



R+X: Retrieval and Execution from Everyday Human Videos

ICRA 2025

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

Jishnu P

Reading Group | [IRVL](#)

1/31/25

Authors



**Georgios
Papagiannis***



**Norman
Di Palo***



**Pietro
Vitiello**

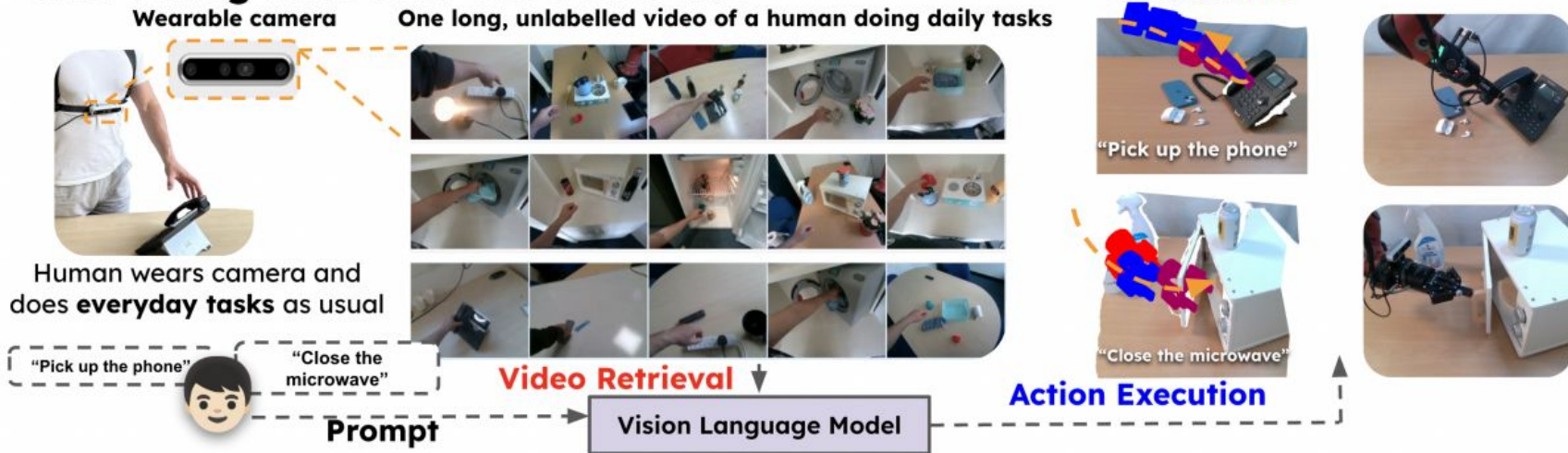


**Edward
Johns**

The Robot Learning Lab
Imperial College London

Problem

R+X learns robot skills from long, unlabelled videos of humans interacting with their environments



