Check for updates

특집논문 (Special Paper)
방송공학회논문지 제30권 제4호, 2025년 7월 (JBE Vol.30, No.4, July 2025)
https://doi.org/10.5909/JBE.2025.30.4.589
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

로봇 조작 정책 학습에서 다중 시점 구성 변화의 정량적 영향 분석

김 도 원[a], 이 상 민[a], 박 성 용[a], 김 희 원[a)‡]

Quantitative Analysis of Multi-Viewpoint Configuration Effects in Robotic Manipulation Policy Learning

Dowon Kim[a], Sangmin Lee[a], Sungyong Park[a], and Heewon Kim[a)‡]

요 약

로봇 조작(Robotic Manipulation) 정책 학습에서 관측 시점의 수와 구성은 폐색 및 정보 손실을 줄이고, 정책의 예측 정확도에 중요한 영향을 미치는 핵심 요소이다. 본 연구는 단순히 다중 시점을 입력으로 사용하는 것에 그치지 않고, 시점 수와 구성 방식이 정책 학습과 일반화에 미치는 영향을 정량적으로 분석하고자 한다. Isaac SIM 시뮬레이션 환경과 PerAct 모델을 기반으로, 동일한 조작 과제를 유지한 상태에서 입력 시점 수와 구성을 변화시키며 정책을 fine-tuning하였다. 이후 과제 성공률 및 예측 위치의 정확도를 기준으로 정책 성능을 비교하였다. 실험 결과, 일부 2-view 구성은 5-view 전체 구성과 유사하거나 더 나은 성능을 보였으며, 단순한 시점 수 증가보다 시점 간 정보 정합성이 성능에 더 큰 영향을 미칠 수 있음을 시사한다. 본 연구는 과제 특성에 맞춘 선택 전략과 정책-연계 최적화 등 향후 실환경 로봇 시스템 설계에 실질적인 기준과 기초 자료를 제공할 수 있다.

Abstract

In robotic manipulation policy learning, the number and configuration of observation viewpoints are critical factors that influence policy performance by mitigating occlusion and reducing information loss. This study goes beyond simply increasing the number of input views and quantitatively analyzes how different combinations of observation viewpoints impact both policy learning and generalization. Using the Isaac SIM simulation environment and the voxel-based PerAct model, we fine-tune policies with varying viewpoint configurations while keeping the manipulation task and model architecture fixed. Policy performance is evaluated in terms of task success rate and 3D grasp prediction accuracy. Experimental results show that certain 2-view configurations outperform or match the full 5-view baseline, while others with more views lead to degraded performance due to redundant or conflicting observations. These findings highlight that viewpoint selection based on task relevance and information complementarity is more important than mere quantity. This work provides practical insights and design guidelines for viewpoint-efficient policy optimization in real-world robotic systems.

Keyword : Robotic Manipulation, Viewpoint Selection, Voxel-based Policy Learning

Copyright ⓒ 2025 Korean Institute of Broadcast and Media Engineers. All rights reserved.
“This is an Open-Access article distributed under the terms of the Creative Commons BY-NC-ND (http://creativecommons.org/licenses/by-nc-nd/3.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited and not altered.”

Ⅰ. 서 론

로봇 조작 정책 학습은 복잡한 환경에서 언어 명령을 이해하고 물리적 상호작용을 수행하는 핵심 기술[3,12,22,26]이다. 특히 시각 기반 정책 학습은 고차원 센서 입력(RGB-D 영상 등)을 통해 환경의 상태를 인식하고 적절한 행동을 선택한다. 이러한 연구는 전통적으로 단일 관측 시점의 카메라 입력에 의존하여 제한된 시야 내에서 객체의 상태를 인식하고 목표 행동을 예측하도록 설계되었다. 그러나 실제 환경[8]에서는 시야 폐색, 부분 관측, 다중 객체 상호작용 등으로 인해 단일 관측 시점 기반 정책은 불안정한 동작을 보이거나 일반화에 실패하는 경우가 빈번하다. 이를 보완하기 위해 최근에는 다중 시점 RGB-D 입력을 활용하여 공간 정보를 보다 완전하게 수집하고, 이를 통합적으로 활용하는 방식이 주목받고 있다.

예를 들어, PerAct[10]는 다중 시점의 RGB-D 영상 3D voxel 표현으로 변환하여, Transformer 기반의 정책 네트워크를 통해 복잡한 물체 조작 과제에서 우수한 성능을 입증하였다. Voxel 기반 표현은 시각 정보를 공간적으로 정합된 3차원 형태로 표현할 수 있어, 객체의 위치와 상태를 보다 정밀하게 파악할 수 있다는 강점을 지닌다.

한편, 최근의 Vision-Language-Action(VLA) 모델들[13-15]은 대규모 인터넷 데이터로 사전 학습된 Vision-Language Model을 기반으로 범용 조작 정책을 학습하는 방향으로 발전하고 있다. RT-2[13], Octo[14], OpenVLA[15] 등의 모델들은 다양한 로봇 플랫폼과 환경에서 일반화 가능한 정책을 학습하며, 언어 조건과 시각 정보를 통합하여 복잡한 조작 과제를 수행할 수 있는 능력을 보여주고 있다.

