# R+X: Retrieval and Execution from Everyday Human Videos

**RSS 2024**

**Project:** robot-learning.uk/r-plus-x

Jishnu P
Reading Group | IRVL
1/31/25

# Authors

Georgios Papagiannis*

Norman Di Palo*

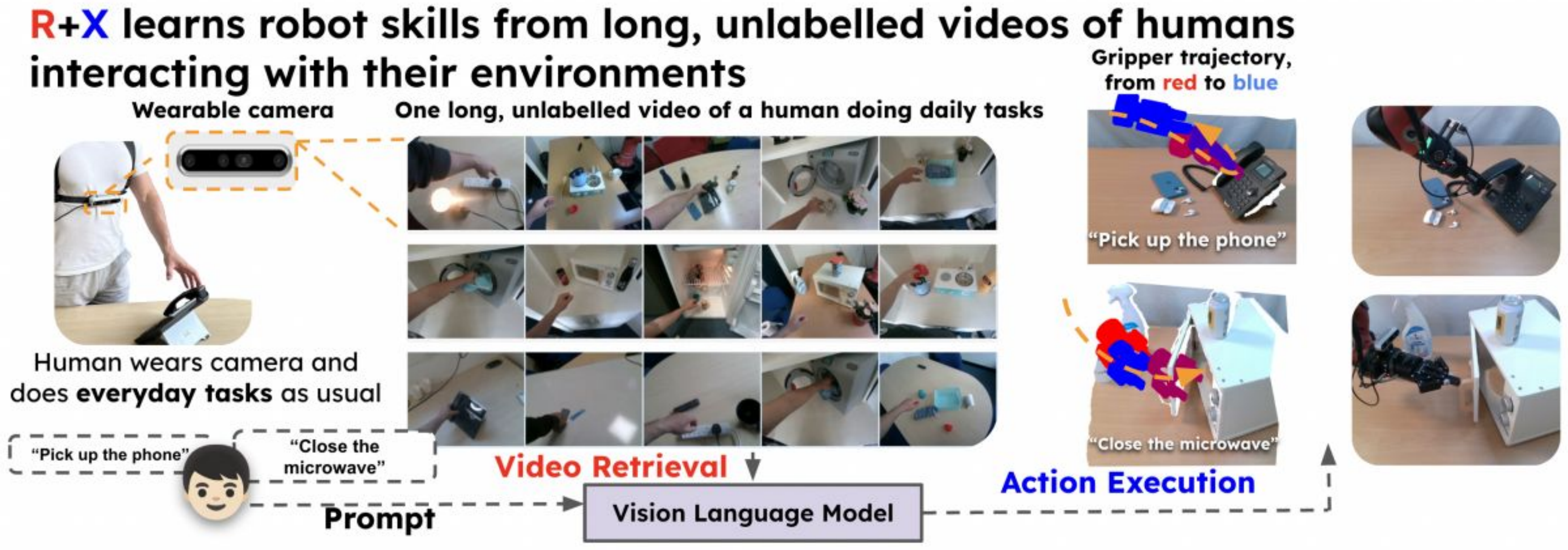
Pietro Vitiello

Edward Johns

The Robot Learning Lab

Imperial College London

# Problem



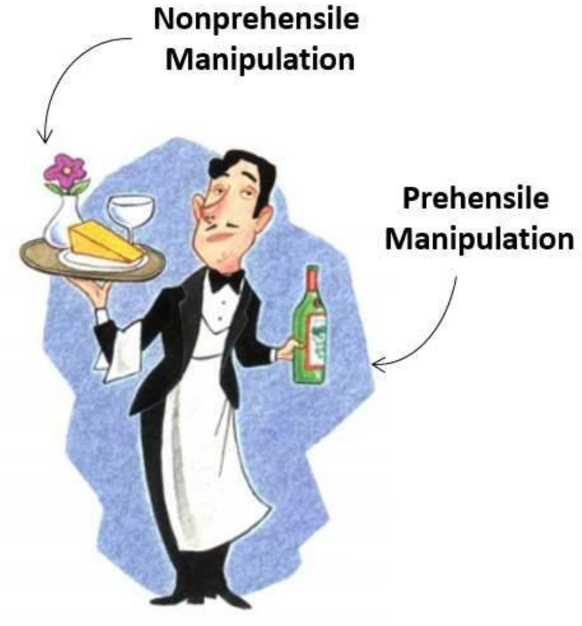

Leverage understanding of large models

- Via video retrieval and understanding
- No Finetuning

Few-Shot In-Context Imitation Learning

# Related Works

| | Multi task no label/align videos | No robot data | Non-prehensile tasks | New obj gener. | Distractors (both train & test) | No MoCap hardware |
|---|---|---|---|---|---|---|
| Vid2Robot | ✗ | ✗ | ✓ | ✓ | ✓ | ✓ |
| WHIRL | ✗ | ✗ | ✓ | ✗ | ✓ | ✓ |
| DITTO | ✗ | ✓ | ✗ | ✗ | ✗ | ✓ |
| ScrewMimic | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ |
| Orion | ✗ | ✓ | ✗ | ✗ | ✗ | ✓ |
| DexCap | ✗ | ✓ | ✓ | ✓ | ✓ | ✗ |
| **R+X** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |



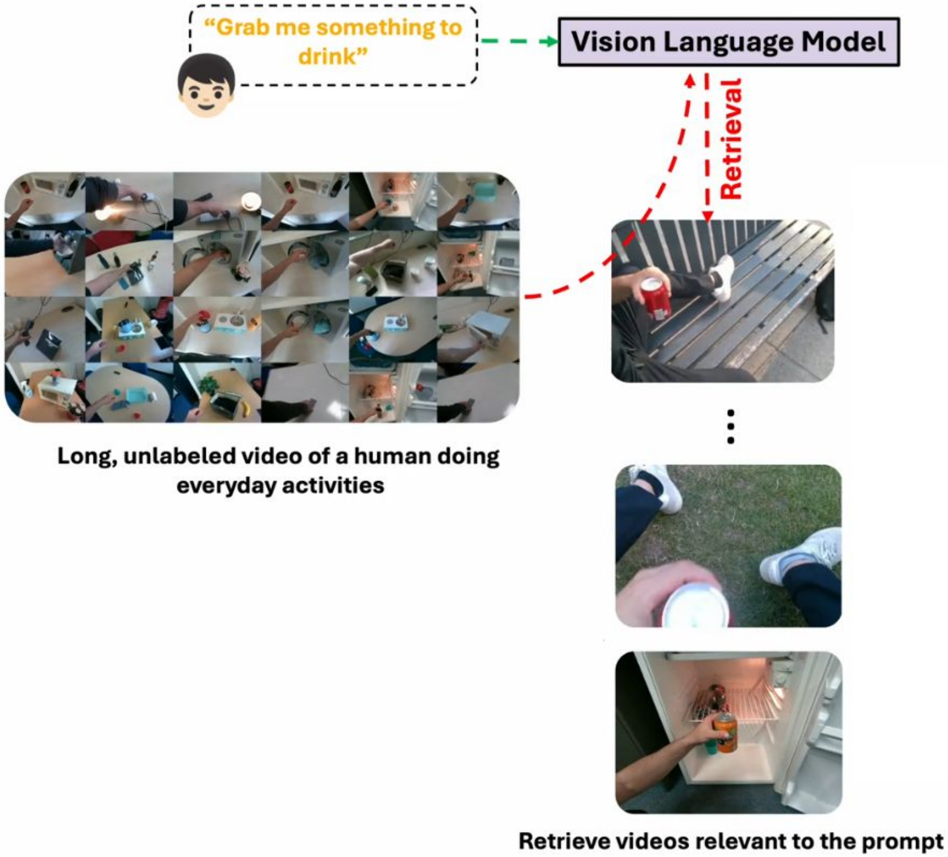
# 1. Get Videos: Record Anywhere, from Multiple Views