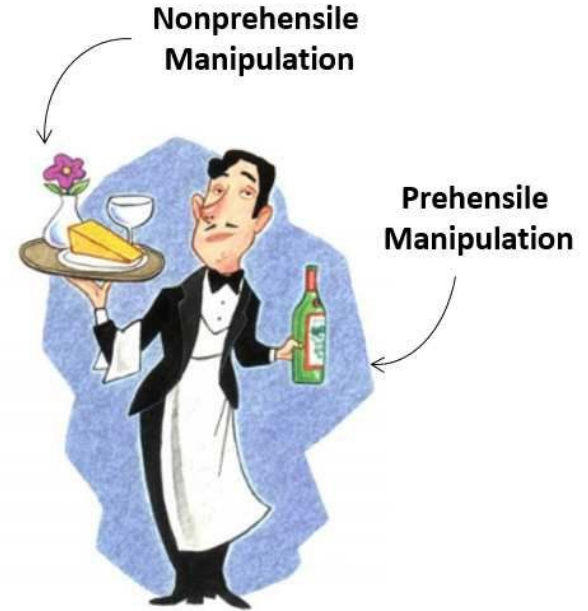
Leverage understanding of large models

- 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



Long, unlabeled video of a human doing everyday activities

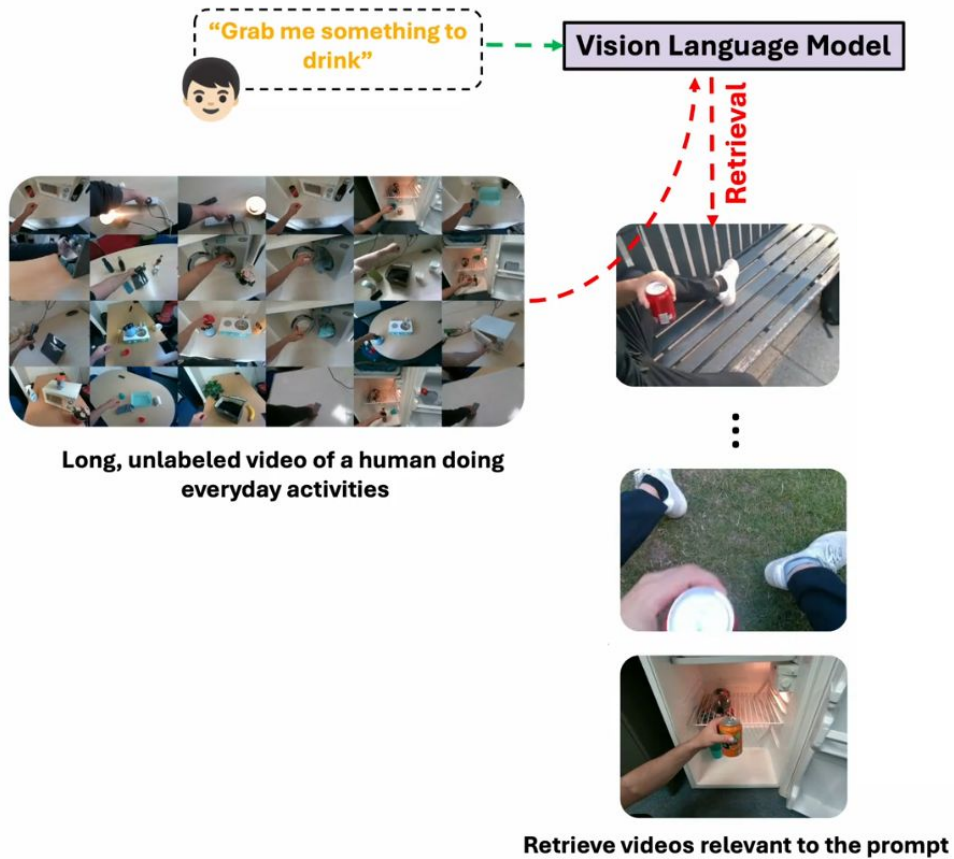


Single Unlabelled Video
with less
clutter/distractors

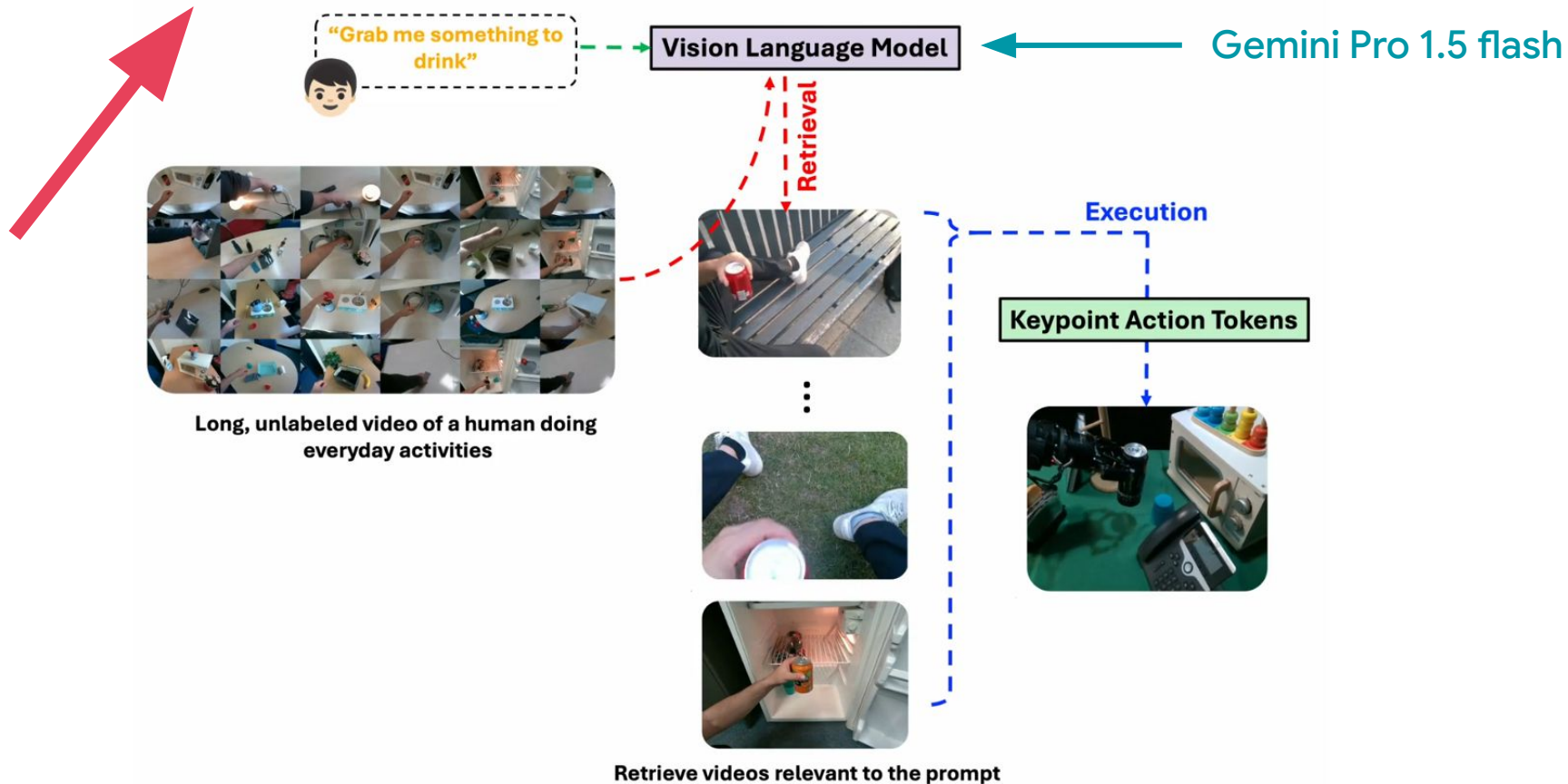




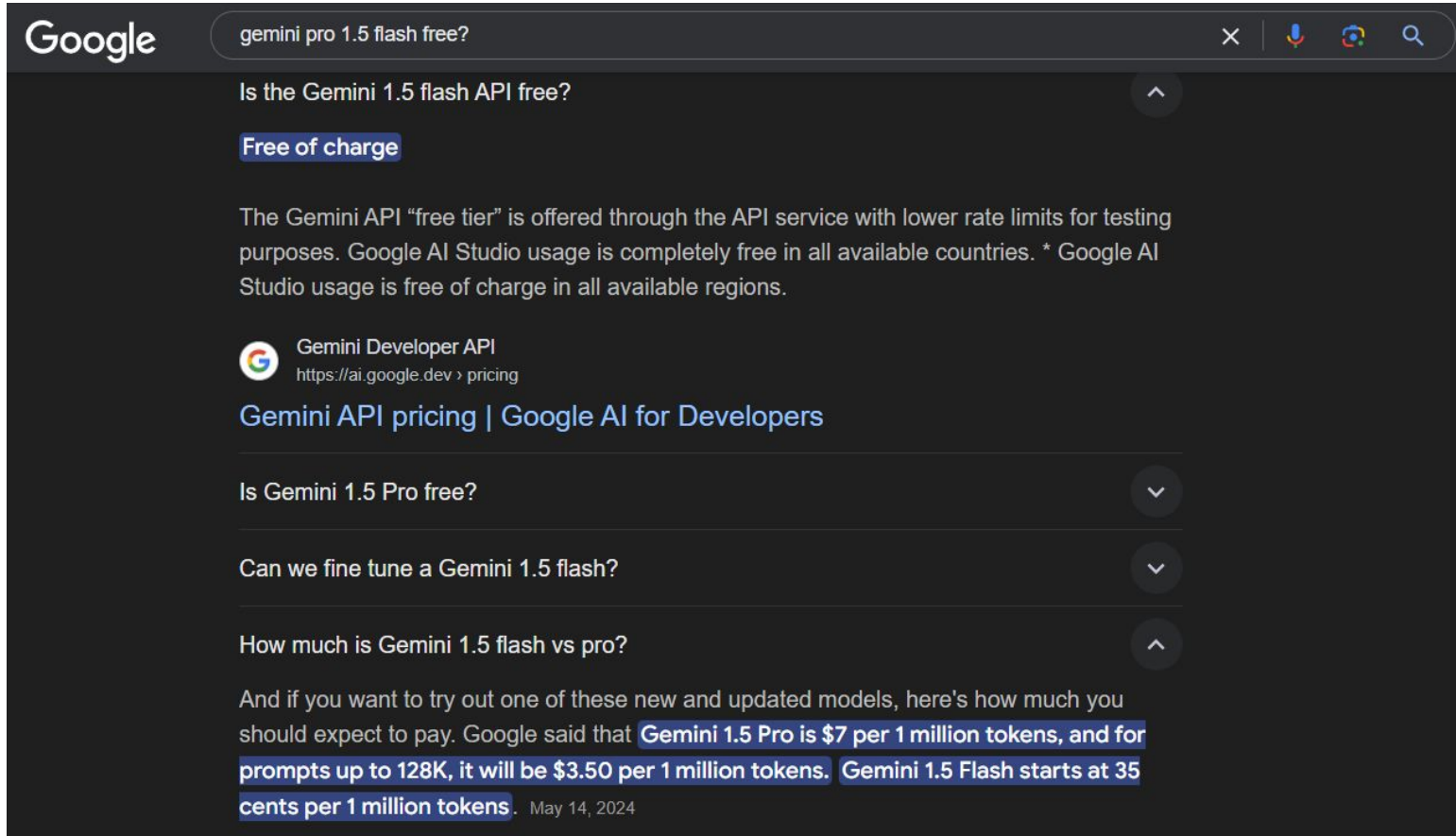
Long, unlabeled video of a human doing everyday activities



R+X : Retrieval and Execution



Google Search as of 1/31/2025



The image shows a Google search interface with a dark theme. The search bar at the top contains the text "gemini pro 1.5 flash free?". Below the search bar, the first search result is displayed. The title of the result is "Is the Gemini 1.5 flash API free?". Below the title, there is a blue pill-shaped badge that says "Free of charge". The main text of the result states: "The Gemini API 'free tier' is offered through the API service with lower rate limits for testing purposes. Google AI Studio usage is completely free in all available countries. * Google AI Studio usage is free of charge in all available regions." Below this text, there is a small circular icon with a 'G' and the text "Gemini Developer API" followed by the URL "https://ai.google.dev > pricing". Underneath the URL is a blue link titled "Gemini API pricing | Google AI for Developers". Below this result, there are three more search results listed as questions: "Is Gemini 1.5 Pro free?", "Can we fine tune a Gemini 1.5 flash?", and "How much is Gemini 1.5 flash vs pro?". The first result has a downward arrow, the second has a downward arrow, and the third has an upward arrow. The text for the third result is partially visible: "And if you want to try out one of these new and updated models, here's how much you should expect to pay. Google said that Gemini 1.5 Pro is \$7 per 1 million tokens, and for prompts up to 128K, it will be \$3.50 per 1 million tokens. Gemini 1.5 Flash starts at 35 cents per 1 million tokens." The date "May 14, 2024" is visible at the bottom of this snippet.


Google

gemini pro 1.5 flash free?

Is the Gemini 1.5 flash API free?

Free of charge

The Gemini API "free tier" is offered through the API service with lower rate limits for testing purposes. Google AI Studio usage is completely free in all available countries. * Google AI Studio usage is free of charge in all available regions.

 Gemini Developer API
<https://ai.google.dev > pricing>

[Gemini API pricing | Google AI for Developers](#)

Is Gemini 1.5 Pro free?

Can we fine tune a Gemini 1.5 flash?

How much is Gemini 1.5 flash vs pro?