이러한 기술적 발전에도 불구하고, 대부분의 연구에서 모델의 성능 향상은 주로 아키텍처 개선이나 학습 데이터 확대[2]에 초점을 맞추고 있으며, 입력 시점의 선택과 구성이 성능에 미치는 영향은 체계적으로 분석되지 않았다. 더욱이, 실제 로봇 시스템 구축 시에는 하드웨어 제약, 비용, 실시간 처리 요구사항 등을 고려해야 하는데, 현재의 연구들은 이러한 실용적 측면을 충분히 다루지 못하고 있다. 예를 들어, 5개의 카메라를 모두 설치하고 동기화하는 것은 상당한 비용과 복잡성을 수반한다.

이러한 한계에 따라 본 연구는 다음과 같은 핵심 질문을 제기한다. “어떤 시점이 정책 성능에 더 큰 기여를 하며, 다중 시점 구성은 예측 성능과 일반화 능력에 어떤 영향을 미치는가?” 이 질문에 대한 답을 구하기 위해, 본 연구는 Isaac Sim 환경에서 PerAct[10] 기반의 조작 정책 모델을 기반으로 시점 구성이 조작 정책 성능에 미치는 영향을 체계적으로 분석한다. 5개의 고정 카메라(Front, Base, Left, Wrist_Bottom, Wrist) 중 다양한 시점 조합을 구성하여, 각 구성의 성공률, 예측 정확도, 일반화 성능에 미치는 영향을 정량적으로 평가한다.

본 연구의 주요 기여는 다음과 같다:

① 관측 시점 구성의 체계적 분석: 다양한 관측 시점의 구성에 대해 포괄적인 성능 비교를 통해, 시점 수와 성능 간의 관계가 단순한 선형 관계가 아님을 보인다.

② 효율적 시점 구성의 발견: 특정 2-view 구성(예: Front+Wrist_Bottom)이 5-view 전체 구성과 동등하거나 더 나은 성능을 보일 수 있음을 실험적으로 증명한다.

③ 일반화 성능 분석: Novel object/scene/state 조건에서의 평가를 통해, 과도한 시점 정보가 일반화 성능을 저해할 수 있음을 보여준다.

본 연구의 결과는 효율적인 로봇 시스템 설계를 위한 기초 자료로 활용될 수 있으며, 향후 과제별 적응적 시점 선택, 능동적 시점 제어 등의 고급 연구로 확장될 수 있을 것으로 기대한다.

a) 숭실대학교 글로벌미디어학부(Global School of Media, Soongsil University)
‡ Corresponding Author : 김희원(Heewon Kim)
E-mail: hwkim@ssu.ac.kr
Tel: +82-2-820-0679
ORCID: https://orcid.org/0000-0001-7777-9823
※ This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the National Program for Excellence in SW (2024-0-00071), the Graduate School of Metaverse Convergence support program (IITP-2025-RS-2024-00430997), and the Convergence security core talent training business support program (IITP-2025-RS-2024-00426853) supervised by the IITP (Institute of Information & communications Technology Planning & evaluation). This work was supported by the Technology Development Program (RS-2024-00510957) funded by the Ministry of SMEs and Startups (MSS, Korea)
· Manuscript May 26, 2025; Revised July 4, 2025; Accepted July 4, 2025.

