical distribution over discrete bins, but with an added continuous offset from the bin-center in the case of *BeT*. Our method, ACT, instead directly predicts continuous actions, motivated by the precision required in fine manipulation. Lastly, ***VINN*** [42] is a non-parametric method that assumes access to the demonstrations at test time. Given a new observation, it retrieves the k observations with the most similar visual features, and returns an action using

BeT [49] 역시 아키텍처로 Transformers를 활용하지만, 다음과 같은 주요 차이가 있다. (1) 행동 청킹이 없다. 즉, 모델은 관측 이력을 바탕으로 1개의 행동을 예측한다. (2) 이미지 관측은 별도로 학습된 고정 시각 인코더로 사전 처리된다. 즉, 지각 네트워크와 제어 네트워크는 공동으로 최적화되지 않는다. RT-1 [7]은 과거 관측의 고정 길이 이력을 바탕으로 하나의 행동을 예측하는 또 다른 Transformer 기반 아키텍처이다. BeT와 RT-1은 모두 행동 공간을 이산화한다. 즉, 출력은 이산 구간에 대한 범주형 분포이며, BeT의 경우 구간 중심으로부터의 연속 오프셋이 추가된다. 우리의 방법인 ACT는 정밀 조작에 필요한 정밀도를 고려하여 대신 연속적인 행동을 직접 예측한다. 마지막으로 VINN [42]은 테스트 시점에 시연 데이터에 접근할 수 있다고 가정하는 비모수적 방법이다.

weighted k-nearest-neighbors. The visual feature extractor is
새로운 관측이 주어지면 시각적 특징이 가장 유사한 k개의 관측을 검색하고, 가중 k-최근접 이웃을 사용하여 행동을 반환한다.
a pretrained ResNet finetuned on demonstration data with unsupervised learning. We carefully tune the hyperparameters
시각적 특징 추출기는 사전 학습된 ResNet을 시연 데이터에 대해 비지도학습으로 미세 조정한 것이다.
of these four prior methods using cube transfer. Details of the
4가지 기존 방법의 하이퍼파라미터를 큐브 전이를 사용하여 신중하게 조정한다.
hyperparameters are provided in Appendix D.
하이퍼파라미터의 세부 사항은 부록 D에 제시한다.

As a detailed comparison with prior methods, we report the average success rate in Table I for two simulated and two real tasks. For simulated tasks, we average performance across 3
기존 방법과의 상세한 비교를 위해 2개의 시뮬레이션 과제와 2개의 실제 과제에 대한 평균 성공률을 표 I에 보고한다.
random seeds with 50 trials each. We report the success rate
시뮬레이션 과제에서는 각각 50회 시도하는 3개의 무작위 시드에 걸쳐 성능을 평균한다.
on both scripted data (left of separation bar) and human data (right of separation bar). For real-world tasks, we run one seed
스크립트 데이터(구분선 왼쪽)와 인간 데이터(구분선 오른쪽) 모두에 대한 성공률을 보고한다.
and evaluate with 25 trials. ACT achieves the highest success
실제 환경 과제에서는 1개의 시드를 사용하고 25회 시도로 평가한다.
rate compared to all prior methods, outperforming the second best algorithm by a large margin on each task. For the two
ACT는 모든 기존 방법보다 높은 성공률을 달성하며, 각 과제에서 2위 알고리즘을 큰 차이로 앞선다.
simulated tasks with scripted or human data, ACT outperforms the best previous method in success rate by 59%, 49%, 29%, and 20%. While previous methods are able to make progress
스크립트 데이터 또는 인간 데이터를 사용하는 2개의 시뮬레이션 과제에서 ACT는 성공률 기준으로 가장 우수한 기존 방법을 59%, 49%, 29%, 20% 앞선다.
in the first two subtasks, the final success rate remains low, below 30%. For the two real-world tasks *Slide Ziploc* and
기존 방법은 처음 2개의 하위 과제에서 진전을 보이지만, 최종 성공률은 30% 미만으로 낮게 유지된다.
Slot Battery, ACT achieves 88% and 96% final success rates respectively, with other methods making no progress past the first stage. We attribute the poor performance of prior methods
2개의 실제 환경 과제인 Slide Ziploc과 Slot Battery에서 ACT는 각각 88%와 96%의 최종 성공률을 달성하는 반면, 다른 방법은 첫 번째 단계 이후 진전을 보이지 않는다.
to compounding errors and non-Markovian behavior in the data: the behavior degrades significantly towards the end of an episode, and the robot can pause indefinitely for certain states. ACT mitigates both issues with action chunking. Our
기존 방법의 저조한 성능은 데이터의 누적 오차와 비마르코프적 행동 때문이라고 본다. 즉, 에피소드가 끝날수록 행동이 크게 저하되고 로봇이 특정 상태에서 무기한 멈출 수 있다. ACT는 행동 청킹을 통해 이 2가지 문제를 완화한다.
ablations in Subsection VI-A also shows that chunking can significantly improve these prior methods when incorporated. In
하위 절 VI-A의 절제 실험에서도 청킹을 적용하면 이러한 기존 방법의 성능이 크게 향상될 수 있음을 확인했다.
addition, we notice a drop in performance for all methods when switching from scripted data to human data in simulated tasks: the stochasticity and multi-modality of human demonstrations make imitation learning a lot harder.
또한 시뮬레이션 과제에서 스크립트 데이터에서 인간 데이터로 전환하면 모든 방법의 성능이 저하되는 것을 확인했다. 인간 시연의 확률성과 다중 양태성 때문에 모방학습이 훨씬 어려워지기 때문이다.

We report the success rate of the 3 remaining real-world tasks in Table II. For these tasks, we only compare with BeT,
나머지 3개의 실제 환경 과제에 대한 성공률을 표 II에 보고한다.
which has the highest task success rate so far. Our method
이러한 과제에서는 지금까지 과제 성공률이 가장 높은 BeT와만 비교한다.
ACT reaches 84% success for *Cup Open*, 20% for *Thread Velcro*, 64% for *Prep Tape* and 92% for *Put On Shoe*, again outperforming BeT, which achieve zero final success on these challenging tasks. We observe relatively low success of ACT
우리의 방법인 ACT는 Cup Open에서 84%, Thread Velcro에서 20%, Prep Tape에서 64%, Put On Shoe에서 92%의 성공률을 달성하여, 이러한 까다로운 과제에서 최종 성공률이 0%인 BeT를 다시 앞선다.
in *Thread Velcro*, where the success rate decreased by roughly half at every stage, from 92% success at the first stage to 20% final success. The failure modes we observe are 1) at
Thread Velcro에서 ACT의 성공률은 비교적 낮았으며, 첫 번째 단계의 성공률 92%에서 최종 성공률 20%까지 각 단계에서 대략 절반씩 감소했다. 관찰한 실패 양상은 다음과 같다.
stage 2, the right arm closes its gripper too early and fails to grasp the tail of the cable tie mid-air, and 2) in stage 3, the
1) 2단계에서 오른쪽 팔이 그리퍼를 너무 일찍 닫아 공중에 있는 케이블 타이의 꼬리를 잡지 못하며,
insertion is not precise enough and misses the loop. In both
2) 3단계에서 삽입이 충분히 정밀하지 않아 고리를 놓친다.
cases, it is hard to determine the exact position of the cable tie from image observations: the contrast is low between the black cable tie and the background, and the cable tie only occupies a small fraction of the image. We include examples of image
두 경우 모두 영상 관측으로부터 케이블 타이의 정확한 위치를 판단하기 어렵다. 검은색 케이블 타이와 배경 사이의 대비가 낮고, 케이블 타이가 영상에서 차지하는 비율도 작기 때문이다.
observations in Appendix B.
영상 관측의 예시는 부록 B에 포함한다.

VI. Ablations
절제 실험

ACT employs action chunking and temporal ensembling to mitigate compounding errors and better handle non-Markovian demonstrations. It also trains the policy as a conditional VAE
ACT는 누적 오차를 완화하고 비마르코프적 시연을 더 잘 처리하기 위해 행동 청킹과 시간적 앙상블을 사용한다.
to model the noisy human demonstrations. In this section, we
또한 노이즈가 있는 인간 시연을 모델링하기 위해 정책을 조건부 VAE로 학습한다.
ablate each of these components, together with a user study that highlights the necessity of high-frequency control in *ALOHA*.
이 절에서는 각 구성 요소에 대한 절제 실험과 함께 ALOHA에서 고주파 제어의 필요성을 보여 주는 사용자 연구를 수행한다.
We report results across a total of four settings: two simulated tasks with scripted or human demonstration.
총 4가지 설정에 걸쳐 결과를 보고한다. 즉, 스크립트 또는 인간 시연을 사용하는 2개의 시뮬레이션 과제이다.

A. Action Chunking and Temporal Ensembling
A. 행동 청킹과 시간적 앙상블

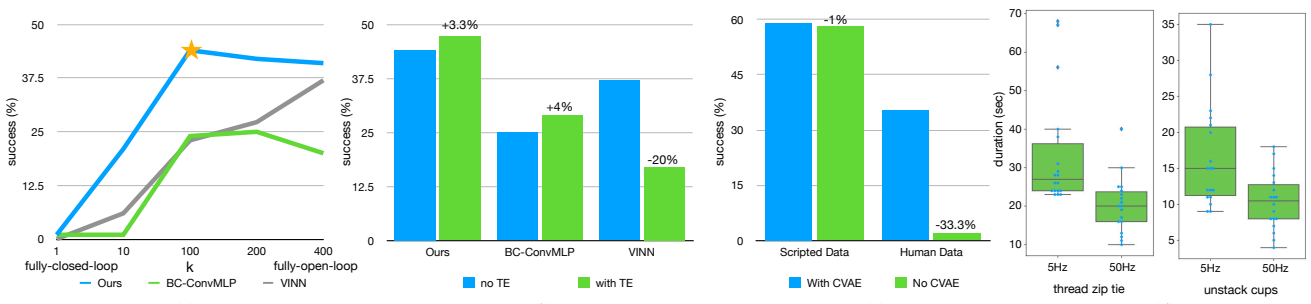
In Subsection V-C, we observed that ACT significantly outperforms previous methods that only predict single-step actions, with the hypothesis that action chunking is the key design choice. Since k dictates how long the sequence in each
하위 절 V-C에서 단일 단계 행동만 예측하는 기존 방법보다 ACT가 크게 우수한 것을 확인했으며, 행동 청킹이 핵심 설계 선택이라는 가설을 세웠다.
"chunk" is, we can analyze this hypothesis by varying k. $k = 1$
k은 각 "청크"의 시퀀스 길이를 결정하므로, k을 변화시켜 이 가설을 분석할 수 있다.
corresponds to no action chunking, and $k = episode\_length$ corresponds to fully open-loop control, where the robot outputs the entire episode's action sequence based on the first observation. We disable temporal ensembling in these
k = 1은 행동 청킹을 사용하지 않는 경우에 해당하고, k = episode_length는 완전한 오픈 루프 제어에 해당한다. 이 경우 로봇은 첫 번째 관측을 바탕으로 전체 episode의 행동 시퀀스를 출력한다.
experiments to only measure the effect of chunking, and trained separate policies for each k. In Figure 8 (a), we plot the success
이 실험에서는 청킹의 효과만 측정하기 위해 시간적 앙상블을 비활성화하고, 각 k에 대해 별도의 정책을 학습했다.
rate averaged across 4 settings, corresponding to 2 simulated tasks with either human or scripted data, with the blue line representing ACT without the temporal ensemble. We observe
그림 8(a)에서는 인간 데이터 또는 스크립트 데이터를 사용하는 2개의 시뮬레이션 과제에 해당하는 4가지 설정에 걸쳐 평균한 성공률을 표시하며, 파란색 선은 시간적 앙상블을 사용하지 않는 ACT를 나타낸다.
that performance improves drastically from 1% at $k = 1$ to 44% at $k = 100$, then slightly tapers down with higher k. This
k = 1에서 1%였던 성능이 k = 100에서 44%로 크게 향상된 후, k가 증가하면 약간 감소하는 것을 확인했다.
illustrates that more chunking and a lower effective horizon generally improve performance. We attribute the slight dip at
이는 일반적으로 청킹을 더 많이 적용하고 유효 시간 범위를 줄일수록 성능이 향상됨을 보여 준다.
$k = 200, 400$ (i.e., close to open-loop control) to the lack of reactive behavior and the difficulty in modeling long action sequences. To further evaluate the effectiveness and generality
k = 200, 400에서 나타나는 약간의 하락(즉, 오픈 루프 제어에 가까운 경우)은 반응적 행동의 부족과 긴 행동 시퀀스를 모델링하기 어려운 데 기인한다고 본다.
of action chunking, we augment two baseline methods with action chunking. For *BC-ConvMLP*, we simply increase the
행동 청킹의 효과와 일반성을 추가로 평가하기 위해 2개의 기준 방법에 행동 청킹을 적용한다.
output dimension to $k * action\_dim$, and for *VINN*, we retrieve the next k actions. We visualize their performance in Figure 8
BC-ConvMLP의 경우 출력 차원을 k∗action_dim으로 단순히 늘리고, VINN의 경우 다음 k개의 행동을 검색한다. 그림 8에서 이들의 성능을 시각화한다.
(a) with different k, showing trends consistent with ACT, where more action chunking improves performance. While ACT still outperforms both augmented baselines with sizable gains, these results suggest that action chunking is generally beneficial for imitation learning in these settings.

