Мультимодальные и поведенческие модели

3D Spatial Reasoning for Manipulation

MIPT VLA 2026 | 19 октября 2026

Алексей Староверов



МФТИ

Center for Cognitive Modeling

План лекции

From Seeing to Doing: Ascending the Ladder of Visual Intelligence

PointWorld, a large-scale 3D world model

BIND: Binding 3D Robot Actions to 2D Image Features

How do we build visually intelligence machines?



Visual Intelligence entails

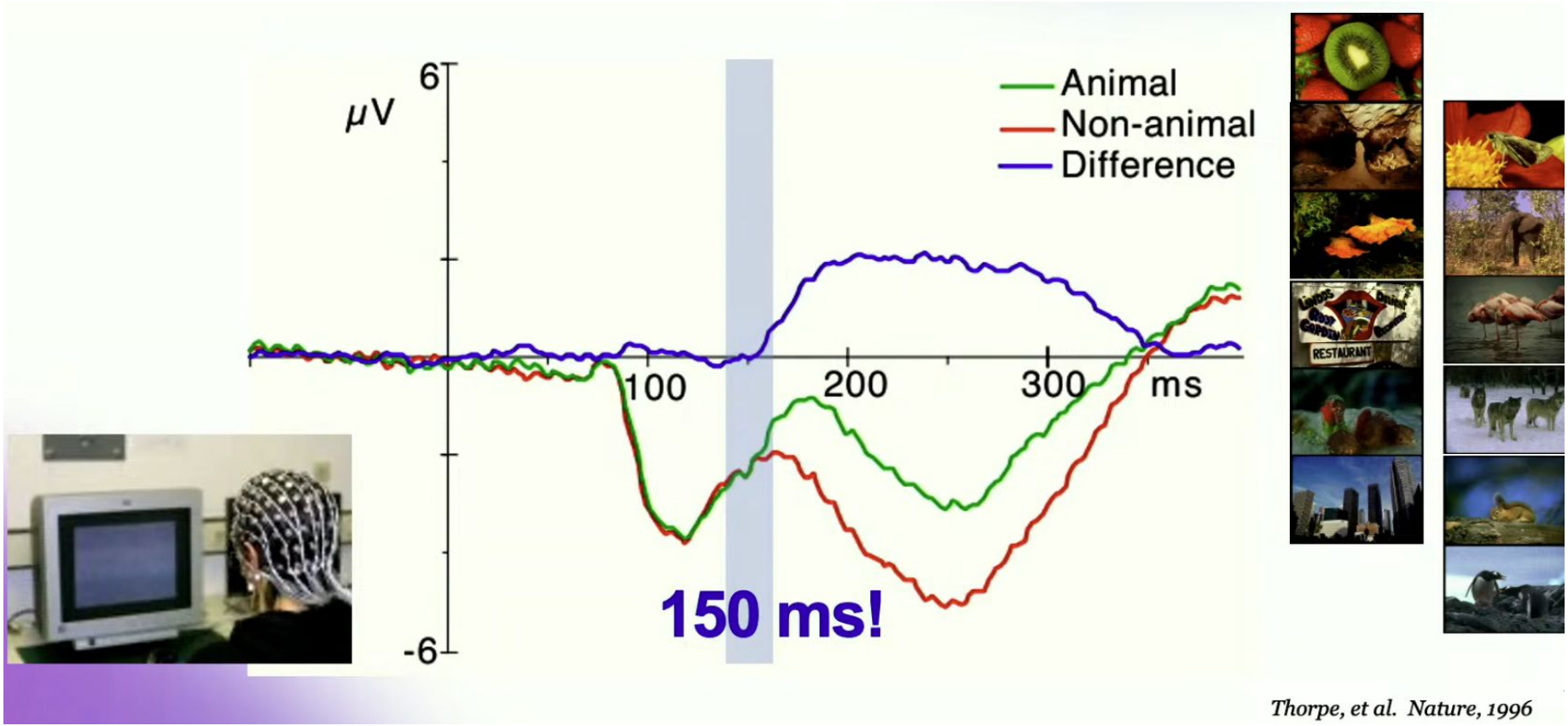
Understanding | Reasoning | Generation

Definition: label the semantic contents or perceptual properties based on the visible pixels

Capabilities: object recognition, segmentation, video classification, human action understanding, etc.

Example questions: draw bounding boxes of all chairs in the image? Find all the basketball videos?