Long, unlabeled video of a human doing everyday activities

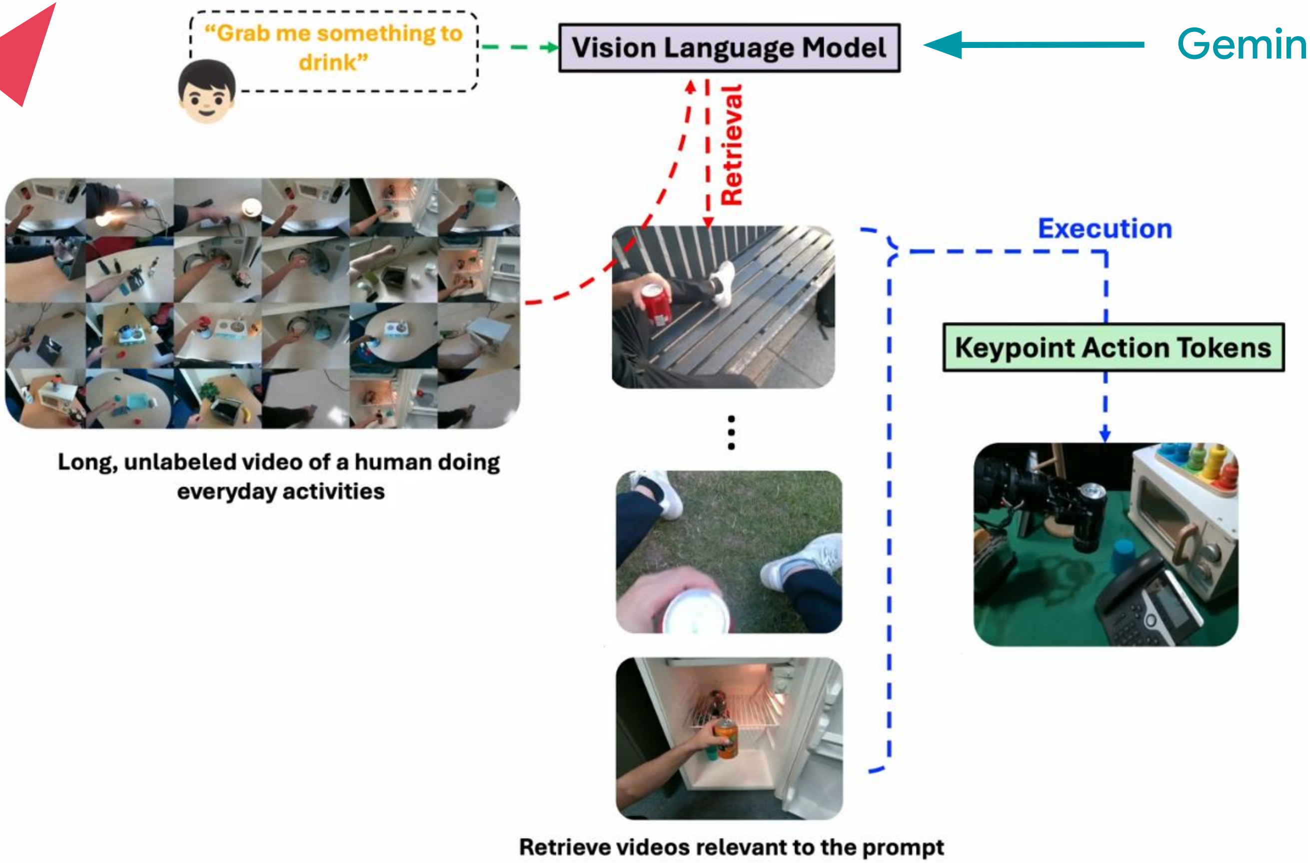
- Multiple rooms, multiple buildings, and even outside
- Chest camera, head camera or a third person camera


Single Unlabelled Video
with less
clutter/distractors


**Long, unlabeled video of a human doing everyday activities**

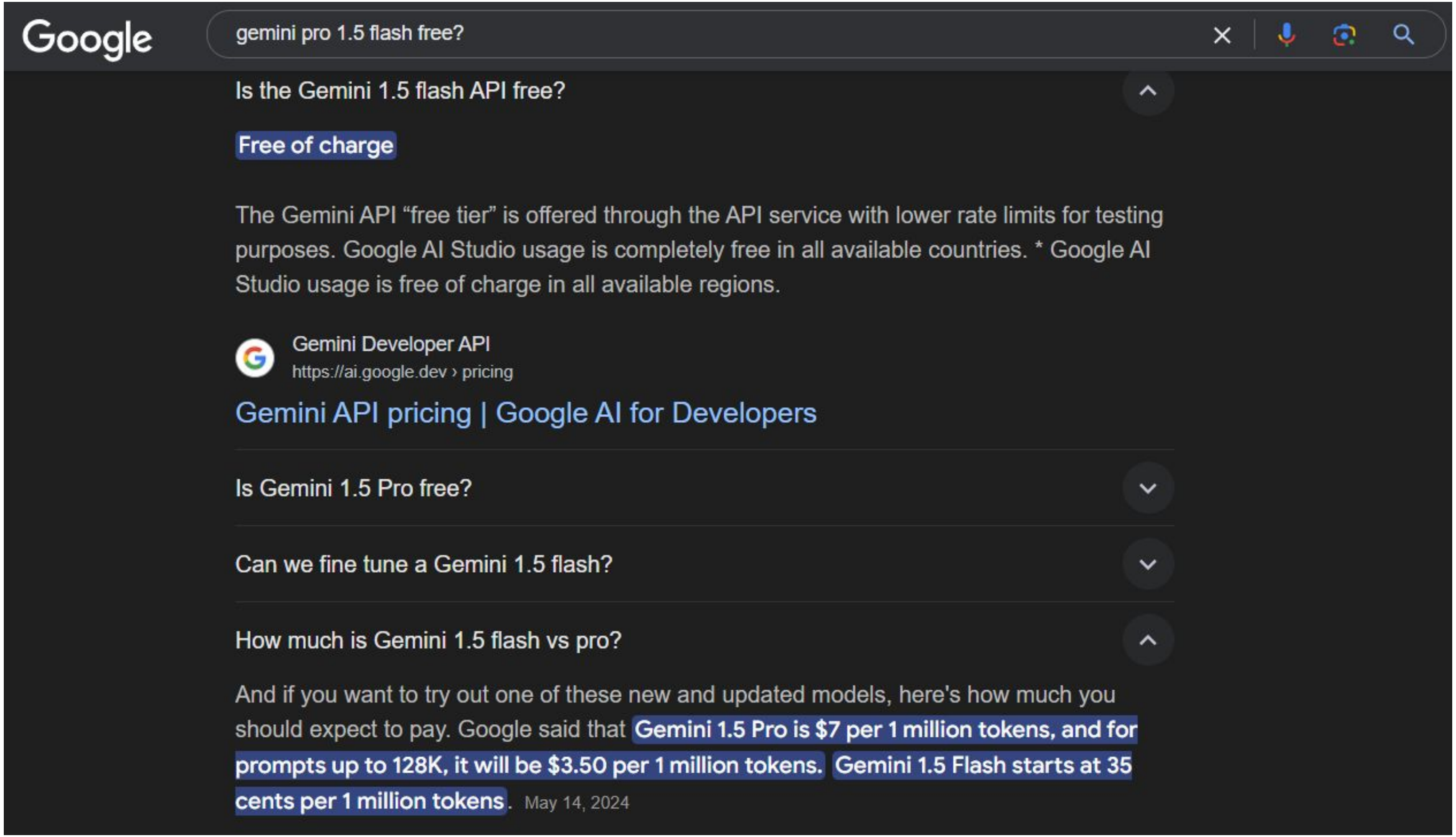

"Grab me something to drink"
Vision Language Model
Retrieval
Long, unlabeled video of a human doing everyday activities
Retrieve videos relevant to the prompt