We then ablate the temporal ensemble by comparing the highest success rate with or without it, again across the 4 aforementioned tasks and different k. We note that experiments with and without the temporal ensemble are separately tuned: hyperparameters that work best for no temporal ensemble may not be optimal with a temporal ensemble. In Figure 8 (b), we show that *BC-ConvMLP* benefits from temporal ensembling the most with a 4% gain, followed by a 3.3% gain for our method. We notice a performance drop for *VINN*, a non-parametric method. We hypothesize that a temporal ensemble mostly benefits parametric methods by smoothing out the modeling errors. In contrast, *VINN* retrieves ground-truth actions from the dataset and does not suffer from this issue.

(a) 서로 다른 k에 대한 결과를 제시하며, 행동 청킹을 더 많이 적용할수록 성능이 향상되는 등 ACT와 일관된 추세를 보인다. ACT는 여전히 두 가지로 확장한 기준 방법보다 상당한 차이로 우수하지만, 이러한 결과는 이러한 설정에서 행동 청킹이 일반적으로 모방학습에 유익하다는 점을 시사한다. 그런 다음 앞서 언급한 4개 과제와 서로 다른 k에 대해 시간적 앙상블을 적용한 경우와 적용하지 않은 경우의 최고 성공률을 비교하여 시간적 앙상블을 절제한다. 시간적 앙상블을 적용한 실험과 적용하지 않은 실험은 별도로 조정했음을 명시한다. 시간적 앙상블을 적용하지 않을 때 가장 잘 작동하는 하이퍼파라미터가 시간적 앙상블을 적용할 때 최적이 아닐 수 있다. 그림 8(b)에서는 BC-ConvMLP가 4% 향상으로 시간적 앙상블의 혜택을 가장 크게 받고, 이어서 우리 방법이 3.3% 향상되는 것을 보인다. 비모수 방법인 VINN에서는 성능이 하락하는 것을 확인한다. 시간적 앙상블은 모델링 오류를 평활화하므로 주로 모수적 방법에 도움이 된다고 가정한다. 반면 VINN은 데이터셋에서 실제 정답 행동을 검색하므로 이러한 문제를 겪지 않는다.

B. Training with CVAE
B. CVAE를 사용한 학습

We train ACT with CVAE objective to model human demonstrations, which can be noisy and contain multi-modal behavior. In this section, we compare with ACT without the
노이즈가 있을 수 있고 다중 모드 행동을 포함하는 인간 시연을 모델링하기 위해 CVAE 목적 함수로 ACT를 학습한다.
CVAE objective, which simply predicts a sequence of actions given current observation, and trained with L1 loss. In Figure 8
이 절에서는 CVAE 목적 함수가 없는 ACT와 비교한다. 이 방법은 현재 관측이 주어졌을 때 행동 시퀀스를 단순히 예측하며 L1 손실로 학습한다. 그림 8에서



(a) (b) (c) (d)

Fig. 8: *(a)* We augment two baselines with action chunking, with different values of chunk size k on the x-axis, and success rate on the y-axis. Both methods significantly benefit from action chunking, suggesting that it is a generally useful technique. *(b)* Temporal Ensemble (TE) improves our method and *BC-ConvMLP*, while hurting *VINN*. *(c)* We compare with and without the CVAE training, showing that it is crucial when learning from human data. *(d)* We plot the distribution of task completion time in our user study, where we task participants to perform two tasks, at 5Hz or 50Hz teleoperation frequency. Lowering the frequency results in a 62% slowdown in completion time.

그림 8: (a) 2개의 기준선에 행동 청킹을 적용하고, x축에는 청크 크기 k의 서로 다른 값을, y축에는 성공률을 표시한다. 두 방법 모두 행동 청킹을 통해 크게 향상되며, 이는 행동 청킹이 일반적으로 유용한 기법임을 시사한다. (b) 시간 앙상블(TE)은 우리의 방법과 BC-ConvMLP의 성능을 향상시키지만 VINN의 성능은 저하시킨다. (c) CVAE 학습을 적용한 경우와 적용하지 않은 경우를 비교한 결과, 인간 데이터로 학습할 때 CVAE 학습이 중요함을 확인한다. (d) 사용자 연구에서 과제 완료 시간의 분포를 도시한다. 참가자들에게 5Hz 또는 50Hz 원격조작 주파수로 2개의 과제를 수행하도록 했다. 주파수를 낮추면 완료 시간이 62% 느려진다.

(c), we visualize the success rate aggregated across 2 simulated tasks, and separately plot training with scripted data and with human data. We can see that when training on scripted data,

(c) 2개의 시뮬레이션 과제에 대해 집계한 성공률을 시각화하고, 스크립트 데이터와 인간 데이터를 사용한 학습 결과를 각각 표시한다.

the removal of CVAE objective makes almost no difference in performance, because dataset is fully deterministic. While for

스크립트 데이터로 학습할 때 데이터셋이 완전히 결정론적이므로 CVAE 목적 함수를 제거해도 성능에는 거의 차이가 없음을 확인할 수 있다.

human data, there is a significant drop from 35.3% to 2%. This

반면 인간 데이터에서는 35.3%에서 2%로 크게 하락한다.

illustrates that the CVAE objective is crucial when learning from human demonstrations.

이는 인간 시연으로 학습할 때 CVAE 목적 함수가 중요함을 보여준다.

C. Is High-Frequency Necessary?

C. 고주파가 필요한가?

Lastly, we conduct a user study to illustrate the necessity of high-frequency teleoperation for fine manipulation. With

마지막으로 정밀 조작에 고주파 원격조작이 필요한 이유를 보여주기 위해 사용자 연구를 수행한다.

the same hardware setup, we lower the frequency from 50Hz to 5Hz, a control frequency that is similar to recent works that use high-capacity deep networks for imitation learning

동일한 하드웨어 설정에서 주파수를 50Hz에서 5Hz로 낮춘다. 이는 고용량 심층 네트워크를 사용하는 최근 모방학습 연구의 제어 주파수와 유사하다 [7, 70].

[7, 70]. We pick two fine-grained tasks: threading a zip cable tie and un-stacking two plastic cups. Both require millimeter-level precision and closed-loop visual feedback. We perform

2개의 세밀한 과제, 즉 케이블 타이를 끼우는 작업과 플라스틱 컵 2개를 분리해 쌓이지 않도록 하는 작업을 선정한다. 두 과제 모두 밀리미터 수준의 정밀도와 폐루프 시각 피드백을 필요로 한다.

the study with 6 participants who have varying levels of experience with teleoperation, though none had used *ALOHA* before. The participants were recruited from among computer science graduate students, with 4 men and 2 women aged 22-25 The order of tasks and frequencies are randomized for each participant, and each participant was provided with a 2 minutes practice period before each trial. We recorded the time it took to perform the task for 3 trials, and visualize the data in Figure 8 (d). On average, it took 33s for participants to thread the zip tie at 5Hz, which is lowered to 20s at 50Hz. For separating plastic cups, increasing the control frequency lowered the task duration from 16s to 10s. Overall, our setup (i.e. 50Hz) allows the participants to perform highly dexterous and precise tasks in a short amount of time. However, reducing the frequency from 50Hz to 5Hz results in a 62% increase in teleoperation time. We then use "Repeated Measures Designs", a statistical procedure, to formally verify that 50Hz teleoperation outperforms 5Hz with p-value <0.001. We include more details about the study in Appendix E.

원격조작 경험 수준이 서로 다른 참가자 6명과 함께 연구를 수행했으며, 이들 중 ALOHA를 사용해 본 사람은 없었다.

참가자는 컴퓨터과학 대학원생 중에서 모집했으며, 22~25세의 남성 4명과 여성 2명으로 구성되었다. 참가자마다 과제와 주파수의 순서를 무작위화했으며, 각 시행 전에 2분의 연습 시간을 제공했다.

3회 시행에서 과제를 수행하는 데 걸린 시간을 기록하고, 그 데이터를 그림 8(d)에 시각화한다.

참가자가 5Hz에서 케이블 타이를 끼우는 데 걸린 시간은 평균 33초였으며, 50Hz에서는 20초로 줄어들었다.

플라스틱 컵을 분리하는 경우 제어 주파수를 높이면 과제 수행 시간이 16초에서 10초로 줄어들었다.

전반적으로 우리의 설정(즉, 50Hz)을 사용하면 참가자가 고도로 손재주가 필요하고 정밀한 과제를 짧은 시간 안에 수행할 수 있다.

그러나 주파수를 50Hz에서 5Hz로 낮추면 원격조작 시간이 62% 증가한다.

그런 다음 통계 절차인 "반복 측정 설계"를 사용하여 50Hz 원격조작이 p값 <0.001로 5Hz보다 우수함을 공식적으로 검증한다.

연구에 대한 자세한 내용은 부록 E에 제시한다.

VII. Limitations and Conclusion

한계와 결론

We present a low-cost system for fine manipulation, comprising a teleoperation system *ALOHA* and a novel imitation learning algorithm *ACT*. The synergy between these two parts allows us to learn fine manipulation skills directly in the real-world, such as opening a translucent condiment cup and slotting a battery with a 80-90% success rate and around 10 min of demonstrations. While the system is quite capable, there exist tasks that are beyond the capability of either the robots or the learning algorithm, such as buttoning up a dress shirt.

정밀 조작을 위한 저비용 시스템을 제시한다. 이 시스템은 원격조작 시스템 ALOHA와 새로운 모방학습 알고리즘 ACT로 구성된다.

이 2개 구성 요소의 시너지 효과를 통해 반투명 양념 컵을 열고 배터리를 삽입하는 작업과 같은 정밀 조작 기술을 실제 환경에서 직접 학습할 수 있으며, 약 10분의 시연만으로 80~90%의 성공률을 달성한다.

이 시스템은 상당한 능력을 갖추고 있지만, 셔츠 단추를 채우는 작업처럼 로봇이나 학습 알고리즘 중 어느 하나의 능력을 넘어서는 과제도 존재한다.

We include a more detailed discussion about limitations in Appendix F. Overall, we hope that this low-cost open-source system represents an important step and accessible resource towards advancing fine-grained robotic manipulation.

한계에 대한 더 자세한 논의는 부록 F에 제시한다. 전반적으로 이 저비용 오픈소스 시스템이 세밀한 로봇 조작 기술의 발전을 위한 중요한 단계이자 접근성 높은 자원이 되기를 기대한다.

Acknowledgement

감사의 글

We thank members of the IRIS lab at Stanford for their support and feedback. We also thank Siddharth Karamcheti, Toki Migimatsu, Staven Cao, Huihan Liu, Mandi Zhao, Pete Florence and Corey Lynch for helpful discussions. Tony Zhao is supported by Stanford Robotics Fellowship sponsored by FANUC, in addition to Schmidt Futures and ONR Grant N00014-21-1-2685.

Stanford의 IRIS lab 구성원들에게 지원과 피드백을 보내준 데 대해 감사드린다.