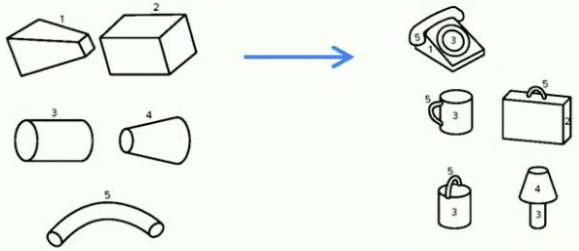
Understanding objects is fundamental to human vision



Thorpe, et al. Nature, 1996

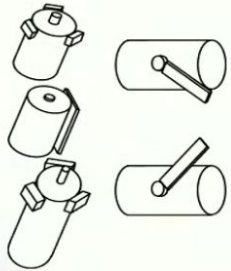
Many early models for Object Recognition

Geons (Recognition by Components)



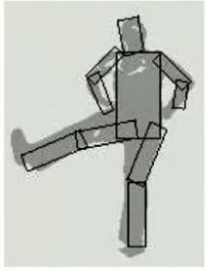
Brooks & Binford, 1979

Generalized Cylinder



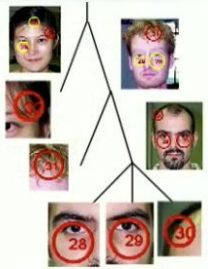
Biederman, 1970s-80s

Pictorial Structure & Constellation Models



Leung et al. 1999; Sivic et al. 2003;
Grauman et al. 2005; Lazebnik et al. 2006;
Fei-Fei et al. 2005

And-Or Graphs



Sudderth et al. 2005
Li et al. 2007

Bag of Words



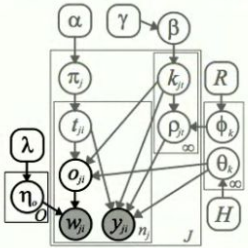
Felzenszwalb et al. 2000
Fergus et al. 2003
Fei-Fei et al. 2003