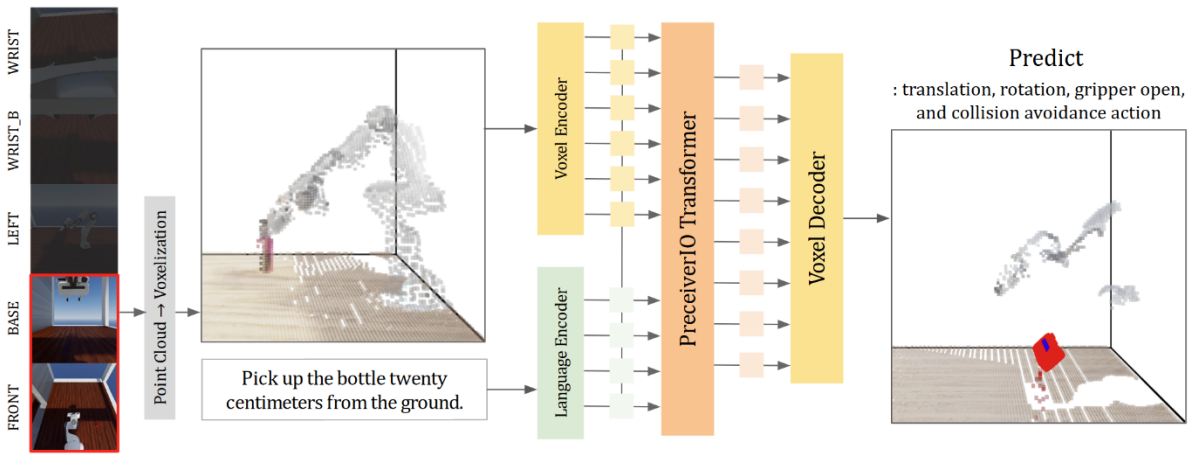

그림 1. 특정 시점 구성을 적용한 PerAct 기반 실험 설정 개요. 예시에서는 사용 가능한 다섯 개의 시점 중 FRONT와 BASE 카메라로부터의 RGB-D 이미지를 선택하여 하나의 포인트 클라우드로 통합한 뒤, 복셀화 과정을 거쳐 3D 격자 형태로 변환된다. 해당 voxel grid는 언어 명령("Pick up the bottle twenty centimeters from the ground")과 함께 PerceiverIO 기반 Transformer에 입력되며, 모델은 로봇의 grasp 위치, 자세, 그리퍼 개폐 명령과 함께, 환경 내 물리적 장애물과의 충돌을 방지하는 회피 동작이 포함된 3D 행동을 예측한다. 예측된 위치는 빨간 voxel로 시각화되어 있다.
Fig. 1. Overview of the experimental setup using the PerAct architecture with a specific viewpoint configuration. In this example, RGB-D images from the FRONT and BASE cameras are selected from the five available viewpoints and fused into a unified point cloud, which is then voxelized. The voxel grid, together with a language instruction ("Pick up the bottle twenty centimeters from the ground"), is processed by a PerceiverIO-based Transformer. The model outputs a predicted 3D action, represented by the red voxel, which encodes the robot's grasp position, orientation, gripper command, and adjustments to avoid physical collisions with obstacles in the environment.

II. 관련 연구

1. 복셀 기반 조작 정책과 3D 표현 학습

복잡한 로봇 조작 환경에서의 정책 학습은 단순한 2D 이미지 기반 표현을 넘어, 3D 공간 정보를 적극적으로 활용하는 방향으로 발전하고 있다. 특히 voxel 기반의 3D 표현은 RGB-D 이미지로부터 생성된 point cloud를 3차원 voxel grid로 변환함으로써, 객체의 위치와 형태를 공간적으로 정합된 형태로 표현할 수 있는 장점을 제공한다[10,16].

대표적인 예로, PerAct[10]는 다중 RGB-D 시점으로부터 생성된 point cloud를 단일 voxel grid로 통합한 후, 언어 명령과 함께 이를 입력으로 받아 PerceiverIO 기반의 구조를 통해 3D 공간 상의 grasp 위치, 자세, 그리퍼 명령 등 복합적인 조작 행동을 예측한다. 이러한 구조는 기존의 2D 픽셀 기반 정책보다 높은 공간 정합성을 가지며, 다중 시점 정보의 통합을 통해 폐색 문제나 시야 제한 등으로 인한 정보 불완전성을 일부 보완할 수 있다.

또한, Transporter-3D[16]는 keypoint-based attention과 point cloud 기반의 위치 매칭을 결합하여, 복셀 표현 없이도 3D 공간 상에서 pick-and-place 위치를 효율적으로 예측하는 구조를 제안하였다. 이 외에도 다양한 point cloud 기반 조작 정책 연구들이 존재하나[2,4-6,8], 대부분 사전 정의된 고정 시점 구성에 의존하며, 시점 수나 시점의 구성 방식이 정책 성능에 미치는 영향에 대한 체계적인 분석은 미비하다.

Perceiver-Actor[10]나 Habitat 기반 embodied AI 플랫폼[1] 또한 유사한 3D 표현 기반의 정책 학습을 위한 구조로 제공하지만, 대부분 고정된 시점 구성에 기반한 실험으로 구성되어 있으며, 시점 수 또는 시점의 구성 변화가 정책 성능에 미치는 민감도 분석은 상대적으로 제한적이다.

2. 범용 로봇 조작 정책과 Vision-Language 기반 학습

로봇 정책 일반화와 관련된 최근 연구들은, 다양한 로봇 플랫폼과 조작 과제에 공통으로 적용 가능한 범용 조작 정

책 학습에 집중하고 있다. 이들 접근법은 일반적으로 대규모 시각-언어 데이터셋과 사전 학습된 비전-언어 모델(vision-language models, VLMs)[11,22,25,27]을 활용하여, 다수의 조작 태스크를 단일 정책 구조에서 학습하고 일반화하는 방식을 채택한다.

예를 들어, RT-2[13]는 웹 스케일의 비전-언어 데이터를 활용해 사전 학습된 모델을 기반으로, 자연어 명령과 시각 입력을 받아 로봇 행동을 직접 예측하는 Vision-Language-Action(VLA) 모델을 제안하였다. 이 모델은 실제 로봇 제어에 VLM의 범용 지식을 효과적으로 전이할 수 있음을 보여주었다. Octo[14]는 다양한 로봇 형태에 걸쳐 수집된 멀티 태스크 조작 데이터를 기반으로, 파라미터 공유를 통해 로봇 간 정책 이식 및 일반화가 가능한 구조를 구현하였다.

최근에는 OpenVLA[15]와 같이 사전 학습된 시각 백본(SigLIP[17], DINOv2[18] 등)과 언어 모델(LLaMA2[19] 등)을 고정한 상태로 활용하면서, 조작 행동을 언어 토큰 시퀀스[20-24]로 표현하여 통합적으로 모델링하는 경량화된 정책 구조도 제안되고 있다. OpenVLA는 end-to-end 방식의 효율적 fine-tuning이 가능하며, 실제 로봇 실험에서 높은 성능을 달성하였다.