또한 유익한 논의를 해준 Siddharth Karamcheti, Toki Migimatsu, Staven Cao, Huihan Liu, Mandi Zhao, Pete Florence 및 Corey Lynch에게 감사드린다.

Tony Zhao는 Schmidt Futures와 ONR Grant N00014-21-1-2685의 지원과 더불어 FANUC가 후원하는 Stanford Robotics Fellowship의 지원을 받는다.

References

참고문헌

[1] Viperx 300 robot arm 6dof. URL https://www.trossenrobotics.com/viperx-300-robot-arm-6dof.aspx.

[1] Viperx 300 로봇 팔 6자유도. URL https://www.trossenrobotics.com/viperx-300-robot-arm-6dof.aspx.

ViperX 300 6자유도 로봇 팔: trossenrobotics.com/viperx-300-robot-arm-6dof.aspx.

[2] Widowx 250 robot arm 6dof. URL https://www.trossenrobotics.com/widowx-250-robot-arm-6dof.aspx.

[2] Widowx 250 로봇 팔 6자유도. URL https://www.trossenrobotics.com/widowx-250-robot-arm-6dof.aspx.

WidowX 250 6자유도 로봇 팔: trossenrobotics.com/widowx-250-robot-arm-6dof.aspx.

[3] Highly dexterous manipulation system - capabilities - part 1, Nov 2014. URL https://www.youtube.com/watch?v=TearcKVj0iY.

[3] 고도로 손재주가 필요한 조작 시스템 - 기능 - 1부, 2014년 11월. URL https://www.youtube.com/watch?v= TearcKVj0iY.

[4] Assembly performance metrics and test methods, Apr 2022. URL https://www.nist.gov/el/intelligent-systems-division-73500/robotic-grasping-and-manipulation-assembly/assembly.

[4] 조립 성능 지표 및 시험 방법, 2022년 4월. URL https://www.nist.gov/el/intelligent-systems-division-73500/robotic-grasping-and-manipulation-assembly/assembly.

[5] Teleoperated robots - shadow teleoperation system, Nov 2022. URL https://www.shadowrobot.com/teleoperation/.

[5] 원격조작 로봇 - shadow 원격조작 시스템, 2022년 11월. URL https://www.shadowrobot.com/teleoperation/.

[6] Sridhar Pandian Arunachalam, Irmak Güzey, Soumith Chintala, and Lerrel Pinto. Holo-dex: Teaching dexterity with immersive mixed reality. *arXiv preprint arXiv:2210.06463*, 2022.

Holo-dex: 몰입형 혼합 현실을 통한 손재주 교육. arXiv 프리프린트 arXiv:2210.06463, 2022년.

[7] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana

Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J. Joshi, Ryan C. Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael S. Ryoo, Grecia Salazar, Pannag R. Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Anand Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Ho Vuong, F. Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-1: Robotics transformer

Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J. Joshi, Ryan C. Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang- Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael S. Ryoo, Grecia Salazar, Pannag R. Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Anand Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Ho Vuong, F. Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, 및 Brianna Zitkovich.

for real-world control at scale. *ArXiv*, abs/2212.06817, 2022.

Rt-1: 실제 환경의 대규모 제어를 위한 로보틱스 Transformer.

ArXiv, abs/2212.06817, 2022.

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko.

[8] 저자 Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, 및 Sergey Zagoruyko.

End-to-end object detection with transformers. *ArXiv*, abs/2005.12872, 2020.

종단 간 Transformer를 이용한 객체 검출.

ArXiv, abs/2005.12872, 2020.

[9] Yuanpei Chen, Yaodong Yang, Tianhao Wu, Shengjie Wang, Xidong Feng, Jiechuan Jiang, Stephen McAleer, Hao Dong, Zongqing Lu, and Song-Chun Zhu. Towards

[9] 저자 Yuanpei Chen, Yaodong Yang, Tianhao Wu, Shengjie Wang, Xidong Feng, Jiechuan Jiang, Stephen McAleer, Hao Dong, Zongqing Lu, 및 Song-Chun Zhu.

human-level bimanual dexterous manipulation with reinforcement learning. *ArXiv*, abs/2206.08686, 2022.

강화학습을 통한 인간 수준의 양손 손재주 조작을 향하여. ArXiv, abs/2206.08686, 2022.

[10] Rohan Chitnis, Shubham Tulsiani, Saurabh Gupta, and Abhinav Kumar Gupta. Efficient bimanual manipulation

[10] 저자 Rohan Chitnis, Shubham Tulsiani, Saurabh Gupta, 및 Abhinav Kumar Gupta. Efficient bimanual manipulation

저자 Abhinav Kumar Gupta.

using learned task schemas. *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1149–1155, 2019.

학습된 작업 스키마를 이용한 효율적인 양손 조작.

2020 IEEE International Conference on Robotics and Automation (ICRA), 1149-1155쪽, 2019.

[11] Sudeep Dasari and Abhinav Kumar Gupta. Transformers

[11] 저자 Sudeep Dasari 및 Abhinav Kumar Gupta. Transformers

for one-shot visual imitation. In *Conference on Robot Learning*, 2020.

원샷 시각 모방을 위한.

Conference on Robot Learning에서, 2020년.

[12] Pim de Haan, Dinesh Jayaraman, and Sergey Levine.

[12] 저자 Pim de Haan, Dinesh Jayaraman, 및 Sergey Levine.

Causal confusion in imitation learning. In *Neural Information Processing Systems*, 2019.

모방학습에서의 인과적 혼동.

Neural Information Processing Systems에서, 2019년.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.

[13] 저자 Jacob Devlin, Ming-Wei Chang, Kenton Lee, 및 Kristina Toutanova.

Bert: 언어 이해를 위한 심층 양방향 Transformer의 사전 학습.

ArXiv, abs/1810.04805, 2019.

[14] Yan Duan, Marcin Andrychowicz, Bradly C. Stadie, Jonathan Ho, Jonas Schneider, Ilya Sutskever, P. Abbeel, and Wojciech Zaremba. One-shot imitation learning.

[14] 저자 Yan Duan, Marcin Andrychowicz, Bradly C. Stadie, Jonathan Ho, Jonas Schneider, Ilya Sutskever, P. Abbeel, 및 Wojciech Zaremba. 원샷 모방학습.

ArXiv, abs/1703.07326, 2017.

ArXiv, abs/1703.07326, 2017.

[15] Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, and Sergey Levine. Bridge data: Boosting

[15] 저자 Frederik Ebert, Yanlai Yang, Karl Schmeckpeper, Bernadette Bucher, Georgios Georgakis, Kostas Daniilidis, Chelsea Finn, 및 Sergey Levine.

generalization of robotic skills with cross-domain datasets. *ArXiv*, abs/2109.13396, 2021.

Bridge data: 도메인 간 데이터셋을 통한 로봇 기술의 일반화 향상.

ArXiv, abs/2109.13396, 2021.

[16] Peter R. Florence, Lucas Manuelli, and Russ Tedrake. Self-supervised correspondence in visuomotor policy learning. *IEEE Robotics and Automation Letters*, 5:492–499, 2019.

[16] 저자 Peter R. Florence, Lucas Manuelli, 및 Russ Tedrake.

자기-

지도 시각운동 정책 학습에서의 대응 관계.

IEEE Robotics and Automation Letters, 5:492-499쪽, 2019년.

[17] Peter R. Florence, Corey Lynch, Andy Zeng, Oscar Ramirez, Ayzaan Wahid, Laura Downs, Adrian S. Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. *ArXiv*, abs/2109.00137, 2021.

[17] 저자 Peter R. Florence, Corey Lynch, Andy Zeng, Oscar Ramirez, Ayzaan Wahid, Laura Downs, Adrian S. Wong, Johnny Lee, Igor Mordatch, 및 Jonathan Tompson.

암시적 행동 복제. ArXiv, abs/2109.00137, 2021.

[18] Aditya Ganapathi, Priya Sundaresan, Brijen Thananjeyan, Ashwin Balakrishna, Daniel Seita, Jennifer Grannen, Minho Hwang, Ryan Hoque, Joseph Gonzalez, Nawid Jamali, Katsu Yamane, Soshi Iba, and Ken Goldberg.

저자: [18] Aditya Ganapathi, Priya Sundaresan, Brijen Thananjeyan, Ashwin Balakrishna, Daniel Seita, Jennifer Grannen, Minho Hwang, Ryan Hoque, Joseph Gonzalez, Nawid Jamali, Katsu Yamane, Soshi Iba, and Ken Goldberg.

저자: Ashwin Balakrishna, Daniel Seita, Jennifer Grannen, Minho Hwang, Ryan Hoque, Joseph Gonzalez, Nawid Jamali, Katsu Yamane, Soshi Iba 및 Ken Goldberg.

Learning dense visual correspondences in simulation to smooth and fold real fabrics. *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11515–11522, 2020.

실제 직물을 매끄럽게 펴고 접기 위해 시뮬레이션에서 밀집 시각 대응을 학습한다.

2021 IEEE International Conference on Robotics and Automation (ICRA), 11515-11522쪽, 2020.

[19] Jennifer Grannen, Priya Sundaresan, Brijen Thananjeyan, Jeffrey Ichnowski, Ashwin Balakrishna, Minho Hwang, Vainavi Viswanath, Michael Laskey, Joseph Gonzalez, and Ken Goldberg. Untangling dense knots by learning

저자: [19] Jennifer Grannen, Priya Sundaresan, Brijen Thananjeyan, Jeffrey Ichnowski, Ashwin Balakrishna, Minho Hwang, Vainavi Viswanath, Michael Laskey, Joseph Gonzalez, and Ken Goldberg. Untangling dense knots by learning

저자: Jeffrey Ichnowski, Ashwin Balakrishna, Minho Hwang, Vainavi Viswanath, Michael Laskey, Joseph Gonzalez 및 Ken Goldberg.

task-relevant keypoints. In *Conference on Robot Learning*, 2020.

과제 관련 키포인트를 학습하여 밀집 매듭을 푼다.

Conference on Robot Learning에서, 2020.

[20] Huy Ha and Shuran Song. Flingbot: The unreasonable

저자: [20] Huy Ha 및 Shuran Song. Flingbot: 천을 펼치는 동적 조작의

effectiveness of dynamic manipulation for cloth unfolding. *ArXiv*, abs/2105.03655, 2021.

비합리적으로 높은 효과.

ArXiv, abs/2105.03655, 2021.

[21] Ankur Handa, Karl Van Wyk, Wei Yang, Jacky Liang, Yu-Wei Chao, Qian Wan, Stan Birchfield, Nathan D. Ratliff, and Dieter Fox. Dexpilot: Vision-based teleoperation

저자: [21] Ankur Handa, Karl Van Wyk, Wei Yang, Jacky Liang, Yu-Wei Chao, Qian Wan, Stan Birchfield, Nathan D. Ratliff, and Dieter Fox. Dexpilot: Vision-based teleoperation

저자: Wei Chao, Qian Wan, Stan Birchfield, Nathan D. Ratliff 및 Dieter Fox.

of dexterous robotic hand-arm system. *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9164–9170, 2019.

Dexpilot: 비전 기반 손재주형 로봇 손-팔 시스템의 원격조작.

2020 IEEE International Conference on Robotics and Automation (ICRA), 9164-9170쪽, 2019.

[22] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015.

[22] Kaiming He, X. Zhang, Shaoqing Ren 및 Jian Sun.

이미지 인식을 위한 심층 잔차 학습.

2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778쪽, 2015.

[23] Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning

[23] Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning

저자: Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed 및 Alexander Lerchner.

basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2016.

beta-vae: 제약된 변분 프레임워크를 이용한 기초 시각 개념 학습.

International Conference on Learning Representations에서, 2016.

[24] Ryan Hoque, Ashwin Balakrishna, Ellen R. Novoseller, Albert Wilcox, Daniel S. Brown, and Ken Goldberg. Thriftydagger: Budget-aware novelty and risk gating for interactive imitation learning. In *Conference on Robot Learning*, 2021.

[24] Ryan Hoque, Ashwin Balakrishna, Ellen R. Novoseller, Albert Wilcox, Daniel S. Brown, and Ken Goldberg. Thriftydagger: Budget-aware novelty and risk gating for interactive imitation learning. In *Conference on Robot Learning*, 2021.