Boosting



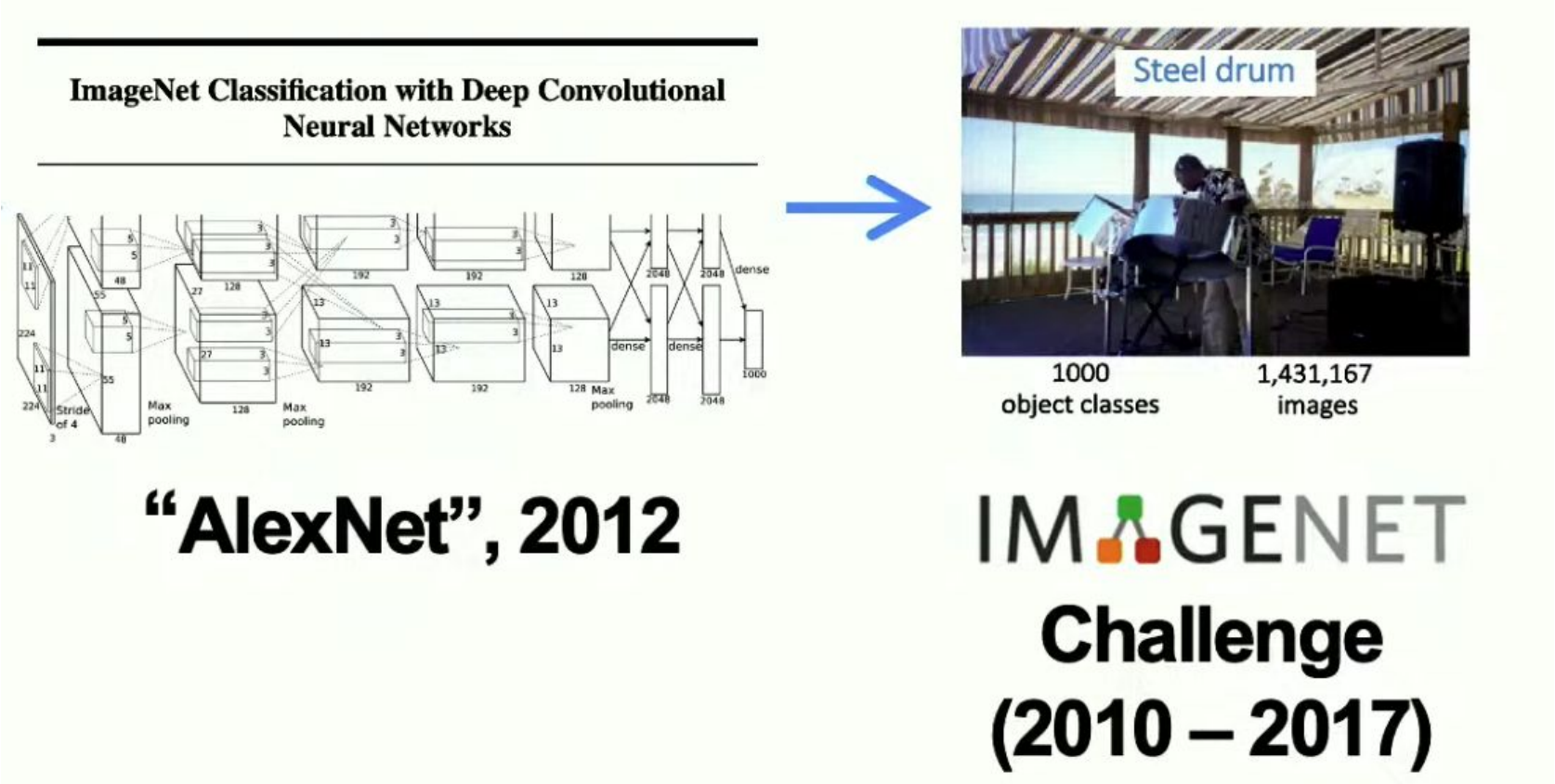
Chen et al. 2006
Zhu et al. 2007

Non-parametric Bayes

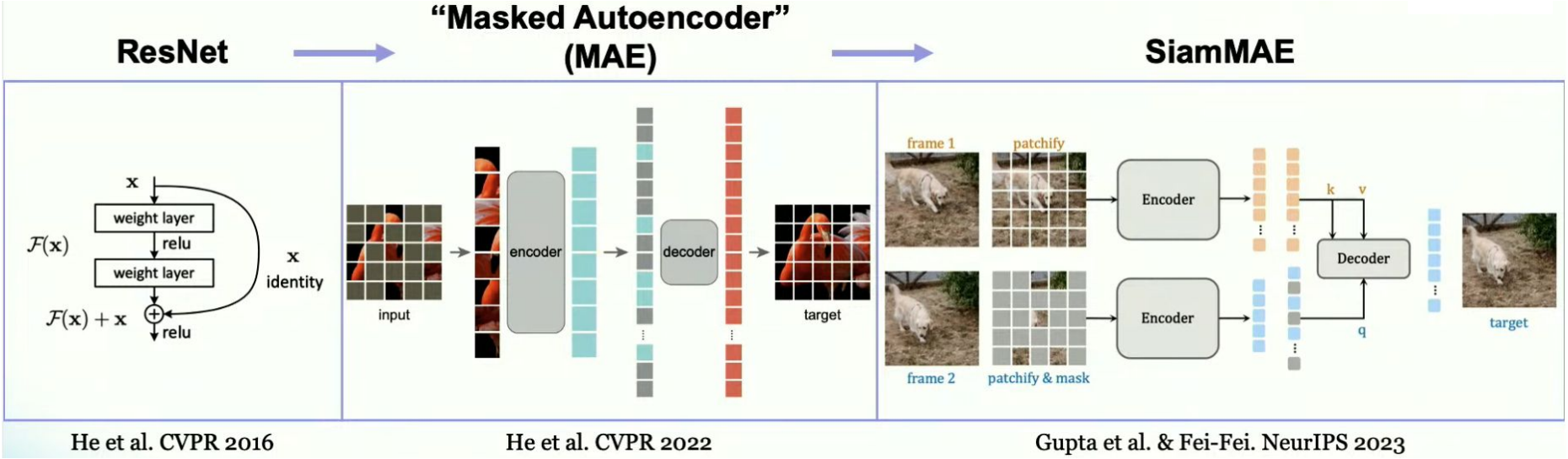


Viola & Jones 2001
Torralba et al. 2004

Challenge drove tremendous Progress in Computer Vision



Representational Learning



He et al. CVPR 2016 | He et al. CVPR 2022 | Gupta et al. & Fei-Fei. NeurIPS 2023

But there is more beyond pixels; Reasoning is key

Understanding | **Reasoning** | Generation

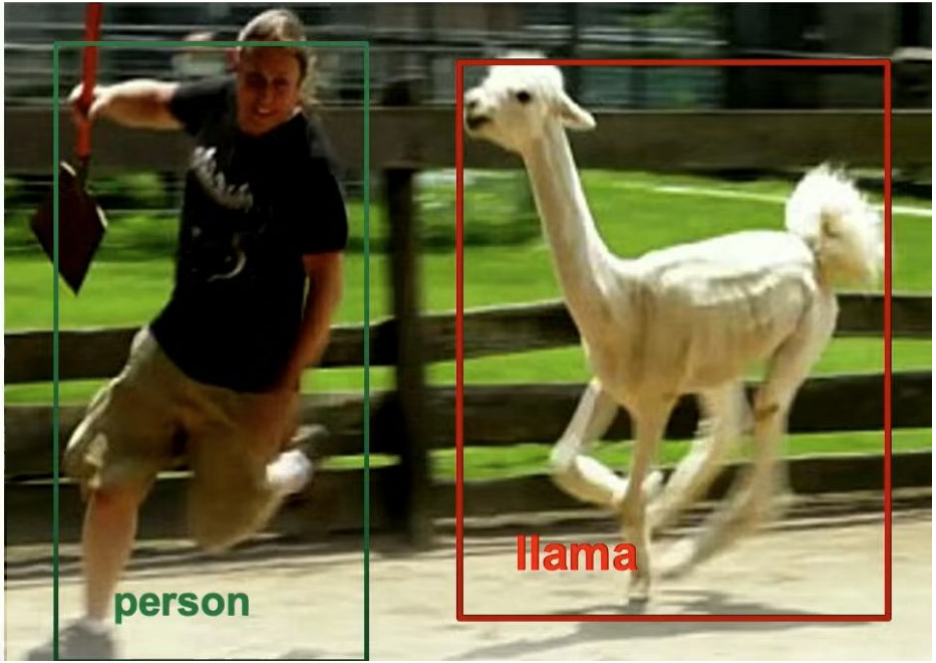
Definition: infer or reason with information beyond visible pixels

Capabilities: object relationship understanding, story telling, QA

Example questions: Is the person sitting on the red chair?
Caption this photo?