그러나 이들 대부분의 모델은 단일 시점 또는 고정된 시야에서 수집된 이미지 입력을 기반으로 학습되며, 입력 시점 수, 시야각 변화, 시점 구성 방식이 정책 학습과 성능에 미치는 영향을 직접적으로 다루지 않는다. 이는 시점 구성 요소의 제어 및 분석 가능성을 제한하며, 로봇의 센서 배열 설계나 관측 전략이 정책 성능에 핵심적으로 작용할 수 있음에도 불구하고, 이에 대한 체계적인 논의는 여전히 부족하다.

Ⅲ. 방법론

1. 실험 목적 및 문제 정의

본 연구는 로봇 조작 정책 학습에서 입력으로 사용되는 관측 시점의 수와 구성 방식이 정책 성능에 미치는 정량적 영향을 분석하는 것을 목적으로 한다. 이를 위해 voxel 기반 조작 정책 모델인 PerAct를 기반으로 하여, 동일한 모델 구조를 유지한 상태에서 관측 시점의 다양한 구성을 적용하여 fine-tuning하고, 성능 변화를 평가하였다. 이는 다중 시점 입력이 항상 성능 향상으로 이어진다는 일반적 가정을 재검토하고, 정보 정합성과 시점 선택 전략의 중요성을 실험적으로 규명하기 위한 시도이다.

2. 시점 구성 및 실험 조건

본 연구에서는 Arnold[7] 데이터에서 제공하는 총 5개의 RGB-D 시점을 기반으로, 관측 시점 수와 구성에 따른 정책 성능의 차이를 비교 실험하였다. 실험에 사용된 관측 시점의 세부 사항은 Figure 2 및 Table 1에 정리되어 있다. 5개의 시점으로 사전 학습된 Baseline 모델을 출발점으로 하여, 각 시점의 구성별로 fine-tuning을 진행하였다. 즉, 동일한 초기 모델을 기반으로 관측 시점의 구성만을 달리하여 fine-tuning함으로써, 시점 구성의 영향만을 분리해 평가할 수 있도록 실험을 설계하였다.

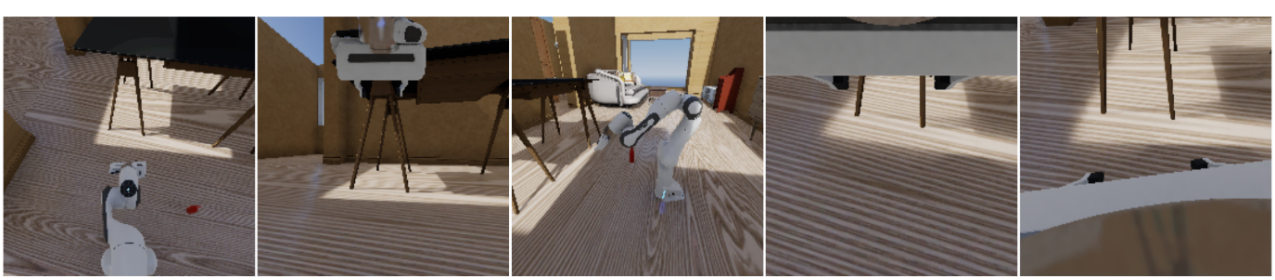

그림 2. 본 실험에 사용된 Arnold Challenge 환경의 RGB-D 시점 구성
Fig. 2. RGB-D viewpoints used in our experiments, captured from the Arnold Challenge environment

표 1. 실험에 사용된 시점 구성 조건 요약
Table 1. Summary of Viewpoint Configurations Used in the Experiments

Configuration Name	Viewpoints Used	# of Views	Description
5-view (baseline)	front, base, left, wrist_bottom, wrist	5	Uses all available viewpoints in Arnold Challenge
1-view	front	1	Single front-facing viewpoint
Ablation	front, wrist_bottom	2	Alternative 2-view configuration for analyzing view importance
2-view	front, base	2	Combines frontal and base-angle views
3-view	front, base, wrist_bottom	3	Adds a wrist-downward view to 2-view configuration

3. 평가 지표

시점 구성에 따른 정책 성능은 다음의 두 가지 정량적 지표를 기반으로 평가하였다:

- 성공률(Success Rate): 각 에피소드에서 정책이 주어진 조작 목표를 성공적으로 완료한 비율. 성공은 로봇이 객체를 올바르게 잡아 들어올렸으며, 이때 예측 위치가 ground truth와의 거리 기준값 이하인 경우로 정의된다.
- 행동 위치 예측 오차(Grasp Prediction Error): 정책이 첫 번째로 예측한 행동 지점을 3D 공간상에서 실제 목표 위치와 비교하여 산출한 유클리드 거리이다. 각 테스트 에피소드에서 산출된 예측 위치 오차의 유클리드 거리를 평균하여 최종 지표로 사용하였다.

Ⅳ. 실험 및 결과

1. 실험 설정

본 연구에서는 Arnold[7] 데이터에서 제공되는 총 5개의 RGB-D 시점을 기반으로, 관측 시점 수 및 구성에 따른 정책 성능을 정량적으로 비교하였다. 5개의 view 입력으로 사전 학습된 Baseline 모델을 출발점으로 하여, 각 시점 조합별로 fine-tuning을 진행하였다. 동일한 초기 모델을 기반으로 관측 시점만을 변경하여 fine-tuning을 수행함으로써, 시점 구성의 효과를 독립적으로 비교할 수 있도록 설계하였다.