저자: Albert Wilcox, Daniel S. Brown 및 Ken Goldberg.

Thriftydagger: 상호작용 모방학습을 위한 예산 인식형 신규성 및 위험 게이팅.

Conference on Robot Learning에서, 2021.

[25] Stephen James, Michael Bloesch, and Andrew J. Davison. Task-embedded control networks for few-shot imitation learning. *ArXiv*, abs/1810.03237, 2018.

[25] Stephen James, Michael Bloesch 및 Andrew J. Davison. Task-embedded control networks for few-shot imitation learning. *ArXiv*, abs/1810.03237, 2018.

소수 샷 모방학습을 위한 과제 임베디드 제어 네트워크. ArXiv, abs/1810.03237, 2018.

[26] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, 2022.

[26] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Conference on Robot Learning*, 2022.

저자: Frederik Ebert, Corey Lynch, Sergey Levine 및 Chelsea Finn.

Bc-z: 로봇 모방학습을 이용한 제로샷 과제 일반화.

Conference on Robot Learning에서, 2022.

[27] R G Jenness and C D Wicker. Master–slave manipulators and remote maintenance at the oak ridge national laboratory, Jan 1975. URL https://www.osti.gov/biblio/4179544.

[27] R G Jenness 및 C D Wicker. 마스터-슬레이브 매니퓰레이터

및 oak ridge national laboratory에서의 원격 유지보수, 1975년 1월. URL https://www.osti.gov/biblio/4179544.

[28] Edward Johns. Coarse-to-fine imitation learning: Robot manipulation from a single demonstration. *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4613–4619, 2021.

[28] Edward Johns. Coarse-to-fine 모방학습: 로봇

단일 시연을 통한 조작.

2021 IEEE International Conference on Robotics and Automation (ICRA), 4613-4619쪽, 2021.

[29] Liyiming Ke, Jingqiang Wang, Tapomayukh Bhattacharjee, Byron Boots, and Siddhartha Srinivasa. Grasping with chopsticks: Combating covariate shift in model-free imitation learning for fine manipulation. In *International*

[29] 저자 Liyiming Ke, Jingqiang Wang, Tapomayukh Bhattacharjee, Byron Boots, and Siddhartha Srinivasa. Grasping with chopsticks: Combating covariate shift in model-free imitation learning for fine manipulation. In *International*

저자 Liyiming Ke, Jingqiang Wang, Tapomayukh Bhattacharjee, Byron Boots 및 Siddhartha Srinivasa.

젓가락을 사용한 파지: 정밀 조작을 위한 모델 비사용 모방학습의 공변량 시프트 대응. International에서

Conference on Robotics and Automation (ICRA), 2021.
Conference on Robotics and Automation (ICRA), 2021년 학회.

[30] Michael Kelly, Chelsea Sidrane, K. Driggs-Campbell,
[30] 저자 Michael Kelly, Chelsea Sidrane, K. Driggs-Campbell,
and Mykel J. Kochenderfer. Hg-dagger: Interactive
및 Mykel J. Kochenderfer.
imitation learning with human experts. *2019 International*
인간 전문가와 함께하는 상호작용 모방학습을 위한 Hg-dagger.
Conference on Robotics and Automation (ICRA), pages
8077–8083, 2018.
2019 International Conference on Robotics and Automation (ICRA), 8077-8083쪽, 2018.

[31] Heecheol Kim, Yoshiyuki Ohmura, and Yasuo Kuniyoshi.
[31] 저자 Heecheol Kim, Yoshiyuki Ohmura, 및 Yasuo Kuniyoshi.
Gaze-based dual resolution deep imitation learning for
high-precision dexterous robot manipulation. *IEEE*
고정밀 손재주 로봇 조작을 위한 시선 기반 이중 해상도 심층 모방학습.
Robotics and Automation Letters, 6:1630–1637, 2021.
IEEE Robotics and Automation Letters 저널, 6:1630-1637, 2021년.

[32] Heecheol Kim, Yoshiyuki Ohmura, and Yasuo Kuniyoshi.
[32] 저자 Heecheol Kim, Yoshiyuki Ohmura, 및 Yasuo Kuniyoshi.
Robot peels banana with goal-conditioned dual-action
deep imitation learning. *ArXiv*, abs/2203.09749, 2022.
목표 조건부 이중 행동 심층 모방학습을 통한 로봇의 바나나 껍질 벗기기. ArXiv, abs/2203.09749, 2022.

[33] Diederik P. Kingma and Max Welling. Auto-encoding
[33] 저자 Diederik P. Kingma 및 Max Welling. 변분 베이즈
variational bayes. *CoRR*, abs/1312.6114, 2013.
자동 인코딩. CoRR, abs/1312.6114, 2013.

[34] Oliver Kroemer, Christian Daniel, Gerhard Neumann,
[34] 저자 Oliver Kroemer, Christian Daniel, Gerhard Neumann,
Herke van Hoof, and Jan Peters. Towards learning
Herke van Hoof, 및 Jan Peters.
hierarchical skills for multi-phase manipulation tasks.
다단계 조작 과제를 위한 계층적 기술 학습을 향하여.
2015 IEEE International Conference on Robotics and
Automation (ICRA), pages 1503–1510, 2015.
2015 IEEE International Conference on Robotics and Automation (ICRA), 1503-1510쪽, 2015.

[35] Lucy Lai, Ann Z Huang, and Samuel J Gershman. Action
[35] 저자 Lucy Lai, Ann Z Huang, 및 Samuel J Gershman. 정책 압축으로서의 행동 청킹
chunking as policy compression, Sep 2022. URL psyarxiv.
2022년 9월. URL psyarxiv.
com/z8yrv.
com/z8yrv.

[36] Michael Laskey, Jonathan Lee, Roy Fox, Anca D. Dragan,
[36] 저자 Michael Laskey, Jonathan Lee, Roy Fox, Anca D. Dragan,
and Ken Goldberg. Dart: Noise injection for robust
및 Ken Goldberg.
imitation learning. In *Conference on Robot Learning*,
강건한 모방학습을 위한 잡음 주입을 수행하는 Dart.
2017.
Conference on Robot Learning에서 2017년.

[37] Alex X. Lee, Henry Lu, Abhishek Gupta, Sergey Levine,
[37] 저자 Alex X. Lee, Henry Lu, Abhishek Gupta, Sergey Levine,
and P. Abbeel. Learning force-based manipulation
and P. Abbeel.
of deformable objects from multiple demonstrations.
다중 시연을 통한 변형 가능한 물체의 힘 기반 조작 학습.
2015 IEEE International Conference on Robotics and
Automation (ICRA), pages 177–184, 2015.
2015 IEEE International Conference on Robotics and Automation (ICRA), 177-184쪽, 2015.

[38] Weiwei Li. Optimal control for biological movement
[38] Weiwei Li. 생물학적 운동 시스템을 위한 최적 제어
systems. 2006.
2006.

[39] Ajay Mandlekar, Danfei Xu, J. Wong, Soroush Nasiriany,
[39] 저자 Ajay Mandlekar, Danfei Xu, J. Wong, Soroush Nasiriany,
Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese,
Yuke Zhu, and Roberto Mart'in-Mart'in. What matters
Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, 및 Roberto Mart'in-Mart'in.
in learning from offline human demonstrations for robot
manipulation. In *Conference on Robot Learning*, 2021.
로봇 조작을 위한 오프라인 인간 시연 학습에서 중요한 요소.
Conference on Robot Learning에서 2021년.

[40] Kunal Menda, K. Driggs-Campbell, and Mykel J. Kochen-
[40] 저자 Kunal Menda, K. Driggs-Campbell, 및 Mykel J. Kochen-
derfer. Ensembledagger: A bayesian approach to safe
imitation learning. *2019 IEEE/RSJ International Confer-*
Ensembledagger: 안전한 모방학습을 위한 베이지안 접근법.
ence on Intelligent Robots and Systems (IROS), pages
5041–5048, 2018.
2019 IEEE/RSJ 지능형 로봇 및 시스템 국제 학술대회(IROS), 5041-5048쪽, 2018.

[41] Samuel Paradis, Minho Hwang, Brijen Thananjeyan,
[41] 저자: Samuel Paradis, Minho Hwang, Brijen Thananjeyan,
Jeffrey Ichnowski, Daniel Seita, Danyal Fer, Thomas
Low, Joseph Gonzalez, and Ken Goldberg. Intermittent
visual servoing: Efficiently learning policies robust to
instrument changes for high-precision surgical manipula-
tion. *2021 IEEE International Conference on Robotics*
간헐적 시각 서보잉: 고정밀 수술 조작을 위한 기기 변경에 강건한 정책을 효율적으로 학습하기.
and Automation (ICRA), pages 7166–7173, 2020.
2021 IEEE 로봇공학 및 자동화 국제 학술대회(ICRA), 7166-7173쪽, 2020.

[42] Jyothish Pari, Nur Muhammad, Sridhar Pandian Arunacha-
[42] 저자: Jyothish Pari, Nur Muhammad, Sridhar Pandian Arunacha-
lam, and Lerrel Pinto. The surprising effectiveness of
representation learning for visual imitation. *arXiv preprint*
시각 모방을 위한 표현 학습의 놀라운 효과.
arXiv:2112.01511, 2021.
arXiv 프리프린트 arXiv:2112.01511, 2021.

[43] Peter Pastor, Heiko Hoffmann, Tamim Asfour, and Stefan
[43] 저자: Peter Pastor, Heiko Hoffmann, Tamim Asfour, 및 Stefan
Schaal. Learning and generalization of motor skills by
learning from demonstration. *2009 IEEE International*
시연으로부터 학습하여 운동 기술을 학습하고 일반화하기.
Conference on Robotics and Automation, pages 763–768,
2009.
2009 IEEE 로봇공학 및 자동화 국제 학술대회, 763-768쪽, 2009.

[44] Dean A. Pomerleau. Alvinn: An autonomous land vehicle
[44] Dean A. Pomerleau. Alvinn: 자율 주행 지상 차량
in a neural network. In *NIPS*, 1988.
신경망에서. NIPS, 1988에서.

[45] Yuzhe Qin, Hao Su, and Xiaolong Wang. From one
[45] 저자: Yuzhe Qin, Hao Su, 및 Xiaolong Wang. 한 손에서
hand to multiple hands: Imitation learning for dexterous
manipulation from single-camera teleoperation. *IEEE*
여러 손으로: 단일 카메라 원격조작을 통한 손재주 조작을 위한 모방학습.
Robotics and Automation Letters, 7:10873–10881, 2022.
IEEE Robotics and Automation Letters 학술지, 7:10873-10881, 2022년.

[46] Rouhollah Rahmatizadeh, Pooya Abolghasemi, Ladislau
[46] 저자: Rouhollah Rahmatizadeh, Pooya Abolghasemi, Ladislau
Bölöni, and Sergey Levine. Vision-based multi-task
manipulation for inexpensive robots using end-to-end
learning from demonstration. *2018 IEEE International*
시연으로부터 엔드투엔드 학습을 사용한 저가 로봇의 비전 기반 다중 작업 조작.
Conference on Robotics and Automation (ICRA), pages
3758–3765, 2017.
2018 IEEE 로봇공학 및 자동화 국제 학술대회(ICRA), 3758-3765쪽, 2017.

[47] Stéphane Ross, Geoffrey J. Gordon, and J. Andrew
[47] 저자: Stéphane Ross, Geoffrey J. Gordon, 및 J. Andrew
Bagnell. A reduction of imitation learning and structured
prediction to no-regret online learning. In *International*
모방학습 및 구조화된 예측을 후회 없는 온라인 학습으로 환원하기.
Conference on Artificial Intelligence and Statistics, 2010.
인공지능 및 통계 국제 학술대회에서, 2010.

[48] Seyed Sina Mirrazavi Salehian, Nadia Figueroa, and Aude
[48] 저자: Seyed Sina Mirrazavi Salehian, Nadia Figueroa, 및 Aude
Billard. A unified framework for coordinated multi-arm
motion planning. *The International Journal of Robotics*
협응 다중 팔 동작 계획을 위한 통합 프레임워크.
Research, 37:1205 – 1232, 2018.
협응 다중 팔 동작 계획을 위한 통합 프레임워크. The International Journal of Robotics Research 학술지, 37:1205 - 1232, 2018년.