# R+X : Retrieval and Execution




Gemini Pro 1.5 flash

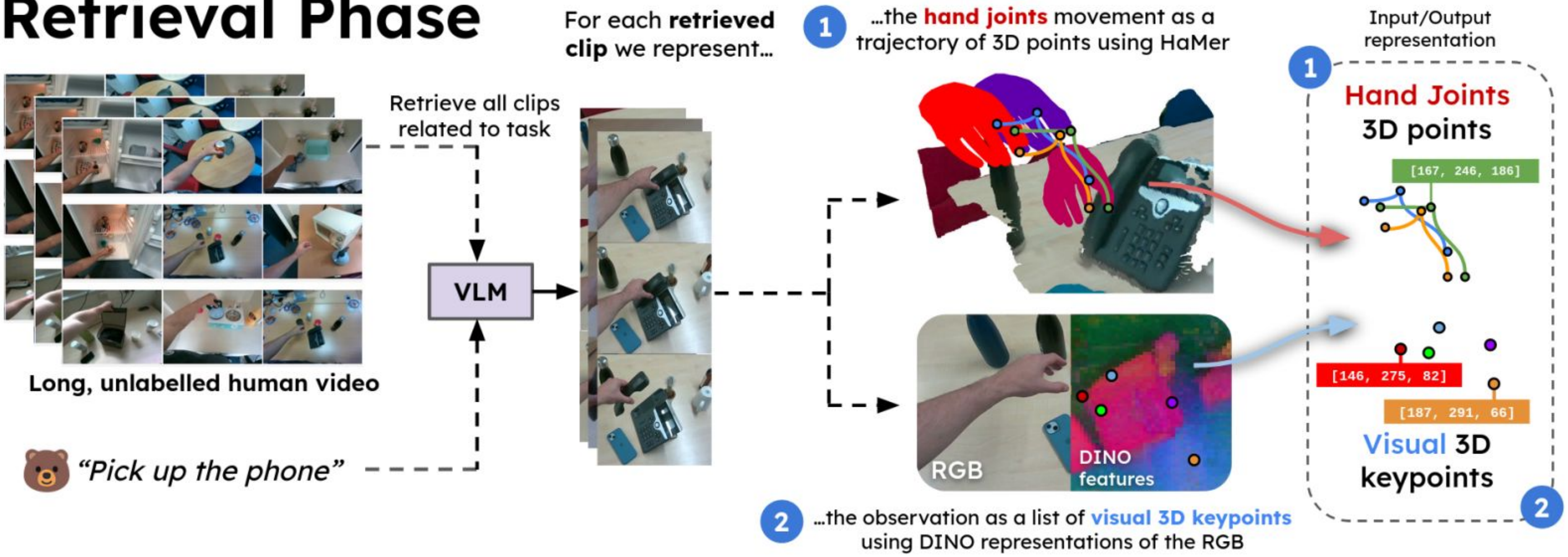
# Google Search as of 1/31/2025



# Deploy Immediately to Novel Environments and Objects

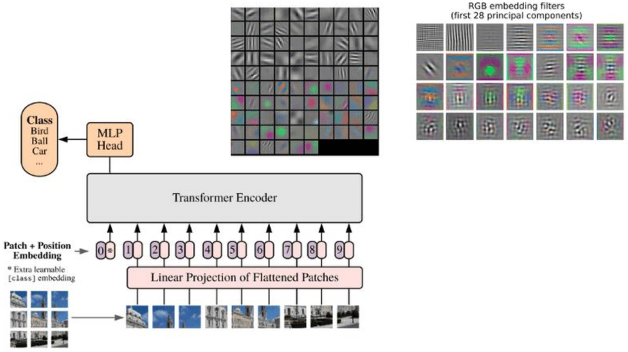
Skills learned from videos can generalize to novel environments, filled with distractors, and even unseen test objects.

# Retrieval Phase




For Scene Semantics + Geometry
Only using first video frame

# Visual Scene Keypoints

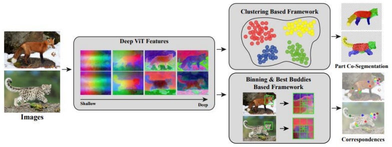



ViT

Deep ViT Features as Dense Visual Descriptors

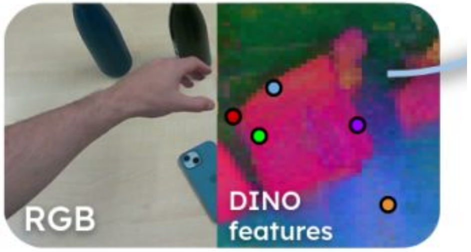
Shir Amir[1], Yossi Gandelsman[2], Shai Bagon[1], and Tali Dekel[1]

[1] Dept. of Computer Science and Applied Math, The Weizmann Inst. of Science
[2] Berkeley Artificial Intelligence Research (BAIR)