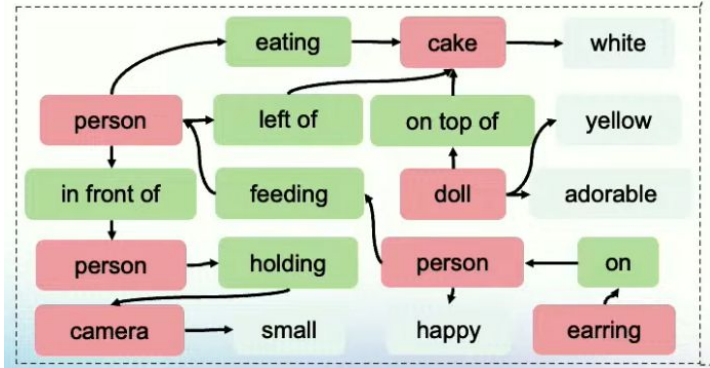
What is beyond object recognition?

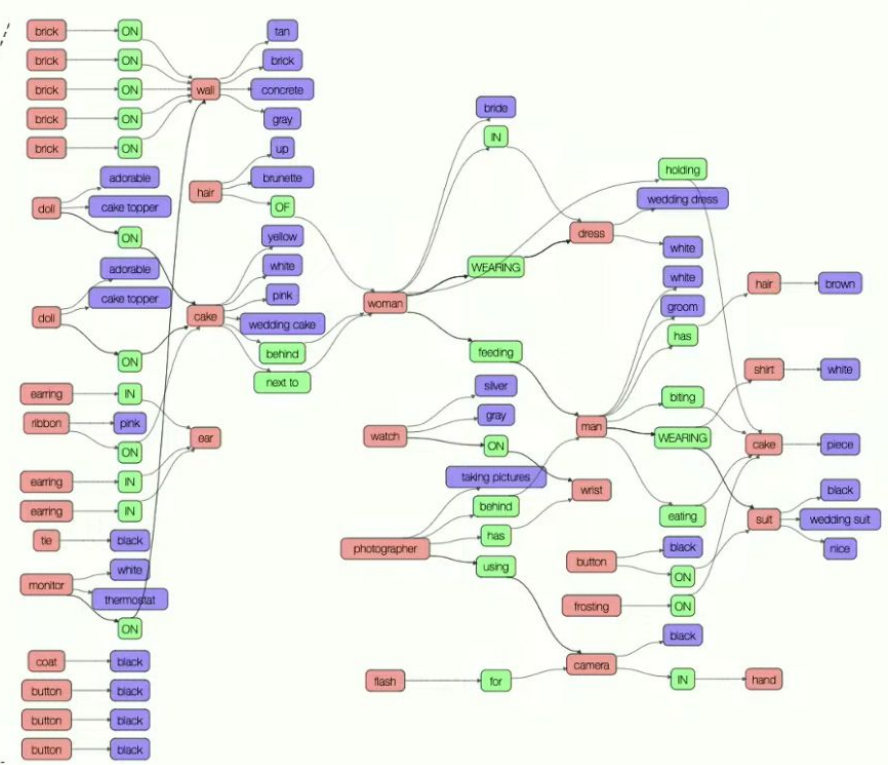




The scene graph representation of visual relationships







Krishna, Zhu, Hata, Johnson, Kravitz, Chen, Li, Shamma, Bernstein, Fei-Fei. IJCV 2017