[49] Nur Muhammad (Mahi) Shafiullah, Zichen Jeff Cui,
[49] 저자: Nur Muhammad (Mahi) Shafiullah, Zichen Jeff Cui,
Ariuntuya Altanzaya, and Lerrel Pinto. Behavior trans-
formers: Cloning k modes with one stone. *ArXiv*,
행동 트랜스포머: 하나의 돌로 k개 모드를 복제하기.
abs/2206.11251, 2022.
ArXiv, abs/2206.11251, 2022.

[50] Kaushik Shivakumar, Vainavi Viswanath, Anrui Gu,
[50] 저자: Kaushik Shivakumar, Vainavi Viswanath, Anrui Gu,
Yahav Avigal, Justin Kerr, Jeffrey Ichnowski, Richard
Cheng, Thomas Kollar, and Ken Goldberg. Sgtm 2.0:
야하브 아비갈, 저스틴 커, 제프리 이흐노프스키, 리처드 쳉, 토머스 콜라르, 켄 골드버그.
Autonomously untangling long cables using interactive
perception. *ArXiv*, abs/2209.13706, 2022.
Sgtm 2.0: 상호작용 인식을 사용하여 긴 케이블을 자율적으로 풀기. ArXiv, abs/2209.13706, 2022.

[51] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport:
[51] 모히트 스리다르, 루카스 마누엘리, 디터 폭스. Cliport:
What and where pathways for robotic manipulation. *ArXiv*,
로봇 조작을 위한 대상과 위치 경로.
abs/2109.12098, 2021.
ArXiv, abs/2109.12098, 2021.

[52] Mohit Shridhar, Lucas Manuelli, and Dieter Fox.
[52] 모히트 스리다르, 루카스 마누엘리, 디터 폭스.
Perceiver-actor: A multi-task transformer for robotic
manipulation. *ArXiv*, abs/2209.05451, 2022.
Perceiver-actor: 로봇 조작을 위한 다중 작업 트랜스포머. ArXiv, abs/2209.05451, 2022.

[53] Aravind Sivakumar, Kenneth Shaw, and Deepak Pathak.
[53] 아라빈드 시바쿠마르, 케네스 쇼, 디팍 파타크.
Robotic telekinesis: Learning a robotic hand imitator by
watching humans on youtube. *RSS*, 2022.
로봇 염동력: YouTube에서 인간을 관찰하여 로봇 손 모방자를 학습하기. RSS, 2022.

[54] Christian Smith, Yiannis Karayiannidis, Lazaros Nal-
[54] 크리스천 스미스, 이아니스 카라야니디스, 라자로스 날-
pantidis, Xavi Gratal, Peng Qi, Dimos V. Dimarogonas,
and Danica Kragic. Dual arm manipulation - a survey.
판티디스, 자비 그라탈, 펑 치, 디모스 V. 디마로고나스, 다니차 크라기치. 양팔 조작 - 조사.
Robotics Auton. Syst., 60:1340–1353, 2012.
Robotics Auton. Syst., 60:1340-1353, 2012.

[55] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning
[55] 기혁 손, 홍락 리, 신천 옌. 학습
structured output representation using deep conditional
generative models. In *NIPS*, 2015.
심층 조건부 생성 모델을 사용한 구조화된 출력 표현. NIPS에서, 2015.

[56] srcteam. Shadow teleoperation system plays jenga,
[56] srcteam. Shadow 원격조작 시스템이 젠가를 플레이하다,
Mar 2021. URL https://www.youtube.com/watch?v=
2021년 3월. URL https://www.youtube.com/watch?
7K9brH27jvM.
v= 7K9brH27jvM.

[57] srcteam. How researchers are using shadow robot's
[57] srcteam. 연구자들이 Shadow robot의
technology, Jun 2022. URL https://www.youtube.com/
기술을 사용하는 방식, 2022년 6월.
watch?v=p36fYIoTD8M.
URL https://www.youtube.com/ watch? v=p36fYIoTD8M.

[58] srcteam. Shadow teleoperation system, Jun 2022. URL
[58] srcteam. Shadow 원격조작 시스템, 2022년 6월.

https://www.youtube.com/watch?v=cx8eznfDUJA.
URL https://www.youtube.com/watch? v=cx8eznfDUJA. 리뷰. Journal of Sport and Health Science 저널, 2022년.

[59] Simon Stepputtis, Maryam Bandari, Stefan Schaal, and Heni Ben Amor. A system for imitation learning of contact-rich bimanual manipulation policies. *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11810–11817, 2022.
[59] Simon Stepputtis, Maryam Bandari, Stefan Schaal, 및 Heni Ben Amor. 모방학습을 위한 접촉이 풍부한 양손 조작 정책 시스템 저자 Heni Ben Amor. 접촉이 풍부한 양손 조작 정책. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 11810-11817쪽, 2022.

[60] Priya Sundaresan, Jennifer Grannen, Brijen Thananjeyan, Ashwin Balakrishna, Jeffrey Ichnowski, Ellen R. Novoseller, Minho Hwang, Michael Laskey, Joseph Gonzalez, and Ken Goldberg. Untangling dense non-planar knots by learning manipulation features and recovery policies. *ArXiv*, abs/2107.08942, 2021.
[60] 저자 Priya Sundaresan, Jennifer Grannen, Brijen Thananjeyan, Ashwin Balakrishna, Jeffrey Ichnowski, Ellen R. Novoseller, Minho Hwang, Michael Laskey, Joseph Gonzalez, 및 Ken Goldberg. 조작 특징과 복구 정책을 학습하여 조밀한 비평면 매듭 풀기. ArXiv, abs/2107.08942, 2021.

[61] Gokul Swamy, Sanjiban Choudhury, J. Andrew Bagnell, and Zhiwei Steven Wu. Causal imitation learning under temporally correlated noise. In *International Conference on Machine Learning*, 2022.
[61] 저자 Gokul Swamy, Sanjiban Choudhury, J. Andrew Bagnell, 및 Zhiwei Steven Wu. 시간적으로 상관된 잡음하에서의 인과적 모방학습. International Conference on Machine Learning에서, 2022.

[62] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5, 2015.
[62] Naftali Tishby 및 Noga Zaslavsky. 딥러닝과 정보 병목 원리. 2015 IEEE Information Theory Workshop (ITW), 1-5쪽, 2015.

[63] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033, 2012.
[63] Emanuel Todorov, Tom Erez, 및 Yuval Tassa. Mujoco: 모델 기반 제어를 위한 물리 엔진. 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, 5026-5033쪽, 2012.

[64] Stephen Tu, Alexander Robey, Tingnan Zhang, and N. Matni. On the sample complexity of stability constrained imitation learning. In *Conference on Learning for Dynamics & Control*, 2021.
[64] Stephen Tu, Alexander Robey, Tingnan Zhang, 및 N. Matni. 안정성이 제약된 모방학습의 샘플 복잡도. Conference on Learning for Dynamics & Control에서, 2021.

[65] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *ArXiv*, abs/1706.03762, 2017.
[65] 저자 Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, 및 Illia Polosukhin. 주의집중만으로 충분하다. ArXiv, abs/1706.03762, 2017.

[66] Solomon Wiznitzer, Luke Schmitt, and Matt Trossen. interbotix_ros_manipulators. URL https://github.com/Interbotix/interbotix_ros_manipulators.
[66] Solomon Wiznitzer, Luke Schmitt, 및 Matt Trossen. 저장소 interbotix_ros_manipulators. URL https://github.com/ Interbotix/interbotix_ros_manipulators.

[67] Fan Xie, A. M. Masum Bulbul Chowdhury, M. Clara De Paolis Kaluza, Linfeng Zhao, Lawson L. S. Wong, and Rose Yu. Deep imitation learning for bimanual robotic manipulation. *ArXiv*, abs/2010.05134, 2020.
[67] 저자 Fan Xie, A. M. Masum Bulbul Chowdhury, M. Clara De Paolis Kaluza, Linfeng Zhao, Lawson L. S. Wong, and Rose Yu. 양손 로봇 조작을 위한 심층 모방학습. ArXiv, abs/2010.05134, 2020.

[68] Andy Zeng, Peter R. Florence, Jonathan Tompson, Stefan Welker, Jonathan Chien, Maria Attarian, Travis Armstrong, Ivan Krasin, Dan Duong, Vikas Sindhwani, and Johnny Lee. Transporter networks: Rearranging the visual world for robotic manipulation. In *Conference on Robot Learning*, 2020.
[68] 앤디 젱, 피터 R. 플로렌스, 조너선 톰프슨, 슈테판 Welker, Jonathan Chien, Maria Attarian, Travis Armstrong, Ivan Krasin, Dan Duong, Vikas Sindhwani, and Johnny Lee. Transporter 네트워크: 로봇 조작을 위한 시각 세계 재배열. Conference on Robot Learning에서 2020년에 발표됨.

[69] Tianhao Zhang, Zoe McCarthy, Owen Jow, Dennis Lee, Ken Goldberg, and P. Abbeel. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–8, 2017.
[69] 톈하오 장, 조이 매카시, 오언 조, 데니스 리, Ken Goldberg, and P. Abbeel. 가상현실 원격조작을 통한 복잡한 조작 과제의 심층 모방학습. 2018 IEEE International Conference on Robotics and Automation (ICRA), 1-8쪽, 2017.

[70] Allan Zhou, Moo Jin Kim, Lirui Wang, Peter R. Florence, and Chelsea Finn. Nerf in the palm of your hand: Corrective augmentation for robotics via novel-view synthesis. *ArXiv*, abs/2301.08556, 2023.
[70] 앨런 저우, 무진 김, 리루이 왕, 피터 R. 플로렌스, and Chelsea Finn. 손바닥 안의 Nerf: 새로운 시점 합성을 통한 로보틱스 교정 증강. ArXiv, abs/2301.08556, 2023.

[71] Áron Horváth, Eszter Ferentzi, Kristóf Schwartz, Nina Jacobs, Pieter Meyns, and Ferenc Köteles. The measurement of proprioceptive accuracy: A systematic literature
[71] 아론 호르바트, 에스테르 페렌치, 크리스토프 슈바르츠, 니나 Jacobs, Pieter Meyns, and Ferenc Köteles. 고유수용성 정확도의 측정: 체계적 문헌

APPENDIX

부록

A. Comparing ALOHA with Prior Teleoperation Setups

A. ALOHA와 기존 원격조작 설정의 비교

In Figure 9, we include more teleoperated tasks that *ALOHA* is capable of. We stress that all objects are taken directly

그림 9에는 ALOHA가 수행할 수 있는 원격조작 과제를 추가로 포함한다.

from the real world without any modification, to demonstrate *ALOHA*'s generality in real life settings.

ALOHA의 실제 환경에서의 일반성을 입증하기 위해 모든 물체는 어떠한 수정도 거치지 않고 실제 환경에서 직접 가져온 것임을 강조한다.

ALOHA exploits the kinesthetic similarity between leader and follower robots by using joint-space mapping for teleoperation.

ALOHA는 원격조작에 관절 공간 매핑을 사용하여 리더 로봇과 팔로어 로봇 간의 동역학적 유사성을 활용한다.

A leader-follower design choice dates back to at least as far as 1953, when Central Research Laboratories built teleoperation systems for handling hazardous material [27]. More recently,

리더-팔로어 설계 방식은 적어도 1953년까지 거슬러 올라가며, 당시 Central Research Laboratories는 위험 물질을 취급하기 위한 원격조작 시스템을 구축했다 [27].

companies like RE2 [3] also built highly dexterous teleoperation systems with joint-space mapping. *ALOHA* is similar to these

최근에는 RE2 [3]와 같은 기업도 관절 공간 매핑을 적용한 고도의 손재주를 갖춘 원격조작 시스템을 구축했다.

previous systems, while benefiting significantly from recent advances of low-cost actuators and robot arms. It allows us to

ALOHA는 이러한 기존 시스템과 유사하면서도 저비용 액추에이터와 로봇 팔의 최근 발전을 크게 활용한다.

achieve similar levels of dexterity with much lower cost, and also without specialized hardware or expert assembly.