본 실험은 Isaac Sim 물리 시뮬레이터에서 구현된 로봇 조작 환경에서 수행되었다. Isaac Sim은 NVIDIA의 고성능 로봇 시뮬레이션 플랫폼으로, 실제와 유사한 물리 엔진과 정밀한 센서 모델링을 제공한다.

학습 및 평가는 다양한 조작 시나리오와 객체 구성을 포함한 Arnold[7] 데이터셋을 기반으로 수행되었다. 특히, 평가 데이터에는 일반화 능력을 평가하기 위한 Novel object/scene/state, any state가 포함되어 있다. 본 연구에서는 이 중 pickup_object 과제를 선택하였으며, 이는 로봇이 언어 명령에 따라 다양한 객체를 지정된 위치에서 집어 올리는 작업이다.

각 카메라는 128×128 해상도의 RGB-D 영상을 제공하며, depth 정보는 카메라 내외부 파라미터를 이용해 3D point cloud로 변환된다. 다중 시점의 point cloud는 100×100×100 크기의 단일 voxel grid로 통합된다. 공정한 비교를 위해 학습률(5e-4), 배치 크기(2), 옵티마이저(Adam) 등 모든 하이퍼파라미터는 고정하였으며, 오직 관측 시점 구성만을 변경하여 실험을 진행하였다.

Baseline은 5개의 관측 시점 구성을 입력으로 사용하여 50,000 iteration 동안 학습된 PerAct[10] 모델로 설정하였다.

2. 관측 시점 구성에 따른 학습 성능 분석

본 실험에서는 학습 중 관측 시점 구성만을 변경하고, 평가 단계에서는 모든 조건에 대해 동일한 5개의 시점 입력을 적용하였다. 이는 PerAct의 본래 구조와 동작 방식을 유지하면서, 학습 단계에서의 관측 시점의 구성 차이가 정책의 일반화 성능에 미치는 영향을 독립적으로 분석하기 위함이다. 즉, 관측 시점 수보다는 정보 구성의 질적 차이가 정책 표현력과 일반화에 어떤 영향을 미치는지를 중점적으로 평

가한다.

일반적으로 학습 시 더 많은 관측 시점을 활용하면, 더 풍부한 3D 관찰 정보를 voxel 공간에 제공함으로써 정책의 예측 정확도를 향상시킬 수 있을 것으로 기대된다. 그러나 학습 시에 관측 시점 구성을 달리한 실험 결과(Table 2)는 이러한 기대가 항상 유효하지 않음을 보여준다. 예를 들어, 학습에 사용된 시점 구성 중 front, base, wrist_bottom의 3-view 구성은 상대적으로 더 많은 관측 정보를 포함하고 있음에도 불구하고, 평가 시 예측 오차가 가장 크게 나타났으며(3.98cm), 성공률 또한 baseline보다 낮은 수준(0.80)에 머물렀다. 이는 관측 시점 간 시야가 보완적이지 않거나, voxel 통합 과정에서 상충하는 정보가 포함될 경우, 오히려 정보 간 간섭이나 노이즈로 인해 정책 성능이 저하될 수 있음을 시사한다. 이러한 결과는 단순히 관측 시점 수를 늘리는 것이 항상 성능 향상을 보장하지 않음을 보여주며, 시점 구성 전략이 정책 학습에 중요한 요소임을 실증적으로 입증한다. 즉, 학습 단계에서의 시점 구성은 단순한 입력 풍부성보다도, 정보 구조의 질적 설계가 정책의 표현력과 일반화 성능에 더욱 결정적인 영향을 미친다.

표 2. 학습 시점 구성만을 변경하고, 평가 단계에서는 모든 조건에 대해 5-view 입력을 공통으로 사용한 실험 결과
Table 2. Experimental results with fixed 5-view evaluation for all conditions

Viewport	Success Rate (avg)	Mean Distance (avg, cm)	Fine-tuned
default (5-view)	0.85	1.21	✗
5-view	0.90	1.17	✓
front	0.95	1.34	✓
front + base	0.80	1.13	✓
front + wrist	0.90	1.17	✓
front + base + wrist	0.80	3.98	✓

반대로, front, base와 front, wrist_bottom의 2-view 구성은 더 적은 입력만을 사용하면서도 5-view 입력과 유사하거나 더 나은 성능을 보였다. 이는 grasp 위치 주변의 정밀한 정보를 제공하는 wrist_bottom 시점과, 장면 전반을 조망할 수 있는 front 시점이 상호보완적으로 작용한 결과로 해석된다. 이 구성은 성공률(0.90)과 예측 오차(1.1745cm) 모두에서 우수한 성능을 기록하였으며, 이는 불필요한 시점을 제거하고 정보적으로 핵심적인 시점만을 선택하는 전략이 입력 효율성과 정책 안정성 측면에서 오히려 유리할 수 있음을 보여준다. 이러한 결과는 향후 과제 중심의 시점 선택 전략이 정책 성능 향상에 있어 중요한 설계 요소가 될 수 있음을 시사한다.