Scene graph models are better at generalizing to novel compositions





person → sitting on → chair

948 training examples

+





fire hydrant → on → field

29 training examples

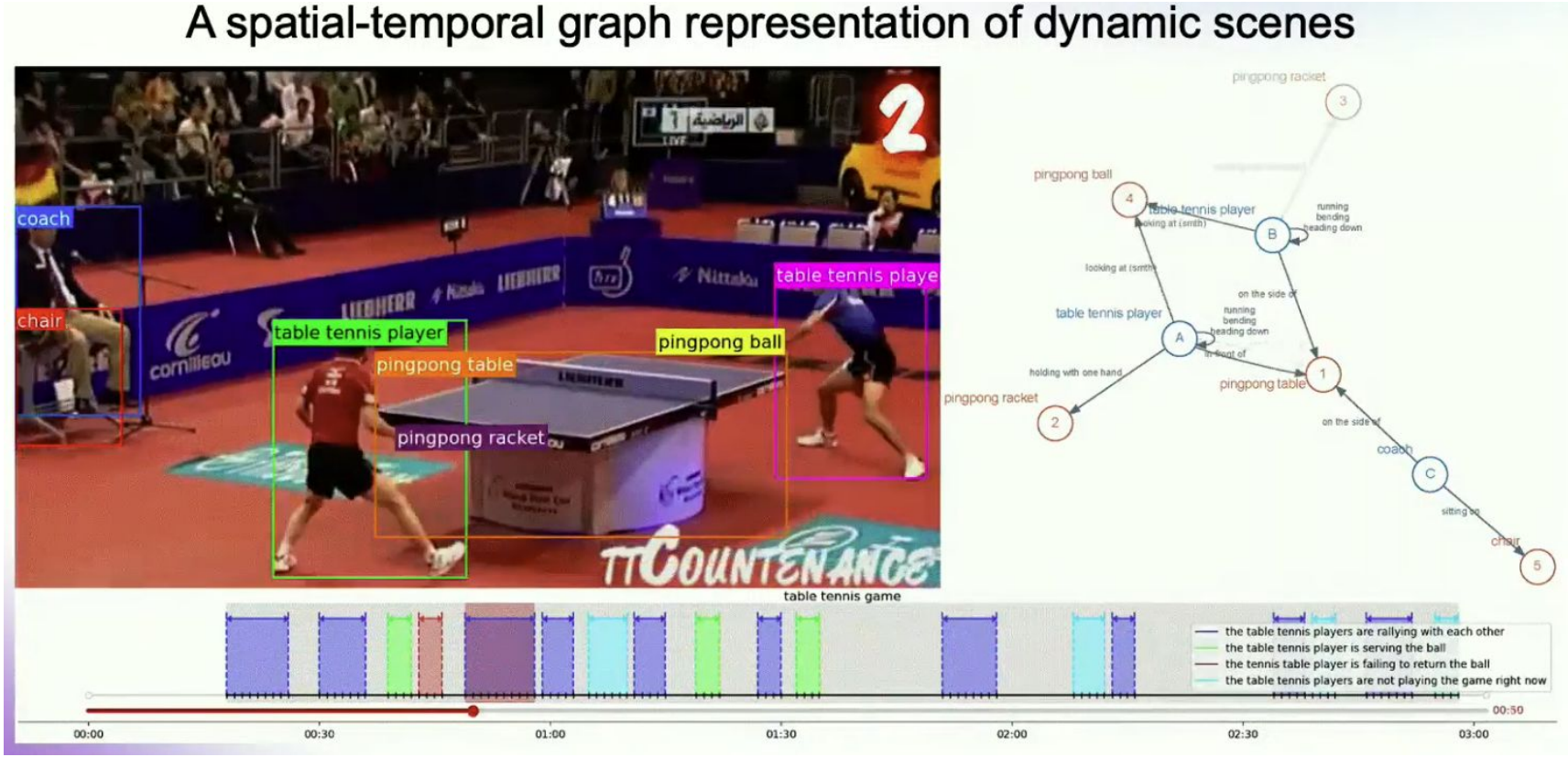


person → sitting on → fire hydrant

A novel composition.
dataset has 0 training examples.

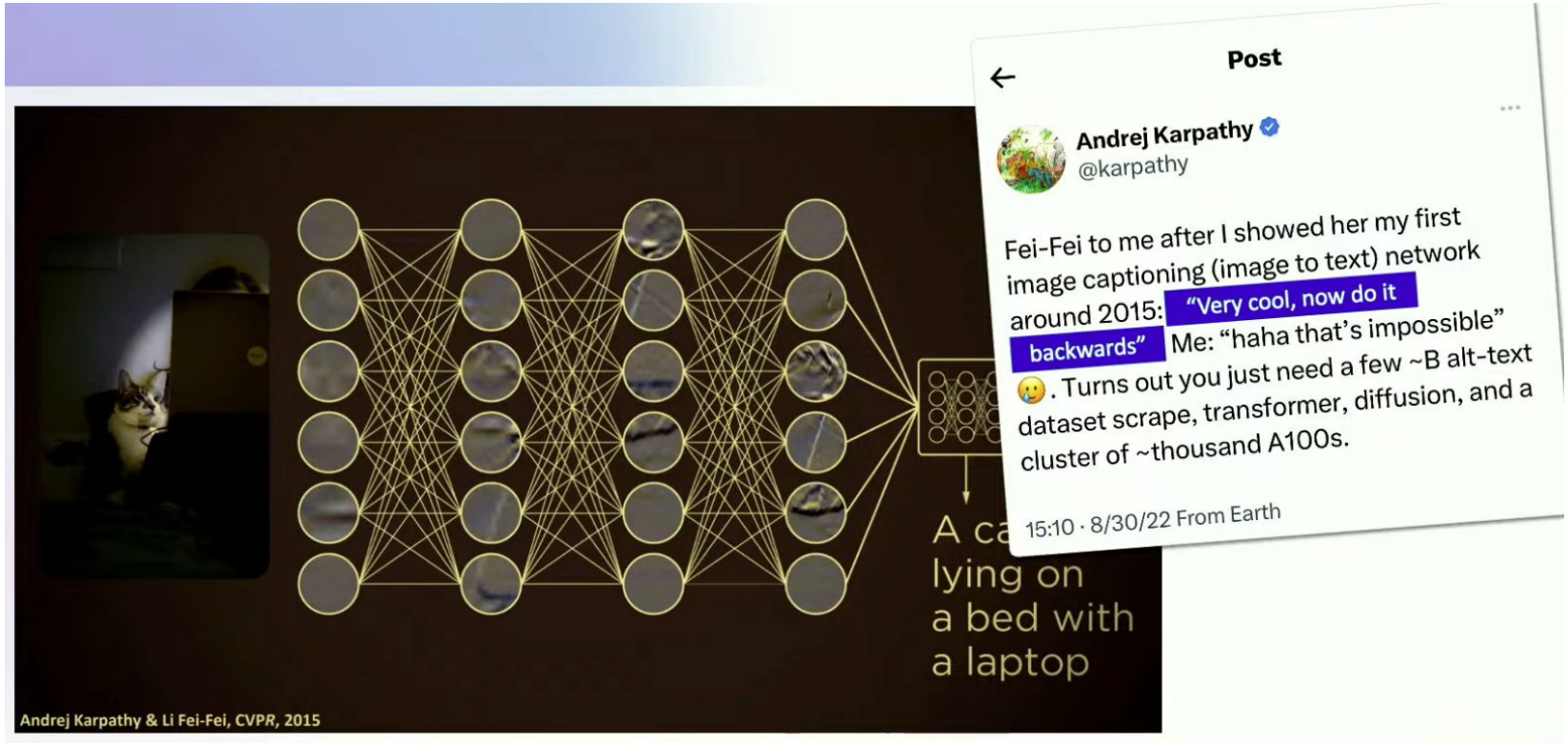
Lu*, Krishna*, Bernstein, Fei-Fei. ECCV 2016