이를 통해 훨씬 낮은 비용으로 유사한 수준의 손재주를 달성할 수 있으며, 특수 하드웨어나 전문적인 조립도 필요하지 않다.

Next, we compare the cost of *ALOHA* to recent teleoperation systems. *DexPilot* [21] controls a dexterous hand using image

다음으로 ALOHA의 비용을 최근 원격조작 시스템과 비교한다.

streams of a human hand. It has 4 calibrated Intel Realsense to

DexPilot [21]는 사람 손의 영상 스트림을 사용하여 손재주를 갖춘 손을 제어한다.

capture the point cloud of a human hand, and retarget the pose to an Allegro hand. The Allegro hand is then mounted to a

이 시스템은 보정된 Intel Realsense 4대를 사용하여 사람 손의 포인트 클라우드를 캡처하고, 손 자세를 Allegro 손에 재지정한다.

KUKA LBR iiwa7 R800. DexPilot allows for impressive tasks

그런 다음 Allegro 손을 KUKA LBR iiwa7 R800에 장착한다.

such as extracting money from a wallet, opening a penut jar, and insertion tasks in NIST board #1. We estimate the system

DexPilot을 사용하면 지갑에서 돈을 꺼내거나, 땅콩 병을 열거나, NIST 보드 #1에서 삽입 작업을 수행하는 등의 인상적인 과제가 가능하다.

cost to be around $100k with one arm+hand. More recent works

팔과 손 1개로 구성된 시스템의 비용은 약 $100k일 것으로 추정한다.

such as Robotic Telekinesis [53, 6, 45] seek to reduce the cost of DexPilot by using a single RGB camera to detect hand pose, and retarget using learning techniques. While sensing

Robotic Telekinesis [53, 6, 45]와 같은 최근 연구는 단일 RGB 카메라로 손 자세를 감지하고 학습 기법으로 이를 재지정하여 DexPilot의 비용을 줄이고자 한다.

cost is greatly reduced, the cost for robot hand and arm remains high: a dexterous hand has more degrees of freedom and is naturally pricier. Moving the hand around would also require

감지 비용은 크게 감소하지만 로봇 손과 팔의 비용은 여전히 높다. 손재주를 갖춘 손은 자유도가 더 많으므로 자연스럽게 더 비싸다.

an industrial arm with at least 2kg payload, increasing the price further. We estimate the cost of these systems to be around

손을 움직이려면 최소 2kg의 페이로드를 지원하는 산업용 팔도 필요하므로 가격이 더욱 상승한다.

$18k with one arm+hand. Lastly, the *Shadow Teleoperation*

이러한 시스템의 비용은 팔과 손 1개로 약 $18k일 것으로 추정한다.

System is a bimanual system for teleoperating two dexterous hands. Both hands are mounted to a UR10 robot, and the hand

마지막으로 Shadow Teleoperation System은 손재주가 뛰어난 두 손을 원격조작하는 양손 원격조작 시스템이다.

pose is obtained by either a tracking glove or a haptic glove.

두 손은 모두 UR10 로봇에 장착되며, 손 자세는 트래킹 장갑 또는 햅틱 장갑으로 얻는다.

This system is the most capable among all aforementioned works, benefitted from its bimanual design. However, it also

이 시스템은 양손 설계 덕분에 앞서 언급한 모든 연구 중 가장 뛰어난 성능을 보인다.

costs the most, at at least $400k. *ALOHA*, on the other hand,

그러나 최소 $400k로 비용도 가장 많이 든다.

is a bimanual setup that costs $18k ($20k after adding optional add-ons such as cameras). Reducing dexterous hands to parallel

반면 ALOHA는 $18k의 비용이 드는 양손 설정이며, 카메라와 같은 선택적 추가 구성품을 더하면 $20k이다.

jaw grippers allows us to use light-weight and low-cost robots, which can be more nimble and require less service.

손재주를 갖춘 손을 평행 죠 그리퍼로 단순화하면 가볍고 저렴한 로봇을 사용할 수 있으며, 이러한 로봇은 더 민첩하고 유지보수도 덜 필요하다.

Finally, we compare the capabilities of *ALOHA* with previous systems. We choose the most capable system as reference: the

마지막으로 ALOHA의 기능을 기존 시스템과 비교한다.

Shadow Teleoperation System [5], which costs more than 10x of *ALOHA*. Specifically, we found three demonstration videos

가장 뛰어난 원격조작 시스템인 Shadow Teleoperation System [5]을 기준으로 삼았으며, 이 시스템의 비용은 ALOHA의 10배를 넘는다.

[56, 57, 58] that contain 15 example use cases of the *Shadow Teleoperation System*, and seek to recreate them using *ALOHA*.

구체적으로 Shadow Teleoperation System을 이용한 15가지 원격조작 사례가 포함된 시연 영상 3개 [56, 57, 58]를 찾아 ALOHA로 이를 재현하고자 했다.

The tasks include playing "beer pong", "jenga," and a rubik's cube, using a dustpan and brush, twisting open a water bottle, pouring liquid out, untying velcro cable tie, picking up an egg and a light bulb, inserting and unplugging USB, RJ45, using a pipette, writing, twisting open an aluminum case, and in-hand rotation of Baoding balls. We are able to recreate 14 out of the

과제에는 "beer pong", "jenga", 루빅스 큐브 놀이, 쓰레받기와 빗자루 사용, 물병 돌려 열기, 액체 따르기, 벨크로 케이블 타이 풀기, 달걀과 전구 집기, USB 및 RJ45 삽입과 분리, 피펫 사용, 글쓰기, 알루미늄 케이스 돌려 열기, 바오딩 공을 손안에서 회전시키기가 포함된다.

15 tasks with similar objects and comparable amount of time.

유사한 물체를 사용하고 비슷한 시간 안에 15개 과제 중 14개를 재현할 수 있었다.

We cannot recreate the Baoding ball in-hand rotation task, as our setup does not have a hand.

우리의 설정에는 손이 없기 때문에 바오딩 공을 손안에서 회전시키는 과제는 재현할 수 없다.

B. Example Image Observations

B. 예시 이미지 관측

We include example image observations taken during policy execution time in Figure 10, for each of the 6 real tasks. From

그림 10에는 6개의 실제 과제 각각에 대해 정책 실행 중 획득한 예시 이미지 관측을 포함한다.

left to right, the 4 images are from top camera, front camera, left wrist, and right wrist respectively. The top and front cameras

왼쪽에서 오른쪽으로 4개의 이미지는 각각 상단 카메라, 전면 카메라, 왼쪽 손목, 오른쪽 손목에서 촬영한 것이다.

are static, while the wrist cameras move with the robots and give detailed views of the gripper. We also rotate the front

상단 카메라와 전면 카메라는 고정되어 있는 반면, 손목 카메라는 로봇과 함께 움직이며 그리퍼를 자세히 보여준다.

camera by 90 degrees to capture more vertical space. For all

또한 더 넓은 수직 영역을 촬영하기 위해 전면 카메라를 90도 회전시킨다.

cameras, the focal length is fixed with auto-exposure on to adjust for changing lighting conditions. All cameras steam at

모든 카메라에서 초점 거리는 고정하고 자동 노출을 켜 변화하는 조명 조건에 맞게 조정한다.

480×640 and 30fps.

모든 카메라는 480 × 640 해상도와 30fps로 스트리밍한다.

C. Detailed Architecture Diagram

C. 상세 아키텍처 다이어그램

We include a more detailed architecture diagram in Figure 11.

그림 11에는 더 상세한 아키텍처 다이어그램을 포함한다.

At training time, we first sample tuples of RGB images and joint positions, together with the corresponding action sequence as prediction target (Step 1: sample data). We then infer style

학습 시에는 먼저 RGB 이미지와 관절 위치의 튜플을 해당 행동 시퀀스와 함께 예측 대상으로 샘플링한다(1단계: 데이터 샘플링).

variable z using CVAE encoder shown in yellow (Step 2: infer z). The input to the encoder are 1) the [CLS] token, which

그런 다음 노란색으로 표시된 CVAE 인코더를 사용하여 스타일 변수 z를 추론한다(2단계: z 추론). 인코더의 입력은 다음과 같다.

consists of learned weights that are randomly initialized, 2)

1) 무작위로 초기화된 학습 가중치로 구성된 [CLS] 토큰,

embedded joint positions, which are joint positions projected to the embedding dimension using a linear layer, 3) embedded

2) 선형 계층을 사용하여 임베딩 차원으로 투영한 관절 위치인 임베딩된 관절 위치,

action sequence, which is the action sequence projected to the embedding dimension using another linear layer. These inputs

3) 또 다른 선형 계층을 사용하여 임베딩 차원으로 투영한 행동 시퀀스인 임베딩된 행동 시퀀스이다.

form a sequence of $(k+2) \times embedding\_dimension$, and is processed with the transformer encoder. We only take the

이러한 입력은 (k + 2) × embedding_dimension의 시퀀스를 이루며, 트랜스포머 인코더로 처리된다.

first output, which corresponds to the [CLS] token, and use another linear network to predict the mean and variance of z's distribution, parameterizing it as a diagonal Gaussian. A

첫 번째 출력만 취하는데, 이는 [CLS] 토큰에 해당하며, 또 다른 선형 네트워크를 사용하여 스타일 변수 z의 분포에 대한 평균과 분산을 예측하고 이를 대각 가우시안으로 매개변수화한다.

sample of z is obtained using reparameterization, a standard way to allow back-propagating through the sampling process so the encoder and decoder can be jointly optimized [33].

z의 샘플은 재매개변수화를 사용하여 얻는다. 재매개변수화는 샘플링 과정을 통한 역전파를 가능하게 하여 인코더와 디코더를 공동으로 최적화할 수 있도록 하는 표준 방법이다 [33].

Next, we try to obtain the predicted action from CVAE decoder i.e. the policy (Step 3: predict action sequence). For

다음으로 CVAE 디코더, 즉 정책에서 예측된 행동을 얻는다(3단계: 행동 시퀀스 예측).

each of the image observations, it is first processed by a ResNet18 to obtain a feature map, and then flattened to get a sequence of features. These features are projected to

각 이미지 관측은 먼저 ResNet18을 통해 처리하여 특징 맵을 얻은 다음, 이를 평탄화하여 특징 시퀀스를 얻는다.

the embedding dimension with a linear layer, and we add a 2D sinusoidal position embedding to perserve the spatial information. The feature sequence from each camera is then

이러한 특징을 선형 계층을 사용하여 임베딩 차원으로 투영하고, 공간 정보를 보존하기 위해 2D 사인파 위치 임베딩을 추가한다.

concatenated to be used as input to the transformer encoder.

그런 다음 각 카메라의 특징 시퀀스를 연결하여 트랜스포머 인코더의 입력으로 사용한다.

Two additional inputs are joint positions and z, which are also projected to the embedding dimension with two linear layers respectively. The output of the transformer encoder are then

추가 입력은 관절 위치와 z이며, 이들 역시 각각 2개의 선형 계층을 사용하여 임베딩 차원으로 투영한다.

used as both "keys" and "values" in cross attention layers of the transformer decoder, which predicts action sequence given encoder output. The "queries" are fixed sinusoidal embeddings

그런 다음 트랜스포머 인코더의 출력을 트랜스포머 디코더의 교차 어텐션 계층에서 "키"와 "값"으로 모두 사용하며, 트랜스포머 디코더는 인코더 출력을 바탕으로 행동 시퀀스를 예측한다.

for the first layer.

첫 번째 계층의 "쿼리"는 고정된 사인파 임베딩이다.

At test time, the CVAE encoder (shown in yellow) is discarded and the CVAE decoder is used as the policy. The

테스트 시에는 CVAE 인코더(노란색으로 표시됨)를 제거하고 CVAE 디코더를 정책으로 사용한다.