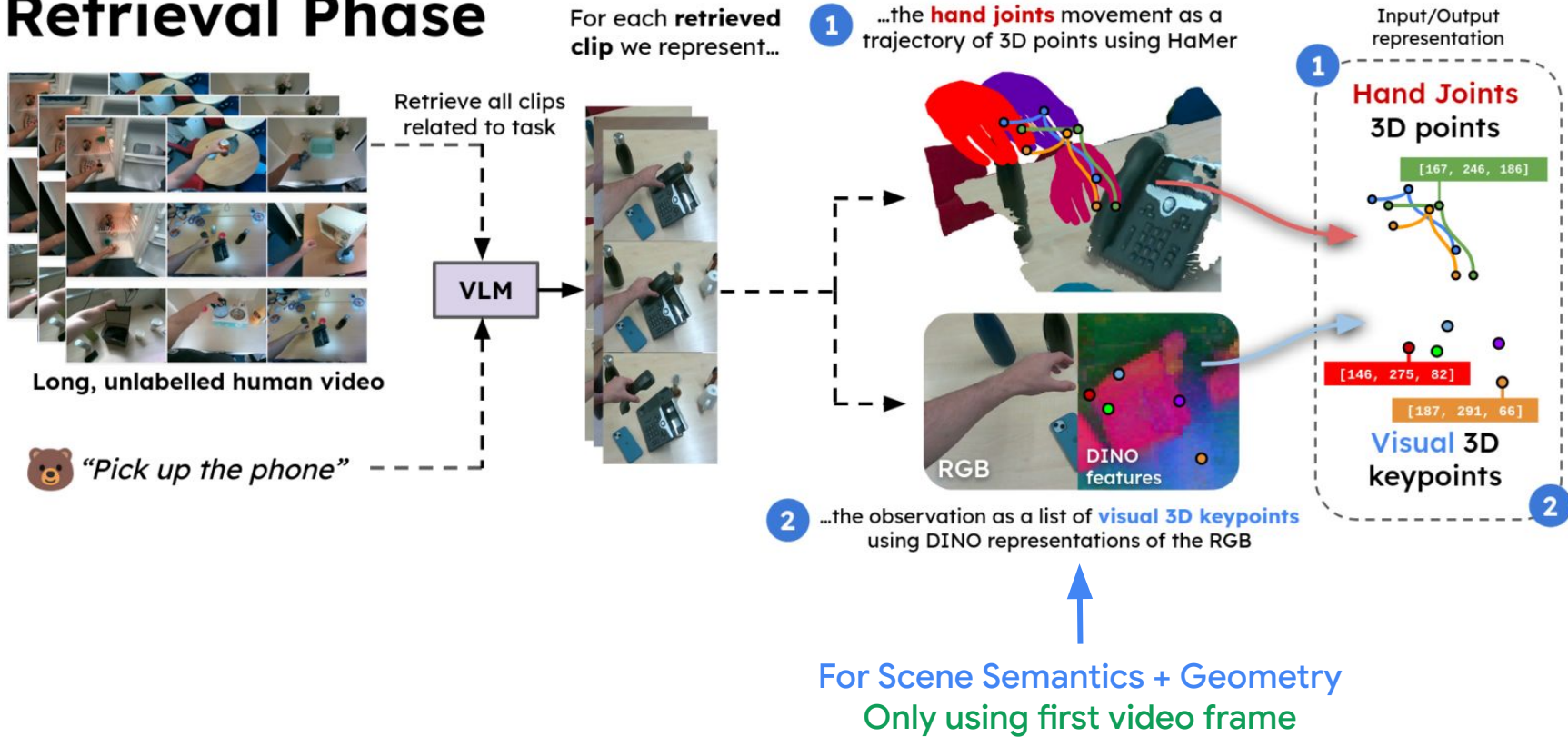
And if you want to try out one of these new and updated models, here's how much you should expect to pay. Google said that **Gemini 1.5 Pro is \$7 per 1 million tokens, and for prompts up to 128K, it will be \$3.50 per 1 million tokens.** **Gemini 1.5 Flash starts at 35 cents per 1 million tokens**. May 14, 2024

Deploy Immediately to Novel Environments and Objects

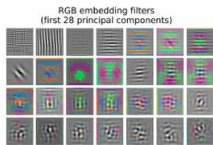
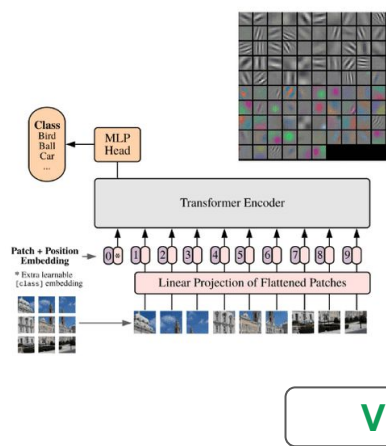
Skills learned from videos can generalize to novel environments, filled with distractors, and even unseen test objects.



Retrieval Phase



Visual Scene Keypoints

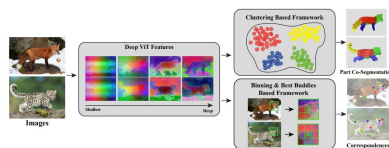


Deep ViT Features as Dense Visual Descriptors

Shir Amir¹, Yossi Gandelsman², Shai Bagon¹, and Tali Dekel¹

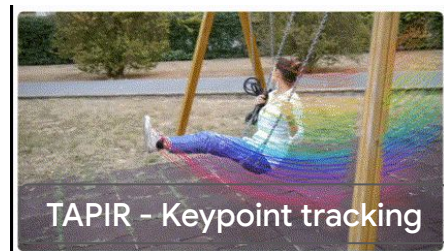
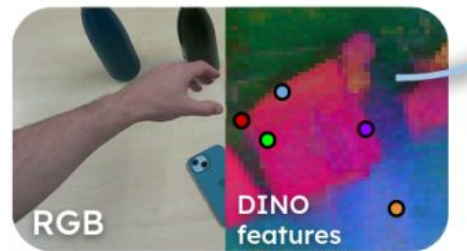
¹ Dept. of Computer Science and Applied Math, The Weizmann Inst. of Science

² Berkeley Artificial Intelligence Research (BAIR)

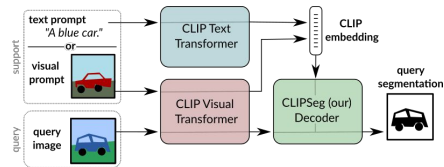


Patch2Pix
Keypoints
 $N_{patch} \times D \rightarrow \text{Cluster} \rightarrow N_{pix} \times D$

First Video Frame



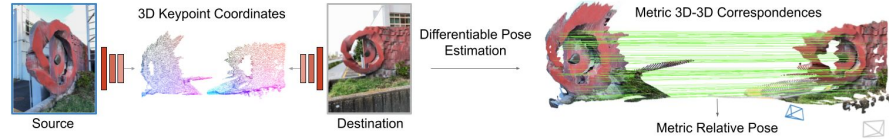
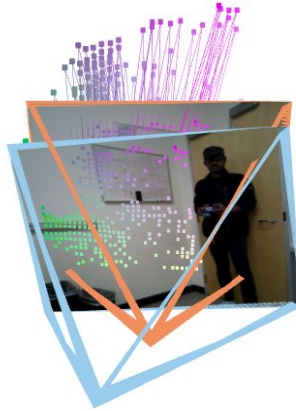
Get Keypoints in the remaining frames



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,.....