Activity Understanding & Reasoning



Luo, Xie, Kapoor, Liang, Cooper, Niebles, Adeli & Fei-Fei, NeurIPS, 2021;
Luo, Durante, Li, Xie, Liu, Jin, Huang, Li, Wu, Niebles, Adeli, & Fei-Fei, NeurIPS, 2022

From Reasoning to Generation



Visual Intelligence entails

Understanding

Reasoning

Generation

Definition: creating new pixels or altering existing pixels

Capabilities: generate an image

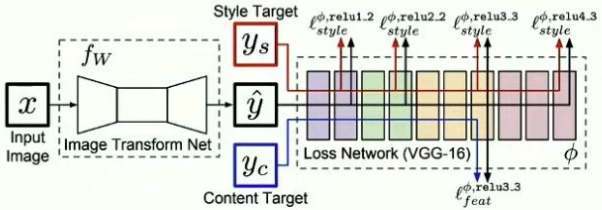
Example questions: draw a picture of a chair? Generate a “sunny beach of palm trees and beach volleyball game”

Artistic Style Transfer



Gatys et al. arXiv 2015, CVPR 2016
Johnson, Alahi, and Fei-Fei, ECCV 2016

Approximate iterative optimization with a feedforward network



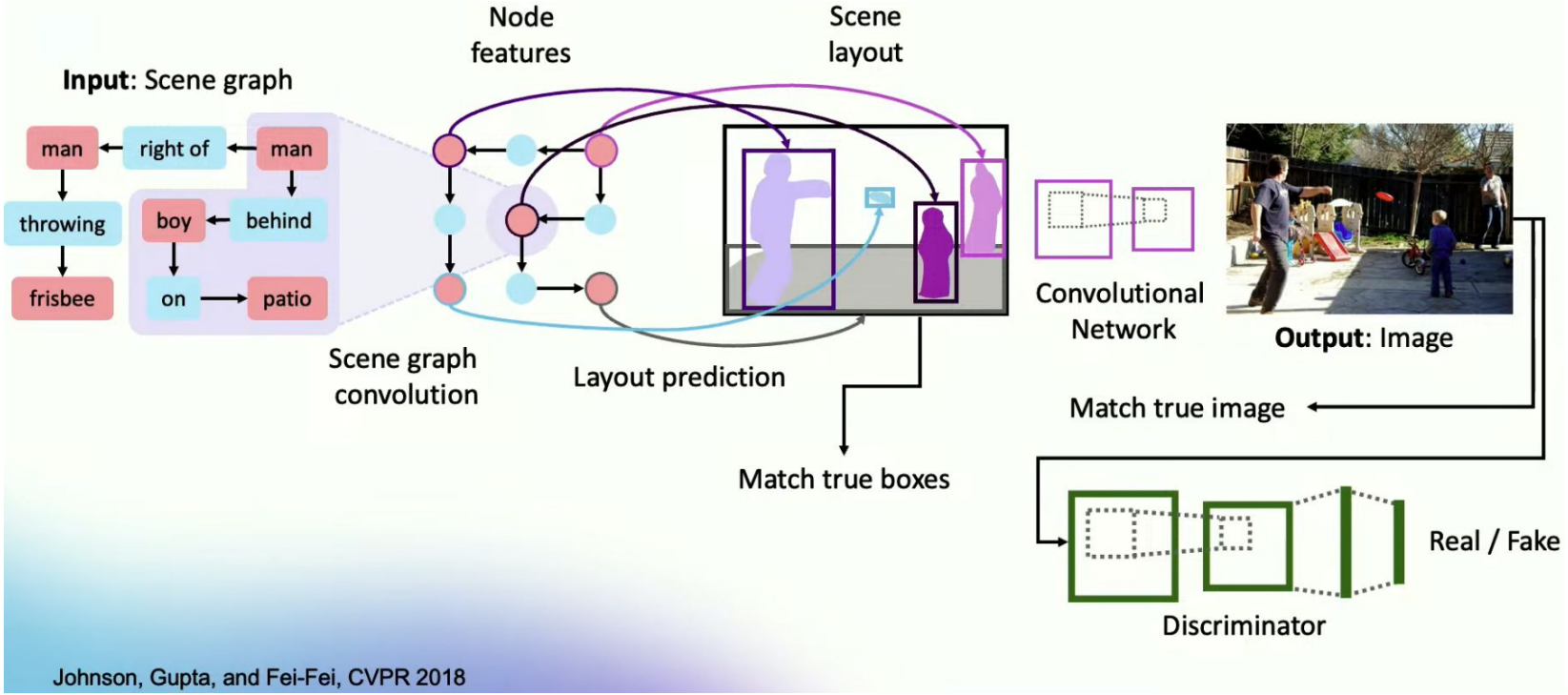
1000x speedup over prior work

Fast enough to run on phones

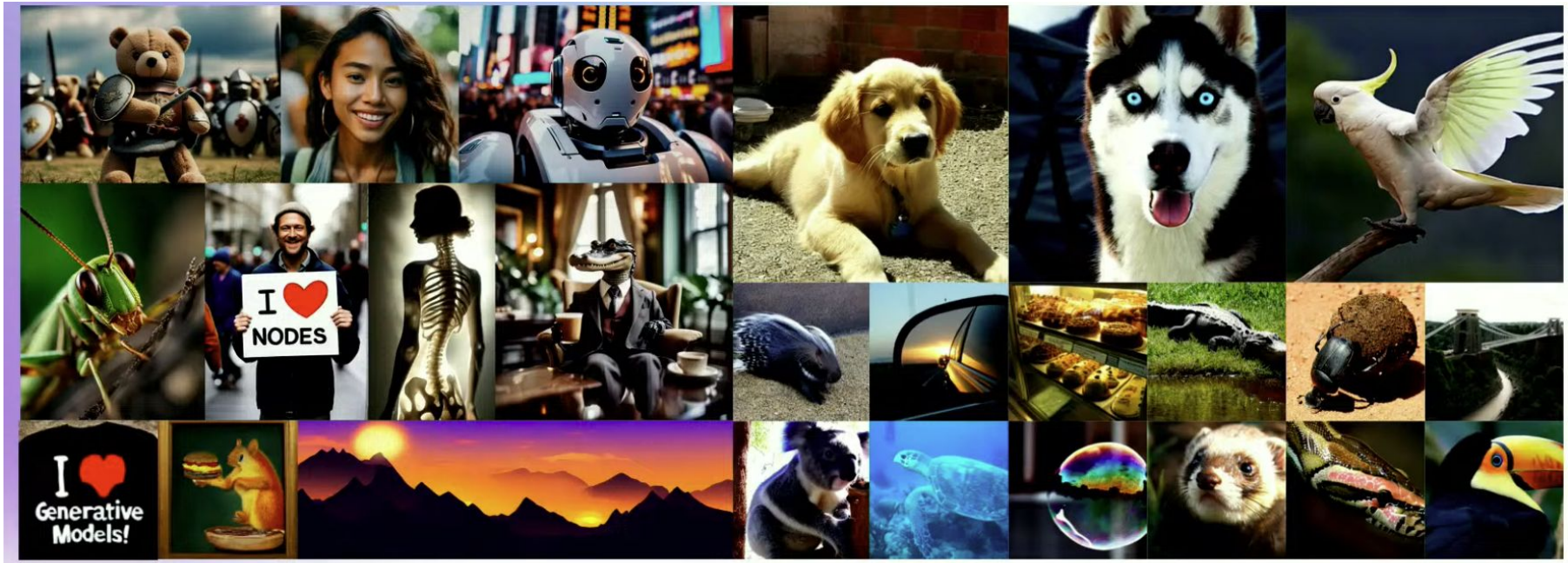
Apple 2016 App of the year



Scene Graph to Image: Model

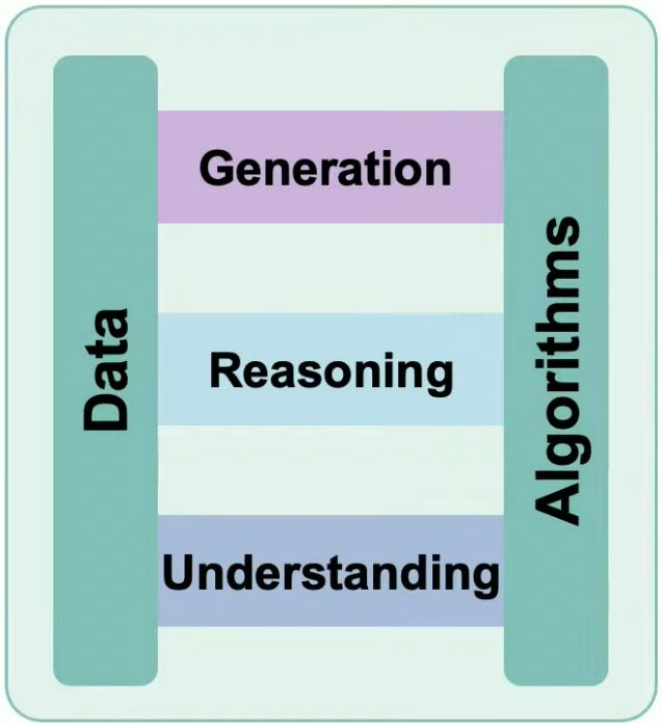


Recent Diffusion-based models



Dhariwal et al. Diffusion models beat gans on image synthesis. NeurIPS 2021; Ho et al. Denoising diffusion probabilistic models. NeurIPS 2020; Ramesh et al. Zero-shot text-to-image generation. ICML 2021; Song et al. Denoising Diffusion Implicit