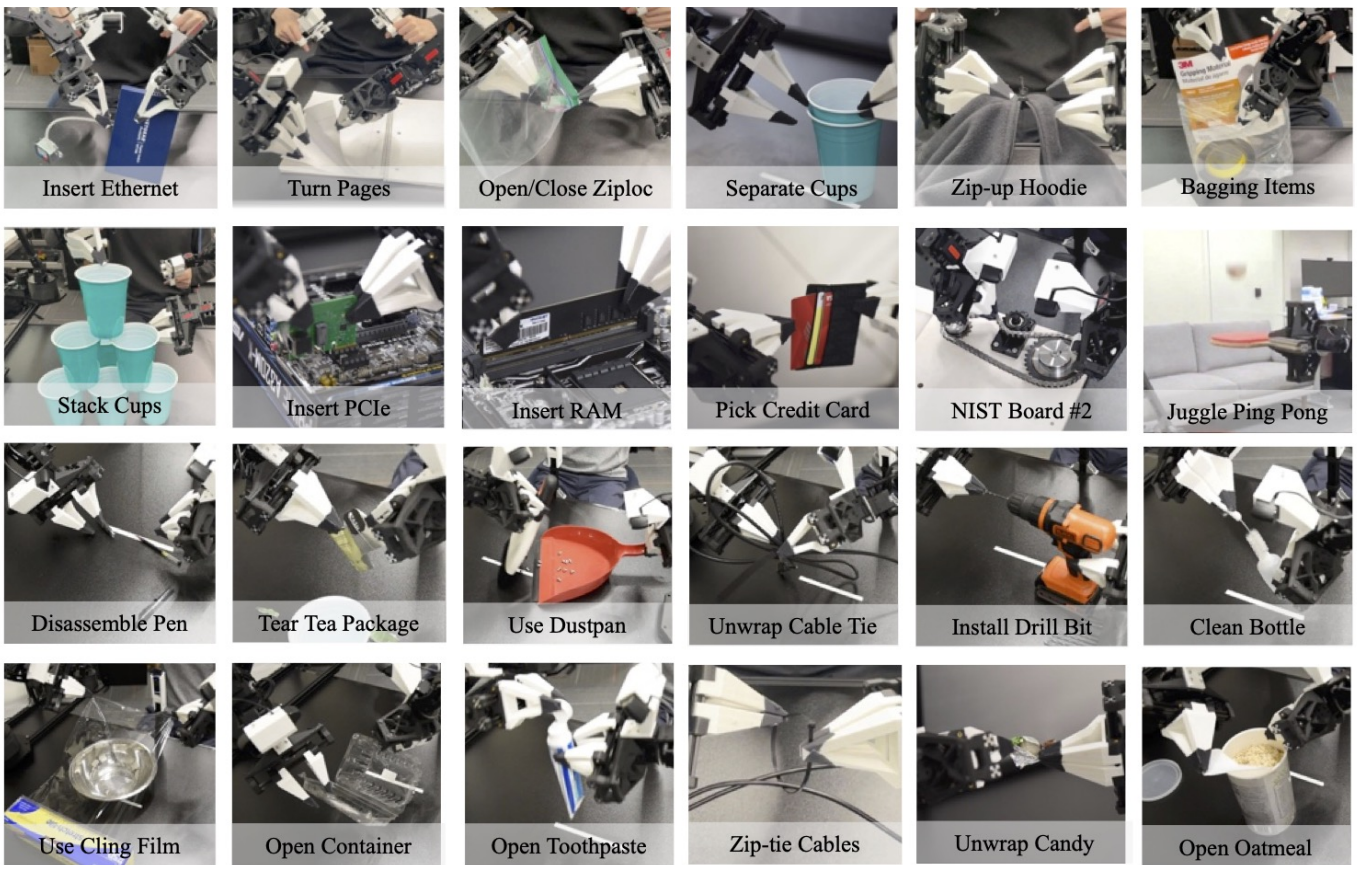


Fig. 9: Teleoperation task examples with *ALOHA*. We include videos on the project website.
그림 9: ALOHA를 사용한 원격조작 작업 예시. 프로젝트 웹사이트에 동영상을 제공한다.

incoming observations (images and joints) are fed into the model in the same way as during training. The only difference
입력되는 관측값(이미지와 관절)은 학습 중과 동일한 방식으로 모델에 입력된다.
is in z, which represents the "style" of the action sequence we want to elicit from the policy. We simply set z to a zero
유일한 차이는 정책에서 유도하려는 "행동 시퀀스"의 "스타일"을 나타내는 z에 있다.
vector, which is the mean of the unit Gaussian prior used during training. Thus given an observation, the output of the
간단히 말해 z을 학습 중 사용한 단위 가우시안 사전분포의 평균인 영벡터로 설정한다.
policy is always deterministic, benefiting policy evaluation.
따라서 관측값이 주어지면 정책의 출력은 항상 결정적이며, 정책 평가에 유리하다.

D. Experiment Details and Hyperparameters
D. 실험 세부 사항 및 하이퍼파라미터

We carefully tune the baselines and include the hyperparameters used in Table III, IV, V, VI, VII. For BeT, we found
기준선의 하이퍼파라미터를 신중하게 조정하고, 사용한 하이퍼파라미터를 표 III, IV, V, VI, VII에 제시한다.
that increasing history length from 10 (as in original paper) to 100 greatly improves the performance. Large hidden dimension
BeT에서는 이력 길이를 10(원 논문과 동일)에서 100으로 늘리면 성능이 크게 향상된다는 사실을 확인했다.
also generally helps. For VINN, the k used when retrieving
큰 은닉 차원도 일반적으로 도움이 된다.
nearest neighbor is adaptively chosen with the lowest validation loss, same as the original paper. We also found that using joint
VINN에서는 최근접 이웃을 검색할 때 사용하는 k를 원 논문과 동일하게 검증 손실이 가장 낮도록 적응적으로 선택한다.
position differences in addition to visual feature similarity improves performance when there is no action chunking, in which case we have state weight = 10 when retrieving actions.
또한 행동 청킹이 없는 경우 시각 특징 유사도에 더해 관절 위치 차이를 사용하면 성능이 향상된다는 사실을 확인했으며, 이 경우 행동을 검색할 때 상태 가중치를 10으로 설정한다.
However, we found this to hurt performance with action chunking and thus set state weight to 0 for action chunking experiments.
그러나 행동 청킹에서는 이 방법이 성능을 저하시킨다는 사실을 확인했으므로 행동 청킹 실험에서는 상태 가중치를 0으로 설정한다.

E. User Study Details
E. 사용자 연구 세부 사항

We conduct the user study with 6 participants, recruited from computer science graduate students, with 4 men and 2 women aged 22-25. 3 of the participants had experience
컴퓨터과학 대학원생 중에서 모집한 참가자 6명(남성 4명, 여성 2명, 연령 22~25세)을 대상으로 사용자 연구를 수행한다.
teleoperating robots with a VR controller, and the other 3 has no prior experience teleoperating. None of the participants used
참가자 중 3명은 VR 컨트롤러를 사용하여 로봇을 원격조작한 경험이 있었고, 나머지 3명은 원격조작 경험이 없었다.
ALOHA before. To implement the 5Hz version of *ALOHA*, we
참가자 중 ALOHA를 이전에 사용한 사람은 없었다.
read from the leader robot at 5Hz, interpolate in the joint space, and send the interpolated positions to the robot at 50Hz. We
ALOHA의 5Hz 버전을 구현하기 위해 리더 로봇의 값을 5Hz로 읽고, 관절 공간에서 보간한 뒤, 보간된 위치를 50Hz로 로봇에 전송한다.
choose tasks that emphasizes high-precision and close-loop visual feedback. We include images of the objects used in
고정밀도와 폐루프 시각 피드백을 강조하는 작업을 선택한다. 그림 12에 사용한 물체의 이미지를 제시한다
Figure 12. For threading zip cable tie, the hole measures 4mm x 1.5mm, and the cable tie measures 0.8mm x 3.5mm with a pointy tip. It is initially lying flat on the table, and the operator needs to pick it up with one gripper, grasp the other end mid-air, then coordinate both hands to insert one end of the cable tie into the hole on the other end. For unstacking cup, we use two single-use plastic cups that has 2.5mm clearance between them when stacked. The teleoperator need to grasp the edge of upper cup, then either shake to separate or use the help from the other gripper. During the user study, we randomize the order in which operators attempt each task, and whether they use 50Hz or 5Hz controller first. We also randomize the initial position of the object randomly around the table center. For each setting, the operator has 2 minutes to adapt, followed by 3 consecutive attempts of the task with duration recorded.

그림 12. 지퍼 케이블 타이를 관통시키는 작업에서 구멍의 크기는 4mm x 1.5mm이고, 케이블 타이의 크기는 뾰족한 끝을 포함하여 0.8mm x 3.5mm이다. 케이블 타이는 처음에 테이블 위에 평평하게 놓여 있으며, 작업자는 한 그리퍼로 이를 집어 들고 다른 쪽 끝을 공중에서 잡은 다음, 양손을 협응하여 케이블 타이의 한쪽 끝을 반대쪽 끝의 구멍에 삽입해야 한다. 컵 분리 작업에서는 쌓았을 때 두 컵 사이에 2.5mm의 여유 공간이 있는 일회용 플라스틱 컵 2개를 사용한다. 원격조작자는 위쪽 컵의 가장자리를 잡은 다음 흔들어 분리하거나 다른 그리퍼의 도움을 받아야 한다. 사용자 연구에서는 작업자가 각 작업을 시도하는 순서와 50Hz 또는 5Hz 컨트롤러 중 어느 것을 먼저 사용하는지를 무작위화한다. 또한 물체의 초기 위치를 테이블 중앙 주변에서 무작위로 정한다. 각 설정에서 작업자는 2분 동안 적응한 후 작업을 3회 연속으로 시도하며, 각 시도의 소요 시간을 기록한다.

F. Limitations
F. 한계

We now discuss limitations of the ALOHA hardware and the policy learning with ACT.
이제 ALOHA 하드웨어와 ACT를 사용한 정책 학습의 한계를 논의한다.

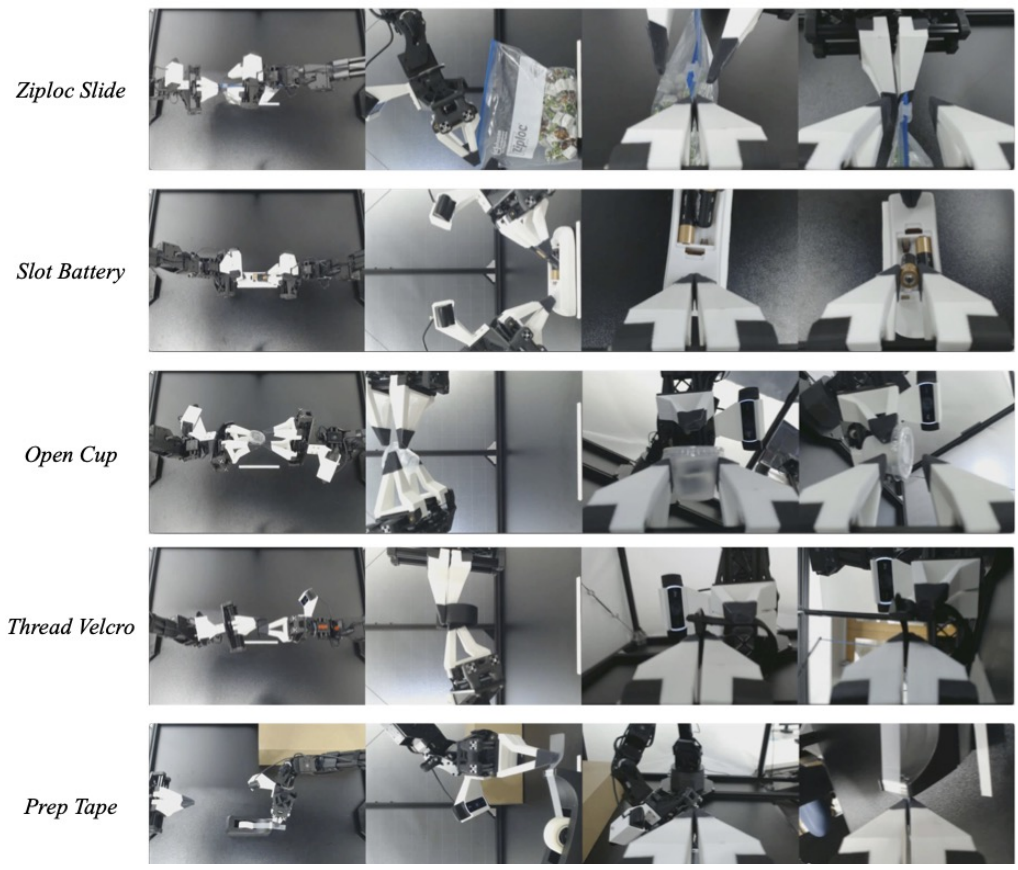


Fig. 10: Image observation examples for 5 real-world tasks. The 4 columns are [top camera, front camera, left wrist camera, right wrist camera] respectively. We rotate the front camera by 90 degree to capture more vertical space.