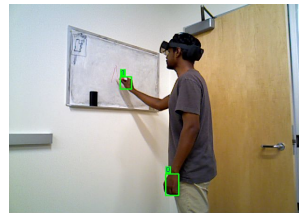


HaMeR

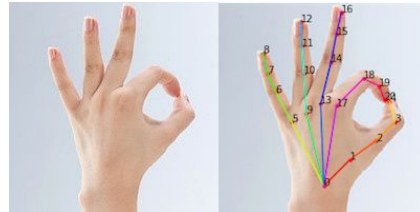
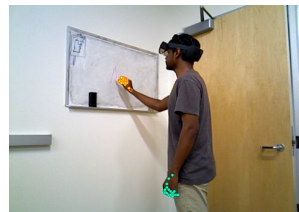
RGB



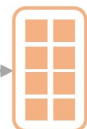
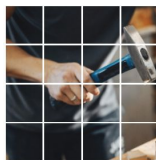
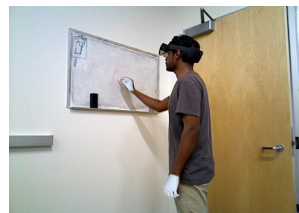
Hand
BBox



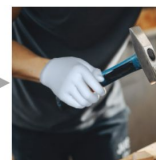
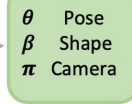
Hand
Pose



Hand
Mesh



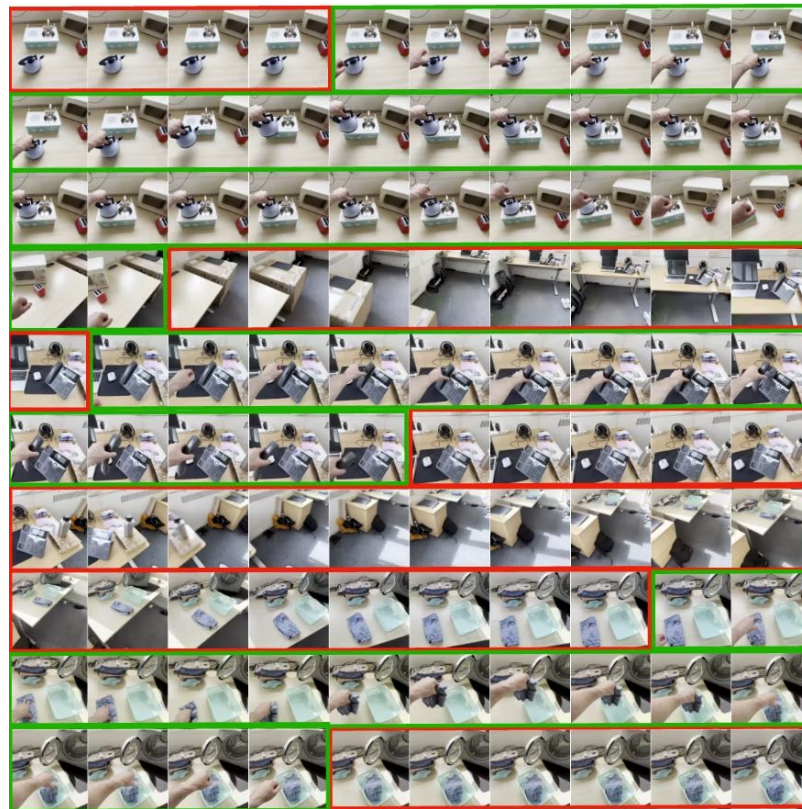
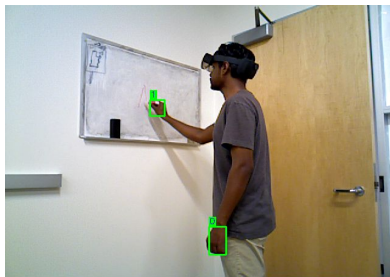
Transformer
Head



HaMeR: Automatic non-hand frame elimination



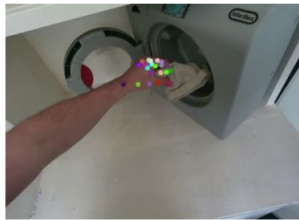
Long, unlabeled video of a human doing everyday activities



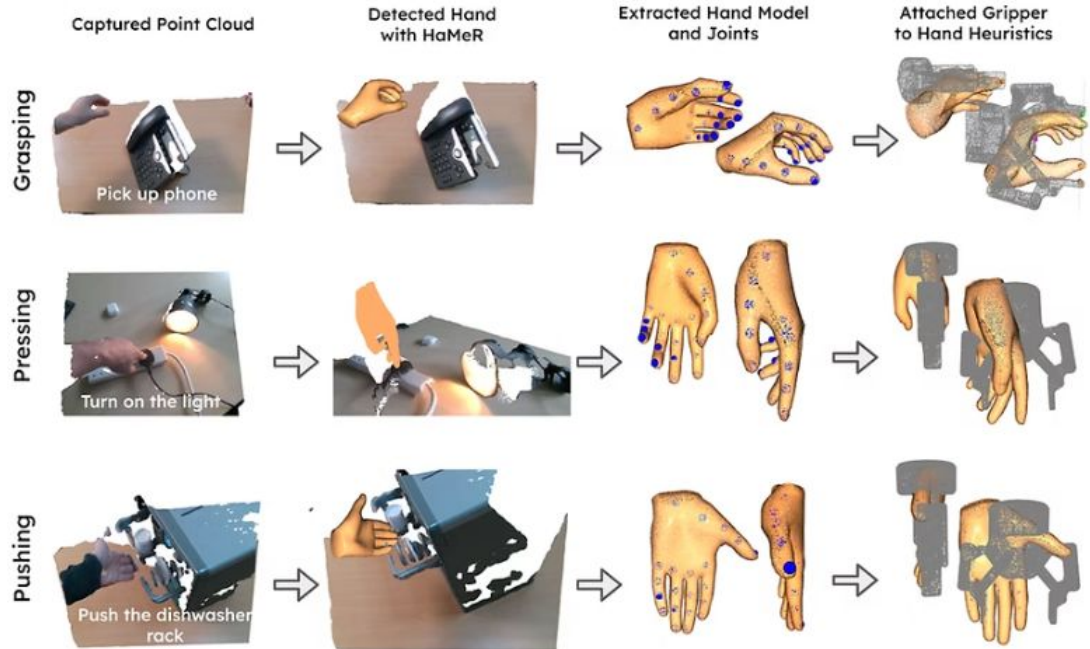
HaMeR: Automatic non-hand frame elimination