3. 학습–평가 시점 구성 일치 조건에서의 정책 성능 분석

앞선 실험에서는 학습 시점 구성만을 변경하고, 평가 단계에서는 모든 조건에서 고정된 5-view 입력을 적용하여 정책 성능을 비교하였다. 그러나 해당 실험 설정은 평가 시 항상 전체 시점을 사용하는 구조로, 실제 로봇 시스템에서 제한된 센서 입력 조건에서의 실행 성능을 반영하기에는 한계가 있다. 이에 따라 본 실험에서는 학습에 사용된 시점 구성과 동일한 입력만을 평가 단계에서도 사용하는 방식으로 실험 설정을 변경하였다. 이를 통해 관측 시점의 구성 방식이 PerAct 기반 정책의 표현 학습과 실행 단계 성능에 미치는 영향을 보다 정밀하게 분석하고자 하였다.

Table 3은 각 시점 구성별 fine-tuning 여부와 평균 성공률을 정리한 결과이다. 실험 결과, 학습과 평가 모두에서 단일 시점인 front만을 사용한 구성에서 가장 높은 성공률(0.90)을 기록하였다. 이는 제한된 관측 시점 입력만으로도 정책이 높은 실행 성능을 유지할 수 있으며, 복잡한 시점 통합이 반드시 필요하지 않음을 시사한다.

2-view 및 3-view 구성은 모두 평균 성공률 0.75를 기록하여, 관측 시점의 수 증가가 반드시 정책 성능 향상으로 이어지지 않음을 시사한다. 특히 front, base, wrist 구성은 가장 많은 입력 정보를 사용했음에도 불구하고, 단일 시점

표 3. 학습과 평가 모두 동일한 시점 구성으로 수행된 실험 결과
Table 3. Policy performance when the same viewpoint configuration is used for both training and evaluation

Viewport	Success Rate (avg)	Fine-tuned
5-view	0.90	✓
front	0.90	✓
front + base	0.75	✓
front + wrist	0.75	✓
front + base + wrist	0.75	✓

구성보다 낮은 성능을 보였다. 이는 시점 간 정보 통합 과정에서 발생할 수 있는 중복 또는 상충된 시각 정보, 그리고 학습 집중도의 분산이 정책 성능에 부정적인 영향을 미쳤을 가능성을 보여준다.

이러한 결과는 시점 수의 단순한 확대가 성능 향상을 보장하지 않으며, 오히려 과제에 적합한 시점 구성을 선택하는 것이 정책 학습과 실행 모두에 있어 핵심적인 설계 요소임을 강조한다.

4. 시점 구성에 따른 일반화 성능 분석

본 실험에서는 학습 시점 구성의 차이가 정책의 일반화 성능에 미치는 영향을 분석하기 위한 것으로, 이를 위해 평가 단계에서는 모든 입력 구성에 대해 동일한 5-view 입력을 사용하였다. 성능 평가는 novel object, novel scene, novel state, any state의 네 가지 일반화 조건에서 성공률을 기준으로 수행되었다(Table 4).

표 4. 다양한 시점 구성에서 학습된 정책의 일반화 성능 비교
Table 4. Generalization performance of each policy trained under different viewpoint configurations

Viewport	Novel Object	Novel Scene	Novel State	Any State
5-view	0.75	0.80	0.00	0.40
front	0.80	0.90	0.00	0.45
front + base	0.75	0.85	0.00	0.50
front + wrist_b	0.75	0.85	0.0	0.4
front + base + wrist_b	0.8	0.75	0.0	0.35

주목할 만한 결과는, front 단일 시점으로 학습된 모델이 novel object(0.8) 및 novel scene(0.9) 조건에서 가장 높은 성공률을 기록했다는 점이다. 이는 과도한 정보 없이도 일관된 표현 학습이 가능하며, 효과적인 일반화가 가능함을 시사한다. 실제로 any state 조건에서도 front 단일 시점 구성은 5-view보다 높은 성공률인 0.45를 나타냈다.

front, base의 2-view 구성은 any state 조건에서 가장 높은 성공률(0.50)을 나타냈으며, front, wrist_bottom 구성 역시 novel 조건 전반에서 안정적인 성능을 유지하였다. 이들은 모두 2개 시점 구성으로, 복잡도는 낮지만 상호보완적인 정보가 주어져 정보 다양성과 학습 안정성 사이에서 균형을 이룬 결과로 해석된다.

반면, 가장 많은 입력 시점이 사용된 front, base, wrist_bottom의 3-view 구성은 any state 조건에서 가장 낮은 성공률인 0.35를 기록하였다.

종합하면, 입력 시점 수보다는 과제 특성에 부합하는 시점 선택과 구성 전략이 모델의 일반화 성능에 더 중요한 영향을 미치는 것으로 나타났다. 이러한 결과는 향후 로봇 시스템에서 센서 자원 효율화 및 시점 구성 기반 정책 최적화를 위한 설계 기준으로 활용될 수 있다.

V. 결 론

본 연구는 로봇 조작 정책 학습에서 관측 시점 수와 구성 방식이 정책의 예측 정확도 및 일반화 성능에 미치는 영향을 분석하였다. PerAct 모델을 기반으로 Arnold Challenge 환경에서 다양한 시점 조합을 실험적으로 구성하였으며, 학습-평가 시점 조건의 일치 여부에 따라 여러 실험을 수행하였다. 각 시점 구성에 대해 성공률과 예측 오차를 측정하여 시점 구성의 효과를 정량적으로 비교하였다.