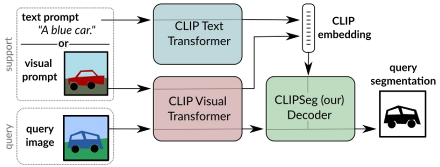

Patch2Pix
Keypoints
$N_{patch}$xD->Cluster->$N_{Pix}$xD

First Video Frame





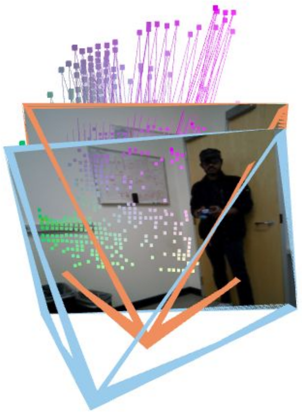
Get Keypoints in the remaining frames




**CLIPSeg**

Only attend to static BG: Table, Wall, Floor + Delete Arm, Person, Hand

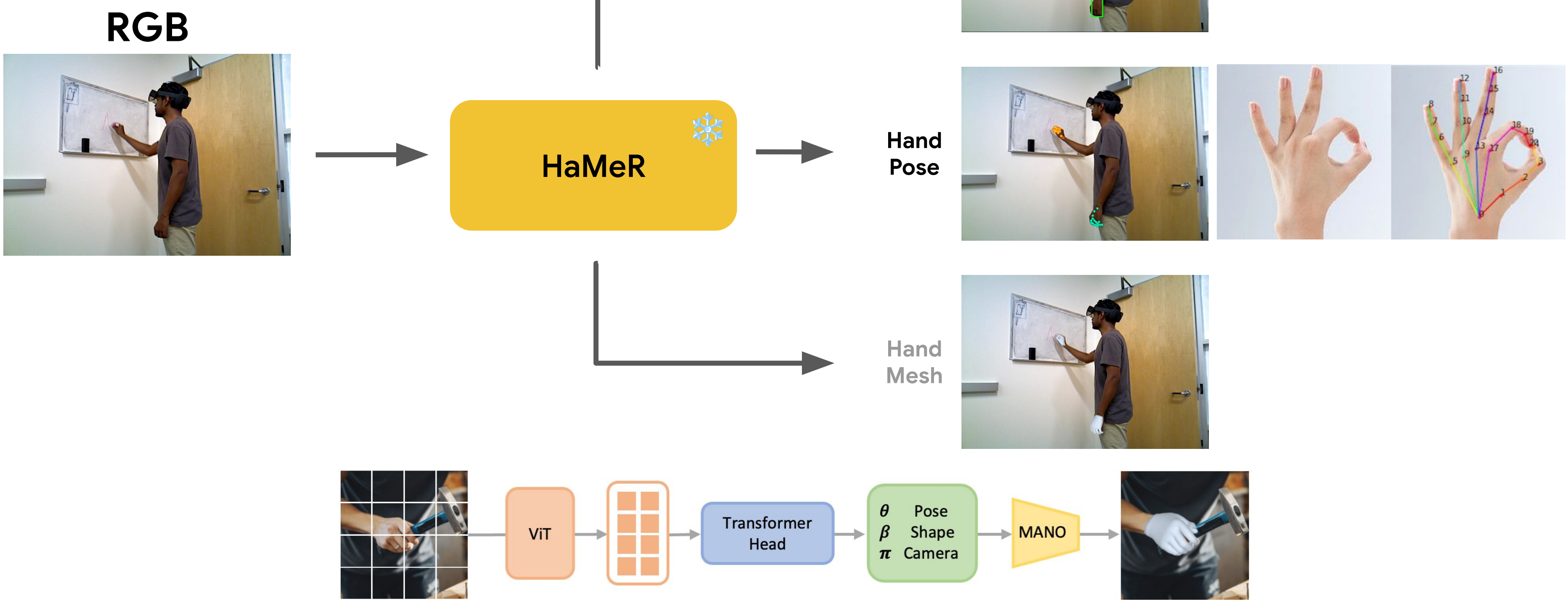
# Rel Camera TF


https://nianticlabs.github.io/mickey
[CVPR2024 Oral]

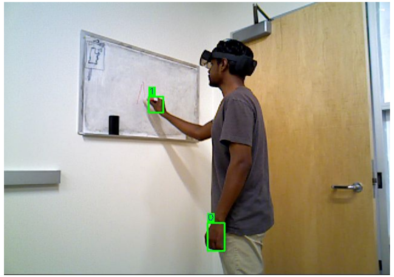
H-Demo: First Frame + Test Frame
Frame-1->Frame-2,........

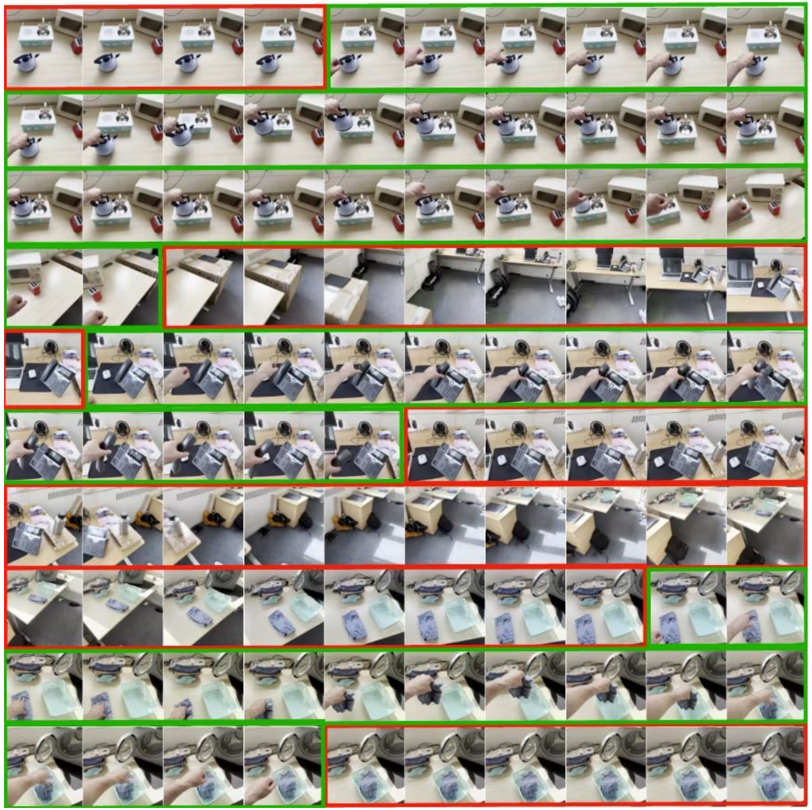
# HaMeR


RGB
HaMeR
Hand BBox
Hand Pose
Hand Mesh
ViT
Transformer Head
θ Pose
β Shape
π Camera
MANO

# HaMeR: Automatic non-hand frame elimination

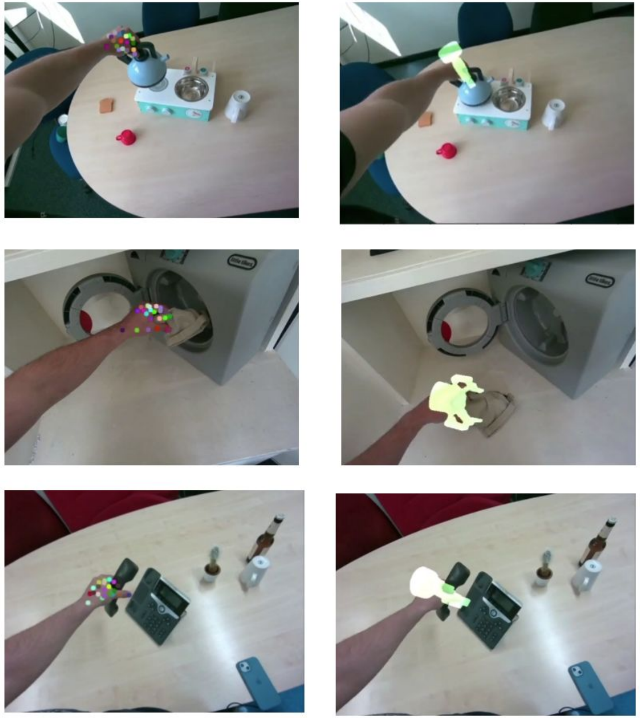
Long, unlabeled video of a human doing everyday activities