Human Hand -> Gripper



Examples of Different Hand to Gripper Actions Heuristics



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 Dasari³ Joey Hejna² Tobias Kreiman¹ Charles Xu¹ Jianlan Luo¹ You Liang Tan¹
Lawrence Yunliang Chen¹ Pannag Sanketi⁴ Quan Vuong⁴ Ted Xiao⁴ Dorsa Sadigh²
Chelsea Finn² Sergey Levine¹

^{*}denotes equal contribution, listed in alphabetical order

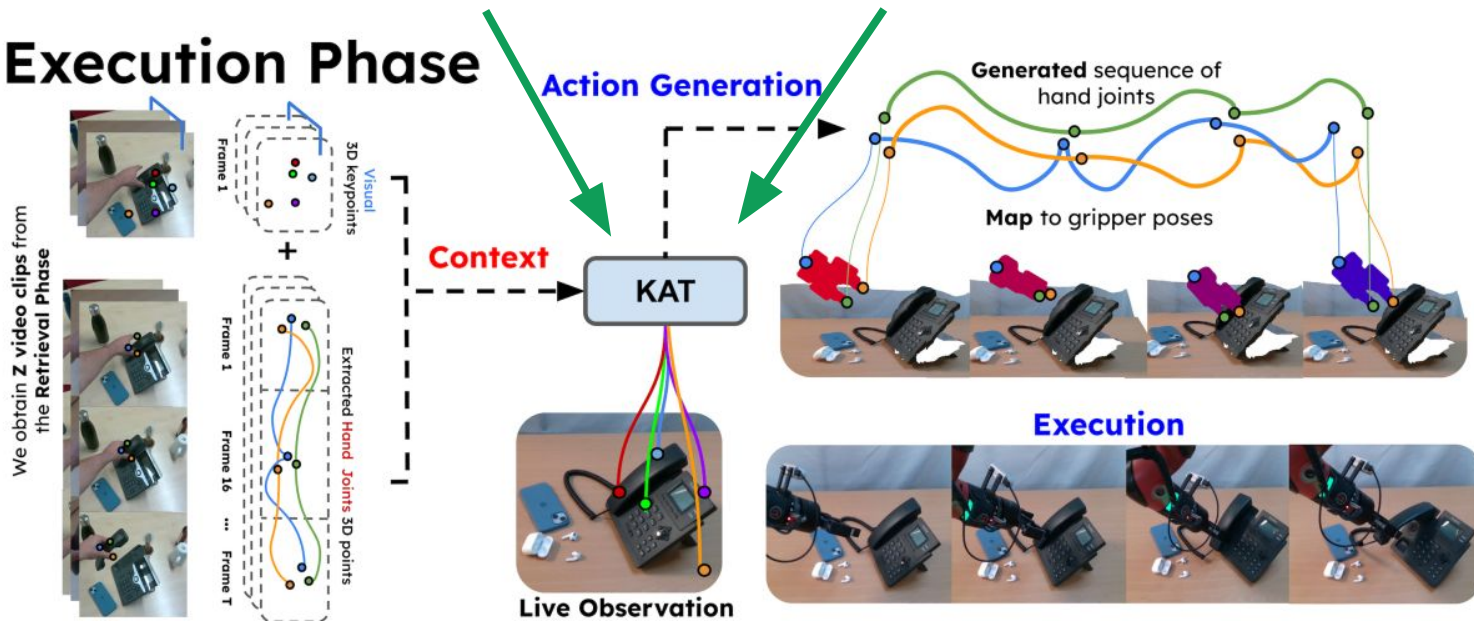
1. UC Berkeley 2. Stanford University 3. Carnegie Mellon University
4. Google DeepMind

3D Trajectory = KAT(3D Visual Keypoints)

3D Traj for gripper movement

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

Execution Phase



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

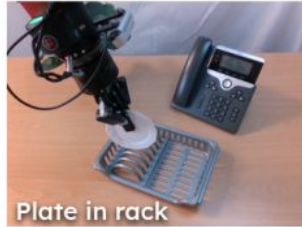
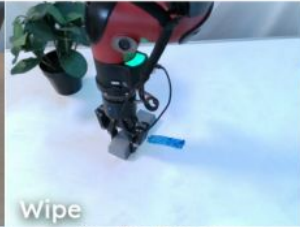


Plate in rack



Push rack in dishwasher



Wipe



Grasp beer



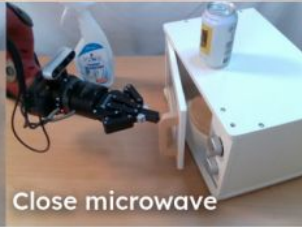
Cloth in washing machine



Open box



Kettle on stove



Close microwave



Cloth in basket



Pick up phone



Grasp can



Turn on light

12 Everyday Tasks

Baselines

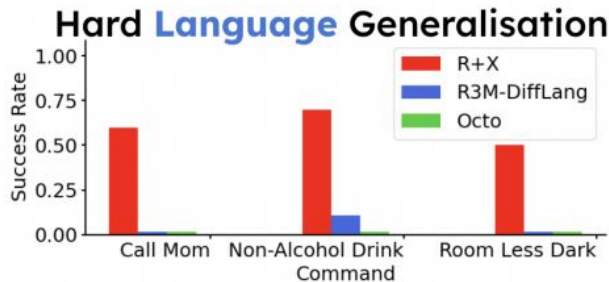
Baselines. We compare R+X, and its retrieval and execution design, to training a single, language-conditioned policy. To obtain language captions from the human video, we use Gemini to autonomously caption snippets of the video, obtaining a (*observation, actions, language*) dataset. We finetune R3M (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