실험 결과, 일부 핵심 시점만으로도 전체 5-view 시점과 유사하거나 더 나은 성능을 보이는 경우가 확인되었다. 이는 관측 시점 선택이 단순히 입력 수의 문제가 아니라, 센서 자원 효율성, 연산 비용, 정책 안정성 측면에서 중요한 설계 변수로 작용함을 시사한다.

최근의 범용 정책 모델들이 시점 구성의 영향을 거의 고려하지 않는 점을 감안할 때, 본 연구는 입력 시점의 구성 조합이 정책 성능에 미치는 영향을 체계적으로 실증했다는 점에서 의의가 있다. 향후 연구는 과제별 시점 구성 최적화, 능동적인 시점 선택 및 이동식 센서 제어, 실환경 센서 제약 하의 시점 설계 전략 등으로 확장될 수 있다.

참 고 문 헌(References)

[1] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E Wijmans, and B. Jain, "Habitat: A Platform for Embodied AI Research," *IEEE/CVF International Conference on Computer Vision*, pp. 9338-9346, 2019.

doi: https://doi.org/10.1109/iccv.2019.00943

[2] S. Lee, S. Park, and H. Kim, “DynScene: Scalable Generation of Dynamic Robotic Manipulation Scenes for Embodied AI,” Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 12166-12175, 2025.

[3] K. Zheng, X. Chen, O. Jenkins, and X. E. Wang, “VLMbench: A Compositional Benchmark for Vision-and-Language Manipulation,” Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2022.
doi: https://doi.org/10.48550/arXiv.2206.08522

[4] E. Chisari, N. Heppert, M. Argus, T. Welschehold, T. Brox, and A. Valada, “Learning Robotic Manipulation Policies from Point Clouds with Conditional Flow Matching,” Conference on Robot Learning, 2024.
doi: https://doi.org/10.48550/arXiv.2409.07343

[5] H. Huang, H. Liu, D. Wang, R. Walters, and R. Platt, “MATCH POLICY: A Simple Pipeline from Point Cloud Registration to Manipulation Policies,” arXiv preprint arXiv:2409.15517, 2025.
doi: https://doi.org/10.48550/arXiv.2409.15517

[6] H. Geng, Z. Li, Y. Geng, J. Chen, H. Dong, and H. Wang, “PartManip: Learning Cross-Category Generalizable Part Manipulation Policy from Point Cloud Observations,” arXiv preprint arXiv:2303.16958, 2023.
doi: https://doi.org/10.48550/arXiv.2303.16958

[7] R. Gong, J. Huang, Y. Zhao, H. Geng, X. Gao, and Q. Wu, “ARNOLD: A Benchmark for Language-Grounded Task Learning With Continuous States in Realistic 3D Scenes,” *IEEE/CVF International Conference on Computer Vision*, pp. 20426-20438, 2023.
doi: https://doi.org/10.1109/ICCV51070.2023.01873.

[8] H. Zhu, Y. Wang, D. Huang, W. Ye, W. Ouyang, and T. He, “Point Cloud Matters: Rethinking the Impact of Different Observation Spaces on Robot Learning,” arXiv preprint arXiv:2402.02500, 2024.
doi: https://doi.org/10.48550/arXiv.2402.02500

[9] S. Choi, S. Park, and H. Kim, “SIDL: A Real-World Dataset for Restoring Smartphone Images with Dirty Lenses,” Proceedings of the AAAI Conference on Artificial Intelligence, 39(3), 2545-2554, 2025
doi: https://doi.org/10.1609/aaai.v39i3.32257

[10] M. Shridhar, L. Manuelli, and D. Fox, “Perceiver-actor: A Multi-Task Transformer for Robotic Manipulation,” *Conference on Robot Learning*, pp. 785-799, 2023.
doi: https://doi.org/10.48550/arXiv.2209.05451

[11] S. Karamcheti, S. Nair, A. Balakrishna, P. Liang, T. Kollar, and D. Sadigh, “Prismatic VLMs: Investigating the Design Space of Visually-Conditioned Language Models,” International Conference on Machine Learning, 2024.
doi: https://doi.org/10.48550/arXiv.2402.07865

[12] Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, and L. Fan, “VIMA: General Robot Manipulation with Multimodal Prompts,” International Conference on Machine Learning, 2023.
doi: https://doi.org/10.48550/arXiv.2210.03094

[13] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, X. Chen, K. Choromanski, T. Ding, D. Driess, A. Dubey, C. Finn, P. Florence, C. Fu, M. G. Arenas, K. Gopalakrishnan, K. Han, K. Hausman, A. Herzog, J. Hsu, B. Ichter, A. Irpan, N. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, I. Leal, L. Lee, T. E. Lee, S. Levine, Y. Lu, H. Michalewski, I. Mordatch, K. Pertsch, K. Rao, K. Reymann, M. Ryoo, G. Salazar, P. Sanketi, P. Sermanet, J. Singh, A. Singh, R. Soricut, H. Tran, V. Vanhoucke, Q. Vuong, A. Wahid, S. Welker, P. Wohlhart, J. Wu, F. Xia, T. Xiao, P. Xu, S. Xu, T. Yu, and B. Zitkovich, "RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control,” arXiv preprint arXiv:2307.15818, 2023.
doi: https://doi.org/10.48550/arXiv.2307.15818