Hardware Limitations. On the hardware front, *ALOHA* struggles with tasks that require multiple fingers from both hands, for example opening child-proof pill bottles with a push tab. To open the bottle, one hand needs to hold the bottle and pushes down on the push tab, with the other hand twisting the lid open. *ALOHA* also struggles with tasks that require high amount of forces, for example lifting heavy objects, twisting open a sealed bottle of water, or opening markers caps that are tightly pressed together. This is because the low-cost motors cannot generate enough torque to support these manipulations. Tasks that requires finger nails are also difficult for *ALOHA*, even though we design the grippers to be thin on the edge. For example, we are not able to lift the edge of packing tape when it is taped onto itself, or opening aluminum soda cans.

Policy Learning Limitations. On the software front, we report all 2 tasks that we attempted where ACT failed to learn the behavior. The first one is unwrapping candies. The steps involves picking up the candy from the table, pull on both ends of it, and pry open the wrapper to expose the candy. We collected 50 demonstrations to train the ACT policy. In our preliminary evaluation with 10 trials, the policy picks up the candy 10/10, pulls on both ends 8/10, while unwraps the candy 0/10. We attribute the failure to the difficulty of perception and lack of data. Specifically, after pulling the candy on both sides, the seam for prying open the candy wrapper could appear anywhere around the candy. During demonstration collection, it is difficult even for human to discern. The operator needs to judge by looking at the graphics printed on the wrapper and find the discontinuity. We constantly observe the policy trying to peel at places where the seam does not exist. To better track the progress, we attempted another evaluation where we give 10 trials for each candy, and repeat this for 5 candies. For this protocol, our policy successfully unwraps 3/5 candies.

Another task that ACT struggles with is opening a small ziploc bag laying flat on the table. The right gripper needs to first pick it up, adjust it so that the left gripper can grasp firmly on the pulling region, followed by the right hand grasping the other side of the pulling region, and pull it open. Our policy trained with 50 demonstrations can consistently pick up the bag, while having difficulties performing the following 3 mid-air manipulation steps. We hypothesize that the bag is hard to perceive, and in addition, small differences in the pick up position can affect how the bag deforms, and result in large differences in where the pulling region ends up. We believe that pretraining, more data, and better perception are promising directions to tackle these extremely difficult tasks.

그림 10: 실제 환경의 5개 과제에 대한 이미지 관측 예시. 4개 열은 [상단 카메라, 전방 카메라, 왼쪽 손목 카메라, 오른쪽 손목 카메라] 순서입니다. 전방 카메라는 더 많은 수직 공간을 포착하기 위해 90도 회전시켰습니다.

하드웨어 측면에서 ALOHA는 양손의 여러 손가락을 사용해야 하는 과제에 어려움을 겪는다. 예를 들어 사탕 포장지를 뜯을 때 포장지의 이음매가 사탕 주변 어디에나 나타날 수 있다. 시연 데이터를 수집하는 동안, 예를 들어 어린이 보호용 알약 용기를 여는 것처럼 손을 사용하는 과제에서는 이음매를 식별하기가 사람에게조차 어렵다.

조작자는 탭을 눌러야 한다.

용기를 열려면 한 손으로 용기를 잡고 누름 탭을 아래로 누르는 동시에, 다른 손으로 뚜껑을 비틀어 열어야 한다.

또한 ALOHA는 무거운 물체를 들어 올리거나, 밀봉된 물병을 비틀어 열거나, 단단히 눌러 붙어 있는 마커 뚜껑을 여는 등 큰 힘이 필요한 과제에서도 어려움을 겪는다.

이는 저비용 모터가 이러한 조작을 수행할 만큼 충분한 토크를 생성하지 못하기 때문이다.

그립퍼의 가장자리를 얇게 설계했음에도 불구하고 손톱을 사용해야 하는 과제 역시 ALOHA에는 어렵다.

예를 들어 포장 테이프가 자기 자신에 붙어 있을 때 그 테이프의 가장자리를 들어 올리거나 알루미늄 탄산음료 캔을 여는 작업을 수행할 수 없다.

정책 학습의 한계

소프트웨어 측면에서는 ACT가 동작을 학습하는 데 실패한, 우리가 시도한 2개 과제를 모두 보고한다. 첫 번째 과제는 사탕 포장지를 벗기는 것이다.

이 과제는 테이블에서 사탕을 집어 들고, 사탕의 양쪽 끝을 잡아당긴 다음, 포장지를 벌려 사탕을 드러내는 단계로 이루어진다.

ACT 정책을 학습하기 위해 50개의 시연을 수집했다.

10회 시행으로 예비 평가를 수행한 결과, 정책은 10/10회 사탕을 집어 들고, 8/10회 양쪽 끝을 잡아당겼지만, 0/10회 사탕 포장지를 벗겼다.

이러한 실패는 인식의 어려움과 데이터 부족 때문이라고 판단한다.

구체적으로는 사탕의 양쪽을 잡아당긴 후 포장지에 인쇄된 그래픽을 보고 불연속을 찾아야 한다.

정책이 이음매가 존재하지 않는 위치에서 계속 벗기기를 시도하는 현상을 관찰했다.

진행 상황을 더 잘 추적하기 위해 사탕 1개당 10회씩 시행하고 이를 5개의 사탕에 반복하는 추가 평가를 수행했다.

이 평가 방식에서 정책은 3/5개의 사탕 포장지를 성공적으로 벗겼다.

ACT가 어려움을 겪는 또 다른 과제는 테이블 위에 납작하게 놓인 작은 지퍼백을 여는 것이다.

먼저 오른쪽 그립퍼로 지퍼백을 집어 들고, 왼쪽 그립퍼가 당기는 부분을 단단히 잡을 수 있도록 위치를 조정한 다음, 오른손으로 당기는 부분의 반대쪽을 잡아 당겨 열어야 한다.

50개의 시연으로 학습한 정책은 지퍼백을 일관되게 집어 들 수 있지만, 이어지는 3개의 공중 조작 단계를 수행하는 데 어려움을 겪는다.

지퍼백은 인식하기 어렵고, 집어 드는 위치의 작은 차이가 지퍼백의 변형 방식에 영향을 미쳐 당기는 부분의 최종 위치에 큰 차이를 초래한다고 가정한다.

이처럼 매우 어려운 과제를 해결하기 위한 유망한 방향으로 사전 학습, 더 많은 데이터, 향상된 인식 기능이 있다고 생각한다.

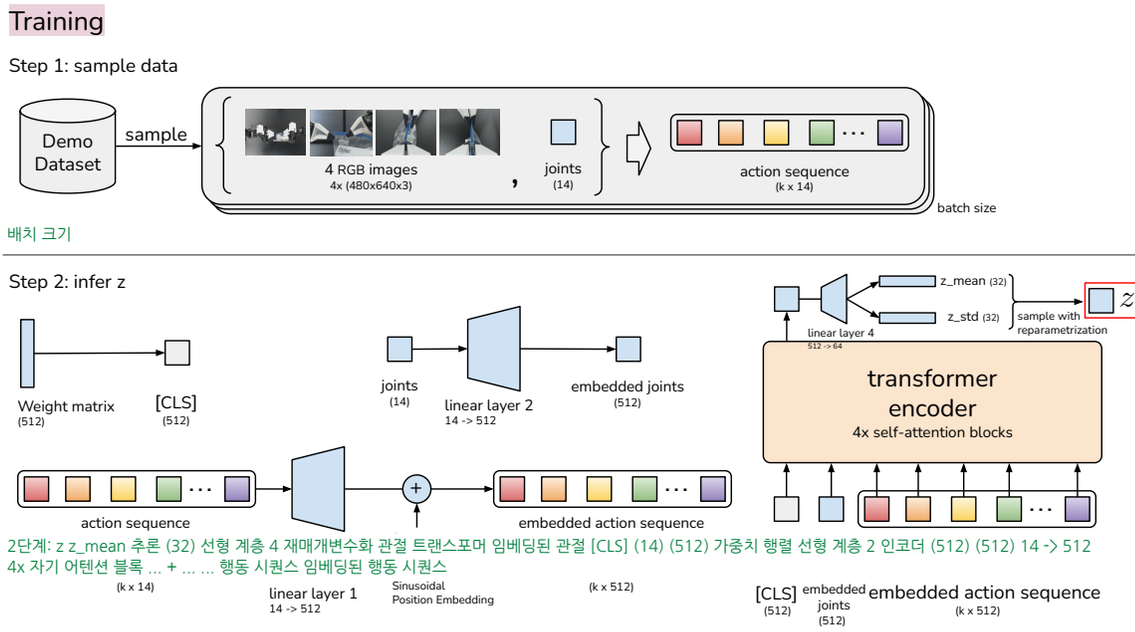


Fig. 11: Detail architecture of Action Chunking with Transformers (ACT).

그림 11: 행동 청킹(Action Chunking with Transformers, ACT)의 상세 아키텍처.

Fig. 12: The cable tie and cups for user study.

그림 12: 사용자 연구에 사용된 케이블 타이와 컵

learning rate	1e-5
batch size	8
# encoder layers	4
# decoder layers	7
feedforward dimension	3200
hidden dimension	512
# heads	8
chunk size	100
beta	10
dropout	0.1

학습률 1e-5 배치 크기 8 인코더 레이어 수 4 디코더 레이어 수 7 피드포워드 차원 3200 은닉 차원 512 헤드 수 8 청크 크기 100 베타 10 드롭아웃 0.1

TABLE III: Hyperparameters of ACT.

표 III: ACT의 하이퍼파라미터

learning rate	3e-4
batch size	128
epochs	100
momentum	0.9
weight decay	1.5e-6

학습률 3e-4 배치 크기 128 에포크 100 모멘텀 0.9 가중치 감쇠 1.5e-6

TABLE IV: Hyperparameters of BYOL, the feature extractor for VINN and BeT.

표 IV: VINN과 BeT의 특징 추출기인 BYOL의 하이퍼파라미터

learning rate	1e-4
batch size	64
# layers	6
# heads	6
hidden dimension	768
history length	100
weight decay	0.1
offset loss scale	1000
focal loss gamma	2
dropout	0.1
discretizer #bins	64

학습률 1e-4 배치 크기 64 레이어 수 6 헤드 수 6 은닉 차원 768 이력 길이 100 가중치 감쇠 0.1 오프셋 손실 스케일 1000 초점 손실 감마 2 드롭아웃 0.1 이산화기 빈 수 64

TABLE V: Hyperparameters of BeT.

표 V: BeT의 하이퍼파라미터

k (nearest neighbour)	adaptive
state weight	0 or 10

k (최근접 이웃) 적응형 상태 가중치 0 또는 10

TABLE VI: Hyperparameters of VINN.

표 VI: VINN의 하이퍼파라미터

learning rate	1e-5
batch size	2
ViT dim head	32
ViT window size	7
ViT mbconv expansion rate	4
ViT mbconv shrinkage rate	0.25
ViT dropout	0.1
RT-1 depth	6
RT-1 heads	8
RT-1 dim head	64
RT-1 action bins	256
RT-1 cond drop prob	0.2
RT-1 token learner num output tokens	8
weight decay	0
history length	6

학습률 1e-5 배치 크기 2 ViT 헤드 차원 32 ViT 윈도 크기 7 ViT mbconv 확장률 4 ViT mbconv 축소율 0.25 ViT 드롭아웃 0.1 RT-1 깊이 6 RT-1 헤드 수 8 RT-1 헤드 차원 64 RT-1 행동 빈 수 256 RT-1 조건 삭제 확률 0.2 RT-1 토큰 학습기 출력 토큰 수 8 가중치 감쇠 0 이력 길이 6

TABLE VII: Hyperparameters of RT-1.

표 VII: RT-1의 하이퍼파라미터