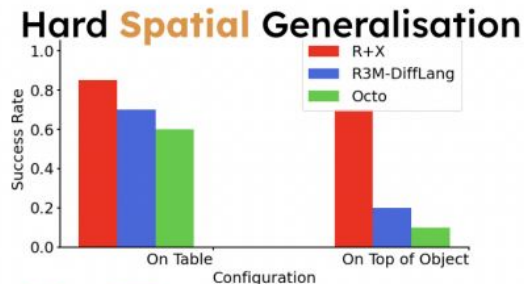
10 episodes (runs)

Spatial, Language and Distractors generalisation

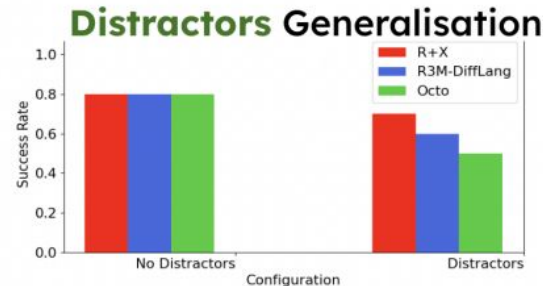
Gripper trajectories move from red to blue.



5 episodes (runs)

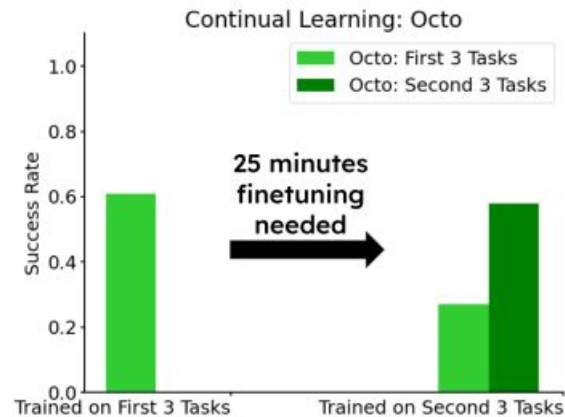
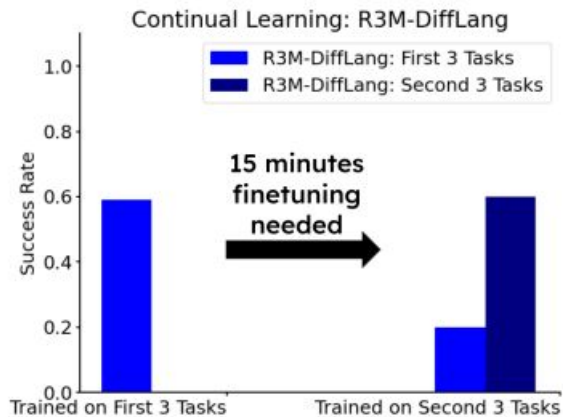
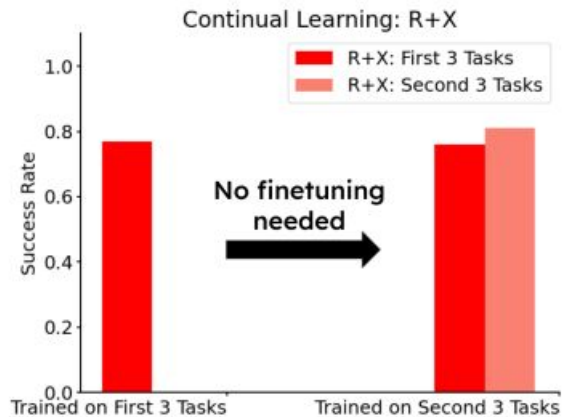


5 episodes (runs)



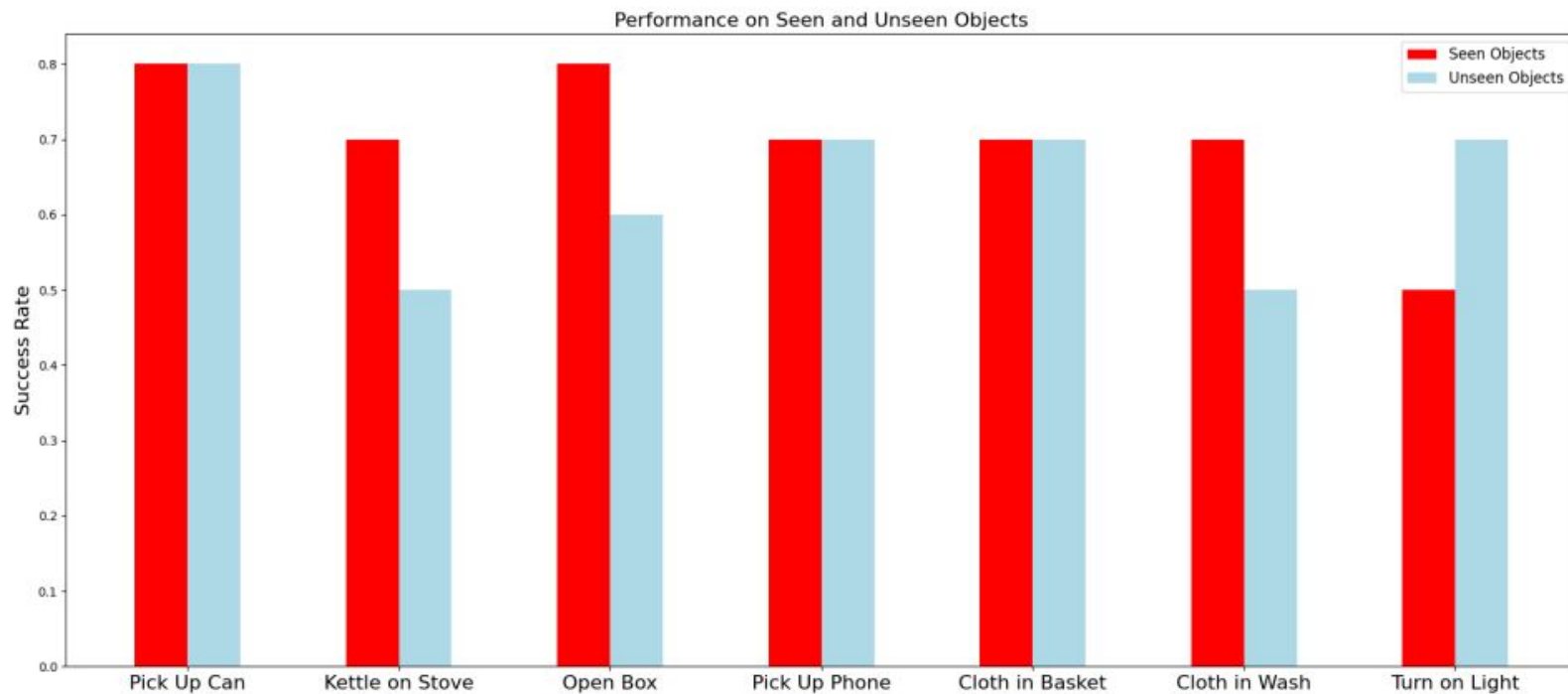
10 episodes (runs)