[14] D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees, S. Dasari, J. Hejna, C. Xu, J. Luo, T. Kreiman, Y. L. Tan, L. Y. Chen, P. Sanketi, Q. Vuong, T. Xiao, D. Sadigh, C. Finn, and S. Levine, “Octo: An Open-Source Generalist Robot Policy,” Proceedings of Robotics: Science and Systems, 2024.
doi: https://doi.org/10.48550/arXiv.2405.12213

[15] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi, Q. Vuong, T. Kollar, B. Burchfiel, R. Tedrake, D. Sadigh, S. Levine, P. Liang, and C. Finn, “OpenVLA: An Open-Source Vision-Language-Action Model,” arXiv preprint arXiv:2406.09246, 2024.
doi: https://doi.org/10.48550/arXiv.2406.09246

[16] A. Zeng, P. Florence, J. Tompson, S. Welker, J. Chien, M. Attarian, T. Armstrong, I. Krasin, D. Duong, V. Sindhwani, and J. Lee, “Transporter Networks: Rearranging the Visual World for Robotic Manipulation,” Conference on Robot Learning, 2020.

[17] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, “Sigmoid Loss for Language Image Pre-Training,” Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11975-11986, 2023.
doi: https://doi.org/10.48550/arXiv.2303.15343

[18] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P. Huang, S. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, “DINOv2: Learning Robust Visual Features without Supervision,” arXiv preprint arXiv:2304.07193, 2024.
doi: https://doi.org/10.48550/arXiv.2304.07193

[19] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, “Llama 2: Open Foundation and Fine-Tuned Chat Models,” arXiv preprint arXiv:2307.09288, 2023.
doi: https://doi.org/10.48550/arXiv.2307.09288

[20] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," Proceedings of the 39th International Conference on Machine Learning, pp. 12888-12900, 2022.
doi: https://doi.org/10.48550/arXiv.2201.12086

[21] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," arXiv preprint arXiv:2301.12597, 2023.
doi: https://doi.org/10.48550/arXiv.2301.12597

[22] C. Lynch and P. Sermanet, "Language Conditioned Imitation Learning over Unstructured Data," arXiv preprint arXiv:2005.07648, 2021.
doi: https://doi.org/10.48550/arXiv.2005.07648

[23] H. Tan and M. Bansal, "LXMERT: Learning Cross-Modality Encoder representations from Transformers," arXiv preprint arXiv:1908.07490, 2019.
doi: https://doi.org/10.48550/arXiv.1908.07490

[24] H. Laurençon, L. Saulnier, L. Tronchon, S. Bekman, A. Singh, A. Lozhkov, T. Wang, S. Karamcheti, A. M. Rush, D. Kiela, M. Cord, and V. Sanh, "OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents," arXiv preprint arXiv:2306.16527, 2023.
doi: https://doi.org/10.48550/arXiv.2306.16527

[25] X. Chen, X. Wang, L. Beyer, A. Kolesnikov, J. Wu, P. Voigtlaender, B. Mustafa, S. Goodman, I. Alabdulmohsin, P. Padlewski, D. Salz, X. Xiong, D. Vlasic, F. Pavetic, K. Rong, T. Yu, D. Keysers, X. Zhai, and R. Soricut, "PaLI-3 Vision Language Models: Smaller, Faster, Stronger," arXiv preprint arXiv:2310.09199, 2023.
doi: https://doi.org/10.48550/arXiv.2310.09199

[26] S. Nair, E. Mitchell, K. Chen, B. Ichter, S. Savarese, and C. Finn, "Learning Language-Conditioned Robot Behavior from Offline Data and Crowd-Sourced Annotation," Proceedings of the 5th Conference on Robot Learning, pp. 1303-1315, 2022.
doi: https://doi.org/10.48550/arXiv.2109.01115

[27] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved Baselines with Visual Instruction Tuning," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26296-26306, 2024.
doi: https://doi.org/10.48550/arXiv.2310.03744

저 자 소 개

김 도 원

- 2020년 ~ 현재 : 숭실대학교 글로벌미디어학부 학사과정
- ORCID : https://orcid.org/0009-0007-8246-5467
- 주관심분야 : 컴퓨터비전, 머신러닝, 로보틱스

이 상 민

- 2018년 ~ 2024년 : 숭실대학교 글로벌미디어학부 학사
- 2024년 ~ 현재 : 숭실대학교 미디어학과 석사과정
- ORCID : https://orcid.org/0009-0007-1713-5197
- 주관심분야 : 컴퓨터비전, 머신러닝, 로보틱스

박 성 용

- 2016년 ~ 2024년 : 숭실대학교 전자정보공학부 학사
- 2024년 ~ 현재 : 숭실대학교 미디어학과 석사과정
- ORCID : https://orcid.org/0009-0006-3818-592X
- 주관심분야 : 컴퓨터비전, 인공지능

저 자 소 개

김 희 원

- 2008년 ~ 2014년 : 서울대학교 전기·정보공학부 학사
- 2017년 ~ 2023년 : 서울대학교 전기·정보공학부 박사
- 2023년 ~ 현재 : 숭실대학교 글로벌미디어학부 조교수
- ORCID : https://orcid.org/0000-0001-7777-9823
- 주관심분야 : 컴퓨터비전, 머신러닝, 로보틱스