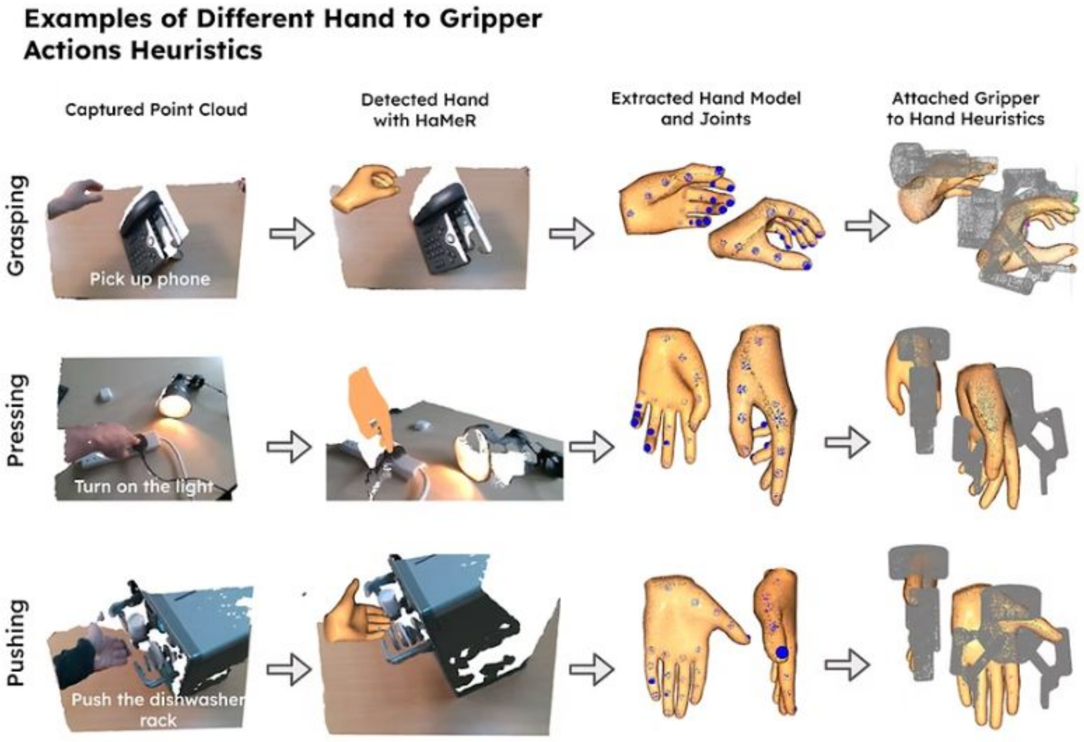




# HaMeR: Automatic non-hand frame elimination

# Human Hand -> Gripper





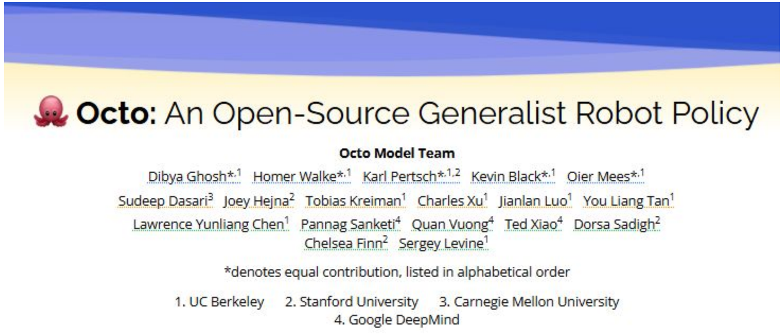
# **Chest Camera Movement** & Scene as a fixed point cloud

Point Cloud before Stabilisation

Point Cloud **after** Stabilisation

Gripper's trajectory before Stabilisation

Gripper's trajectory **after** Stabilisation



# Octo: An Open-Source Generalist Robot Policy

Octo Model Team


Dibya Ghosh*[,1] Homer Walke*[,1] Karl Pertsch*[,1,2] Kevin Black*[,1] Oier Mees*[,1]
Sudeep Dasari[3] Joey Hejna[2] Tobias Kreiman[1] Charles Xu[1] Jianlan Luo[1] You Liang Tan[1]
Lawrence Yunliang Chen[1] Pannag Sanketi[4] Quan Vuong[4] Ted Xiao[4] Dorsa Sadigh[2]
Chelsea Finn[2] Sergey Levine[1]

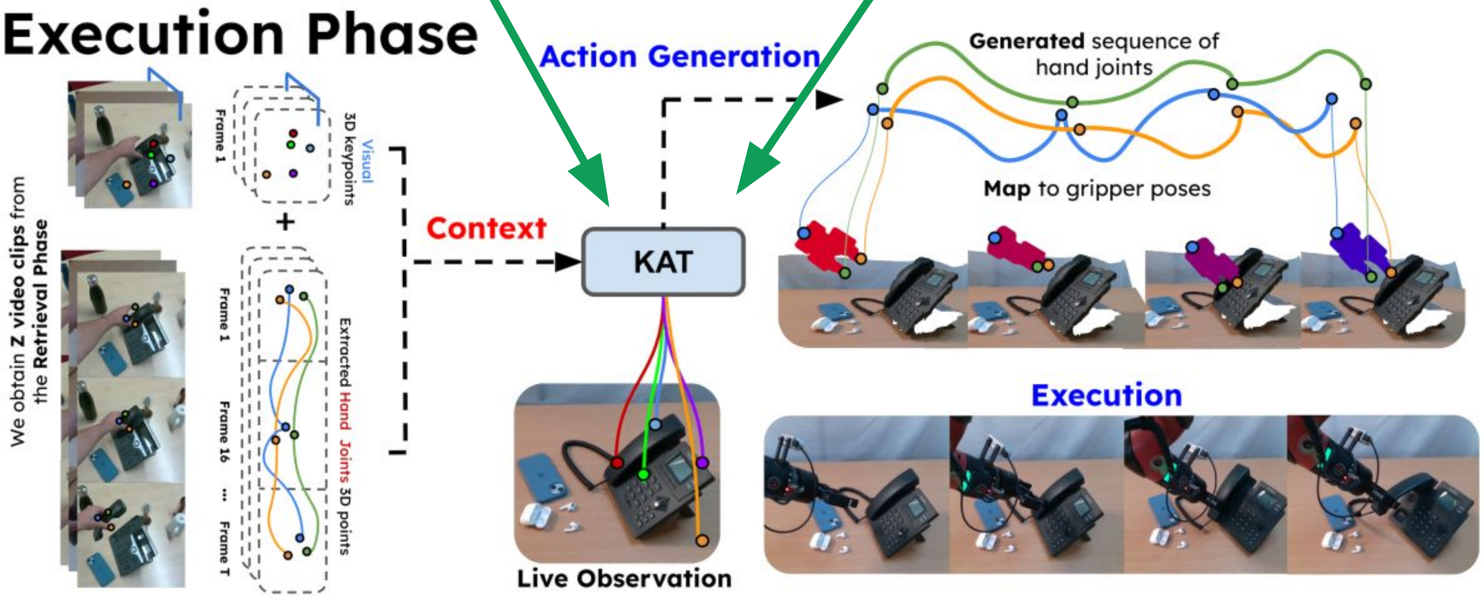
*denotes equal contribution, listed in alphabetical order

1. UC Berkeley 2. Stanford University 3. Carnegie Mellon University
4. Google DeepMind

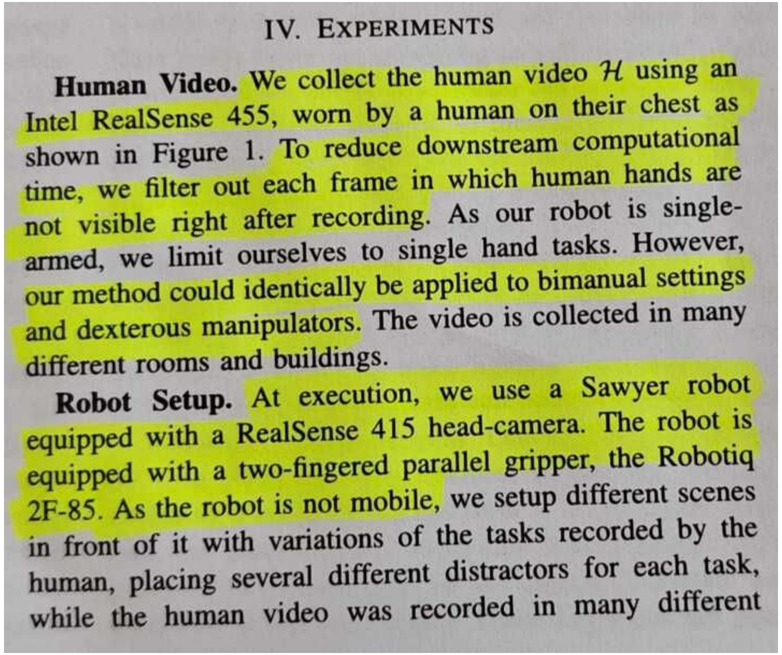

3D Trajectory = KAT(3D Visual Keypoints)