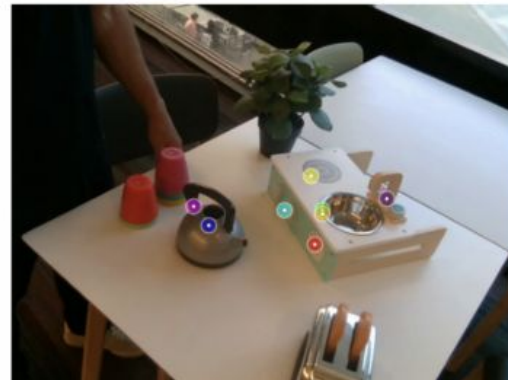
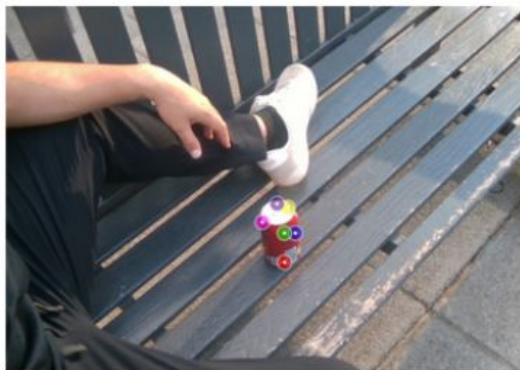
Can R+X learn task sequentially over time?



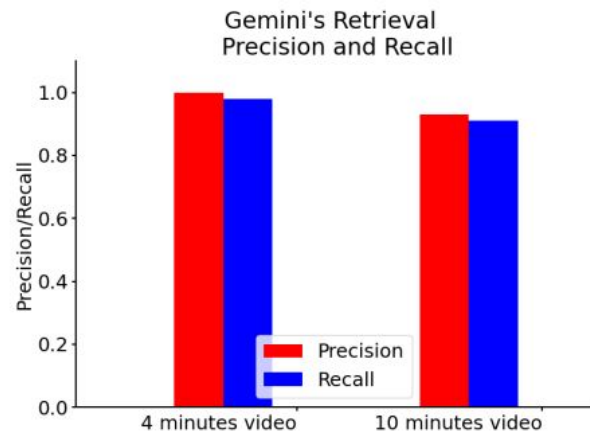
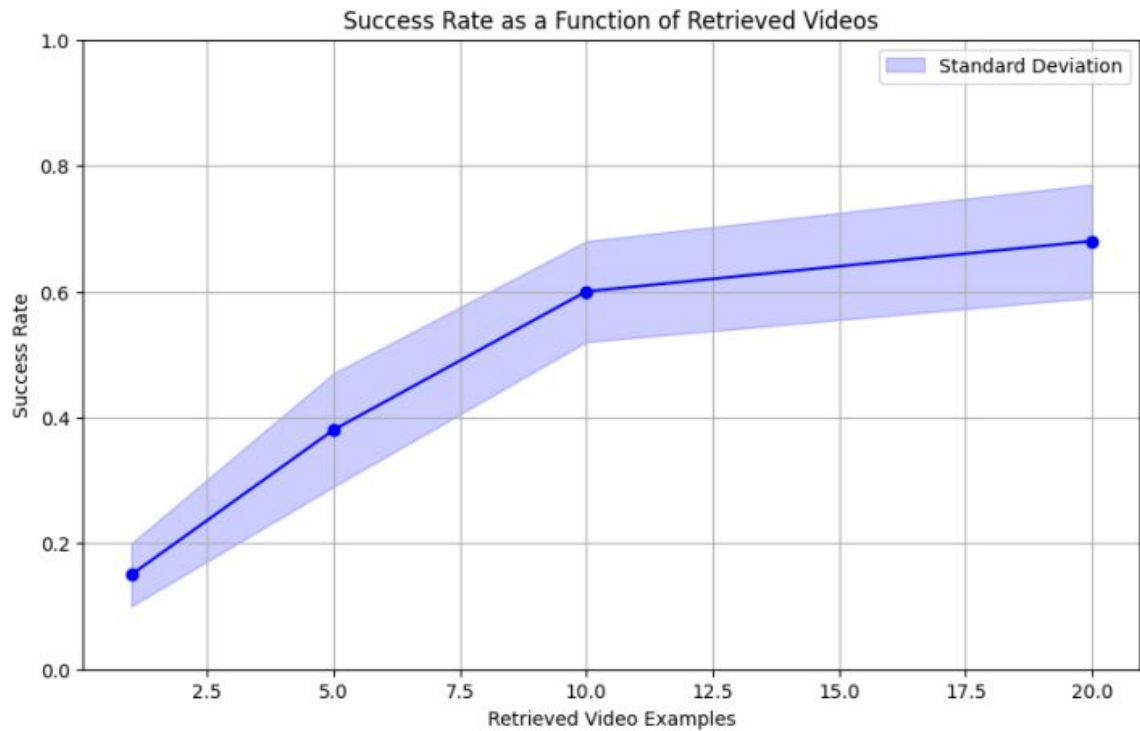
10 demos for 3 new tasks

Success rate on Seen & Unseen Objects





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



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?