3D Traj for gripper movement

No Finetuning
Use off the shelf LLMs' in context learning (few-shot) ability



# Hardware



IV. EXPERIMENTS

**Human Video.** We collect the human video $\mathcal{H}$ using an Intel RealSense 455, worn by a human on their chest as shown in Figure 1. To reduce downstream computational time, we filter out each frame in which human hands are not visible right after recording. As our robot is single-armed, we limit ourselves to single hand tasks. However, our method could identically be applied to bimanual settings and dexterous manipulators. The video is collected in many different rooms and buildings.

**Robot Setup.** At execution, we use a Sawyer robot equipped with a RealSense 415 head-camera. The robot is equipped with a two-fingered parallel gripper, the Robotiq 2F-85. As the robot is not mobile, we setup different scenes in front of it with variations of the tasks recorded by the human, placing several different distractors for each task, while the human video was recorded in many different

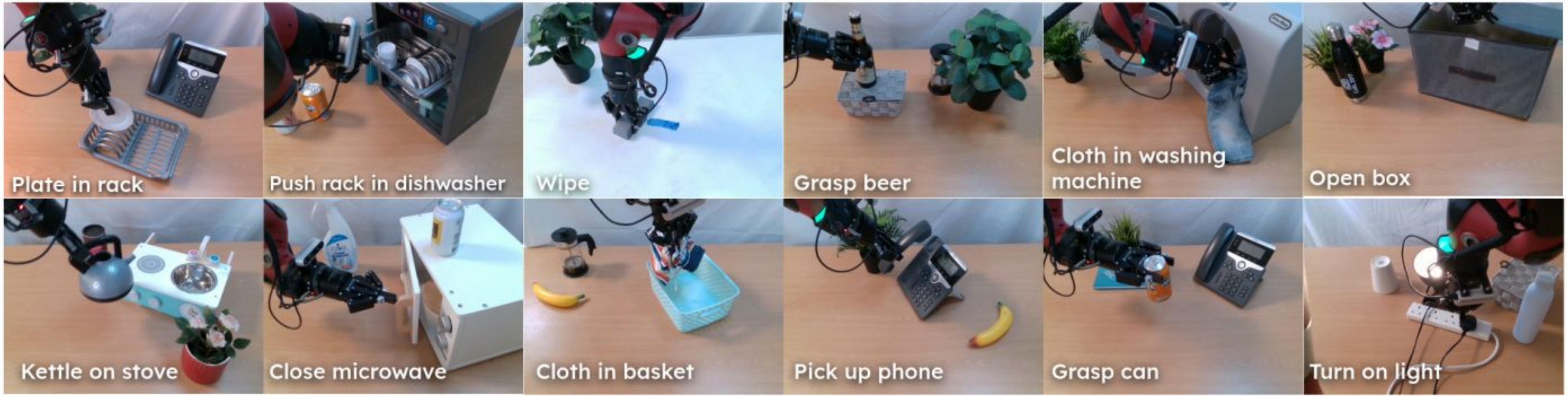
Human 455



Sawyer Robot with fixed base


WristCam 415; not used in the work

# Tasks




12 Everyday Tasks

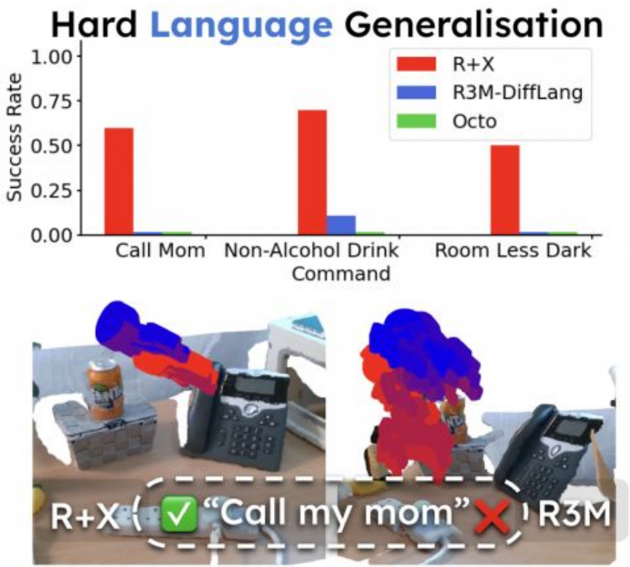
# Baselines

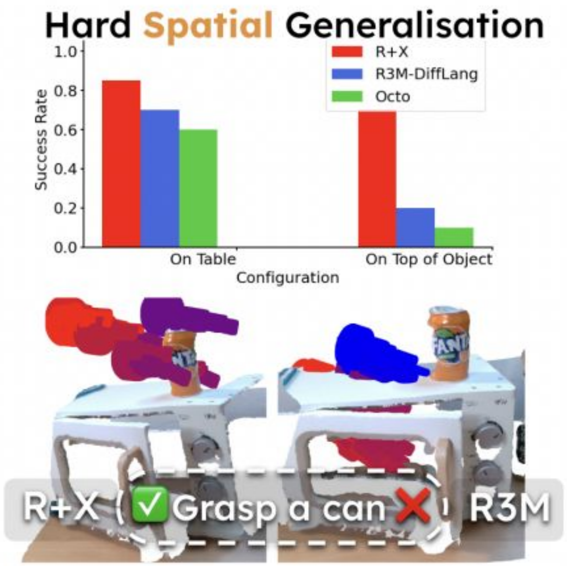
**Baselines.** We compare R+X, and its retrieval and execution design, to training a single, language-conditioned policy. To obtain language captions from the human video, we use Gemini to autonomously caption snippets of the video, obtaining a *(observation, actions, language)* dataset. We finetune R3M (ResNet-50 version [28]) [29] and Octo [30] on this data. We extend R3M to also encode language via SentenceBERT and use a Diffusion Policy [31] head to predict actions from intermediate representations. We denote this version as R3M-DiffLang.

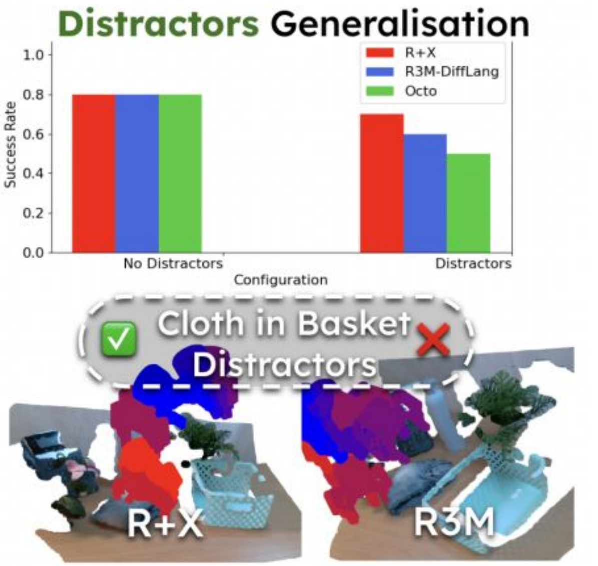
| **Method / Task** | Plate | Push | Wipe | Beer | Wash | Box | Kettle | Micro. | Basket | Phone | Can | Light | **Avg.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **R3M-DiffLang** | 0.5 | 0.7 | 0.4 | 0.7 | 0.5 | 0.5 | 0.4 | 0.8 | 0.7 | 0.4 | 0.7 | 0.3 | 0.55 |
| **Octo** | 0.5 | 0.8 | 0.5 | 0.6 | 0.5 | 0.5 | 0.4 | 0.7 | 0.6 | 0.4 | 0.6 | 0.3 | 0.53 |
| **R+X** | **0.6** | 0.8 | **0.7** | **0.8** | **0.6** | **0.7** | **0.6** | 0.8 | 0.7 | **0.7** | **0.8** | **0.6** | **0.7** |

10 episodes (runs)

# Spatial, Language and Distractors generalisation
# Gripper trajectories move from red to blue.



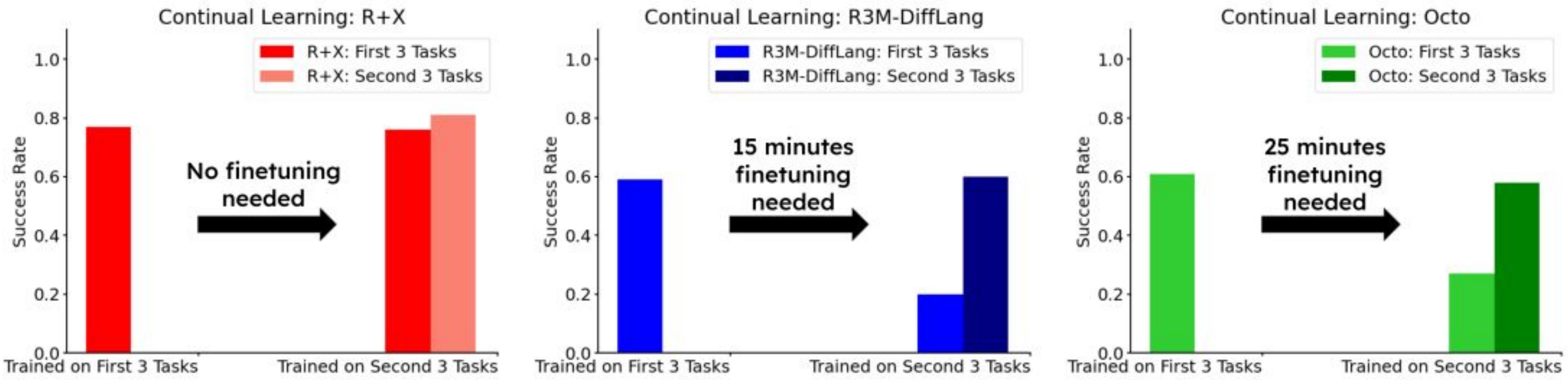




5 episodes (runs)

5 episodes (runs)

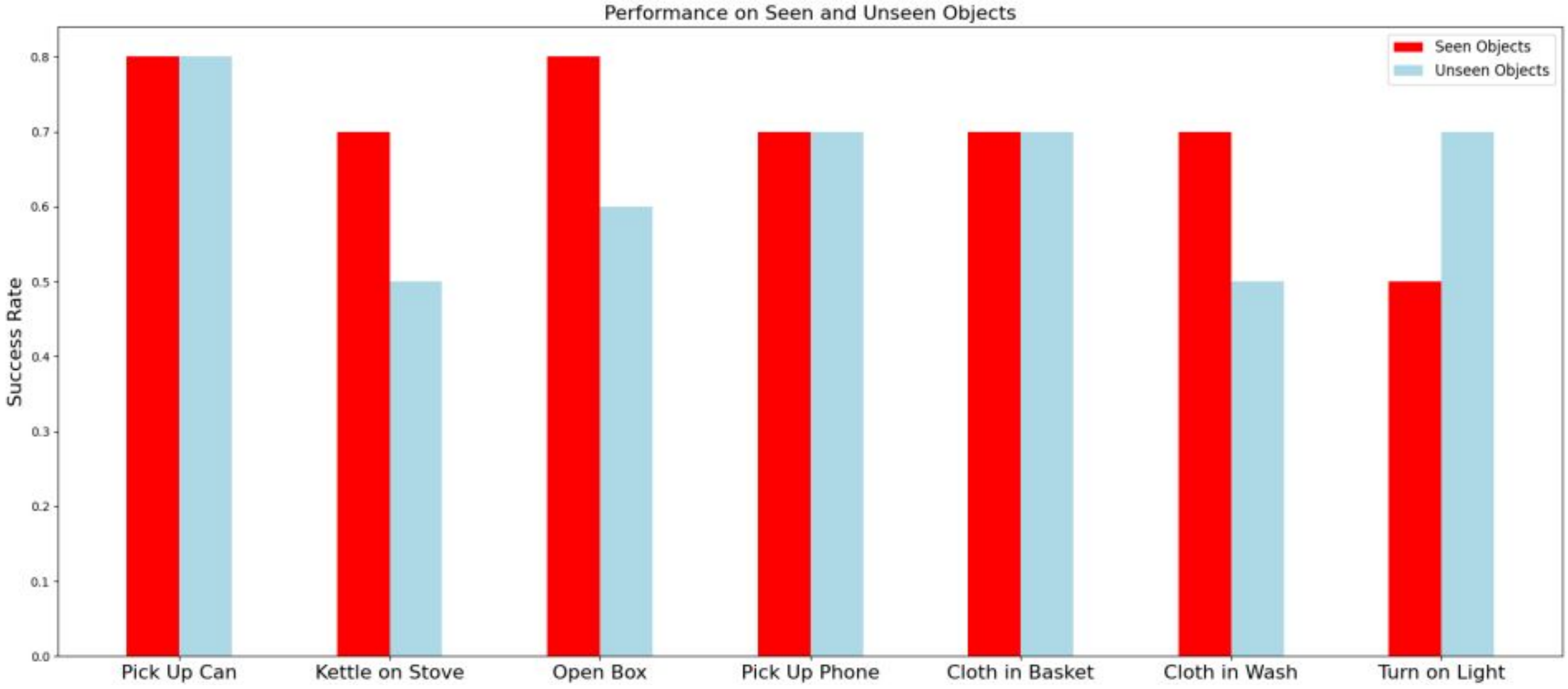
10 episodes (runs)

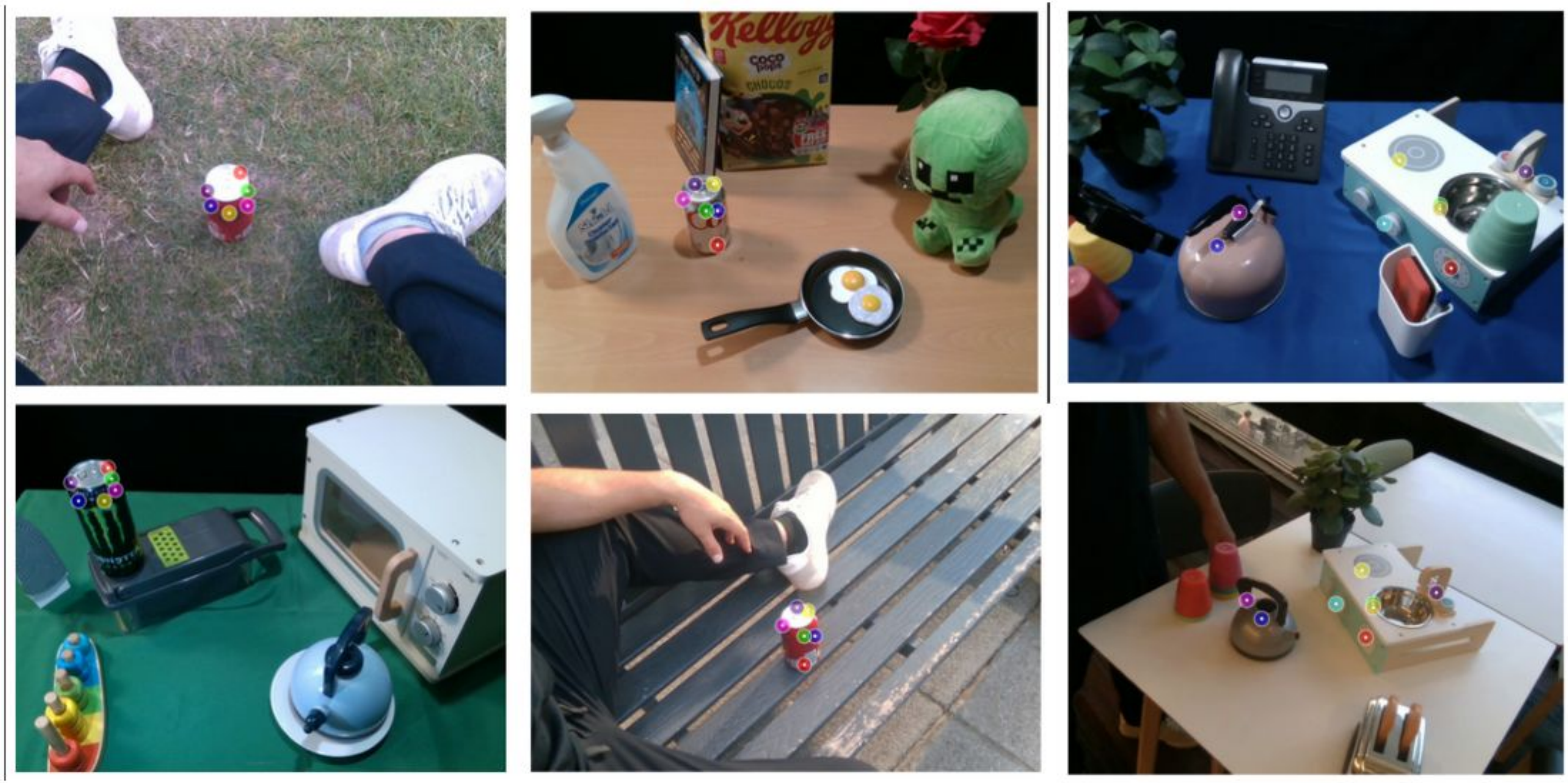
# Can R+X learn task sequentially over time?



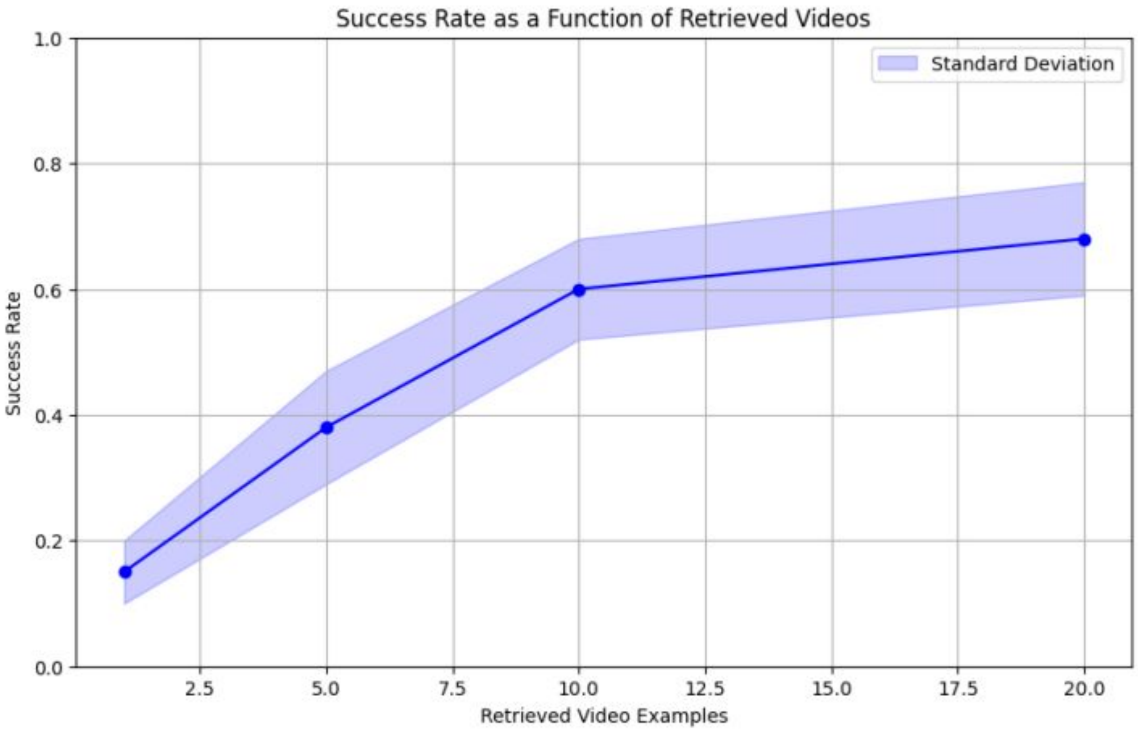

10 demos for 3 new tasks

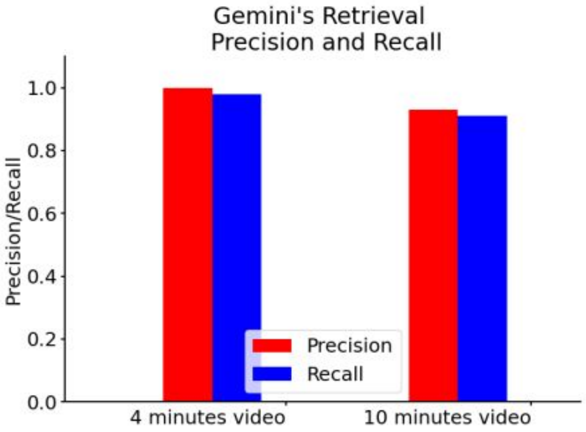
# Success rate on Seen & Unseen Objects





Examples of **keypoints** extracted for the same tasks, but with different views, settings, and target objects


Success Rate as a Function of Retrieved Videos
Standard Deviation
1.0
0.8
0.6
0.4
0.2
0.0
Success Rate
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Retrieved Video Examples


Gemini's Retrieval
Precision and Recall
1.0
0.8
0.6
0.4
0.2
0.0
Precision/Recall
Precision
Recall
4 minutes video
10 minutes video

**Gemini**

# Takeaways

High time to use the reasoning capability of Large Multi Modal Models (LMMs)

Leverage LMMs' few-shot in context learning ability for generalization purposes

Latent plan pre-training benefits multi-task learning. [MimicPlay, LAPA]

Similarly, nuanced inputs like Keypoints are good for generalization instead of direct RGBD or text

# Questions?