Models. ICLR 2021; Rombach et al. High-resolution image synthesis with latent diffusion models. CVPR 2022; Saharia et al. Photorealistic text-to-image diffusion models with deep language understanding. NeurIPS 2022; Peebles et al. Scalable diffusion models with transformers. ICCV 2023; Podell et al. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. ICLR 2024

“A flat earth” problem



interaction

Lots of graphics problem failure

Limited capability

Weird artifacts

Difficult reasoning example

Image generation

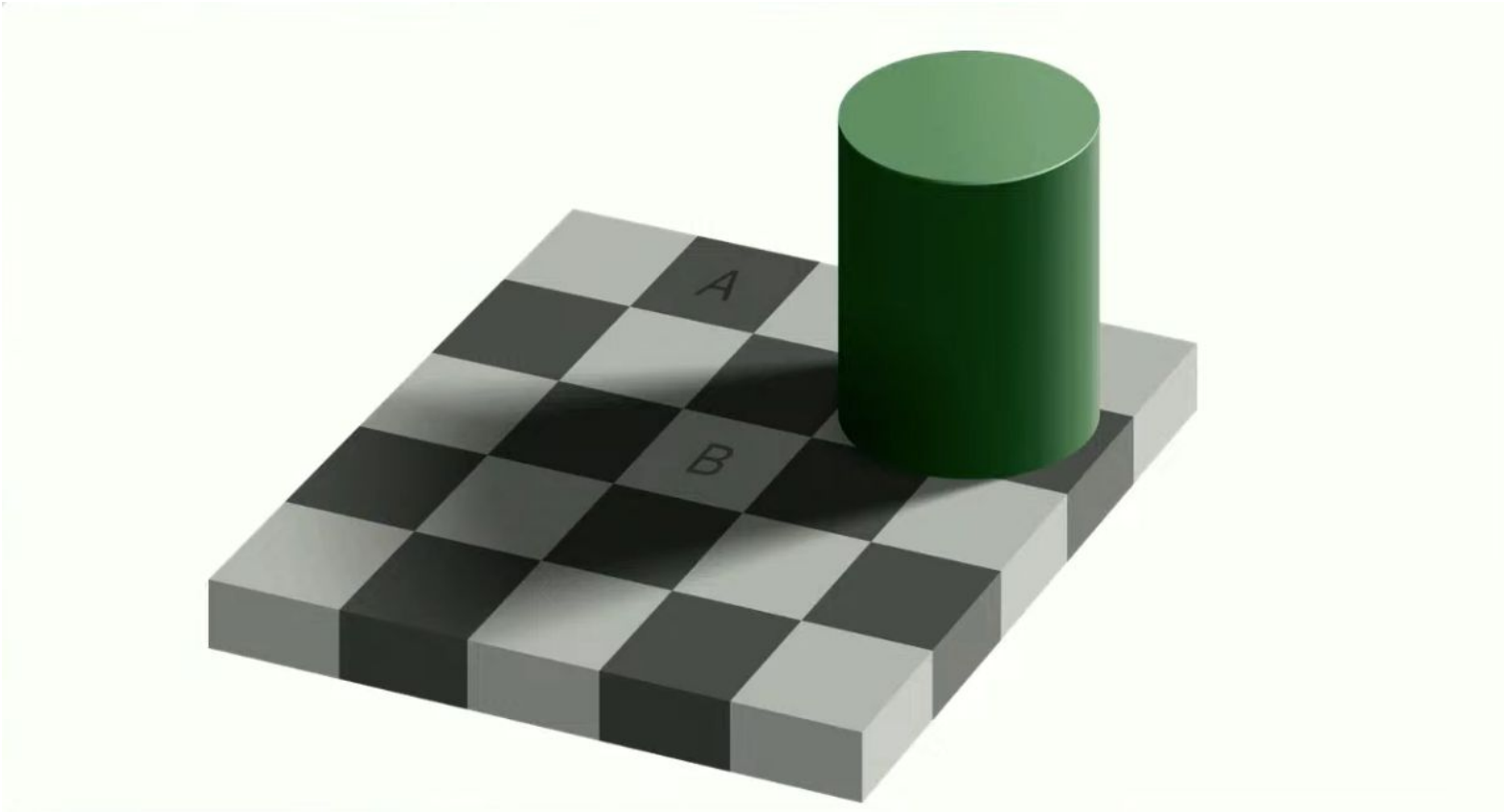
Hard to deal with occlusion

Limited ability to interact

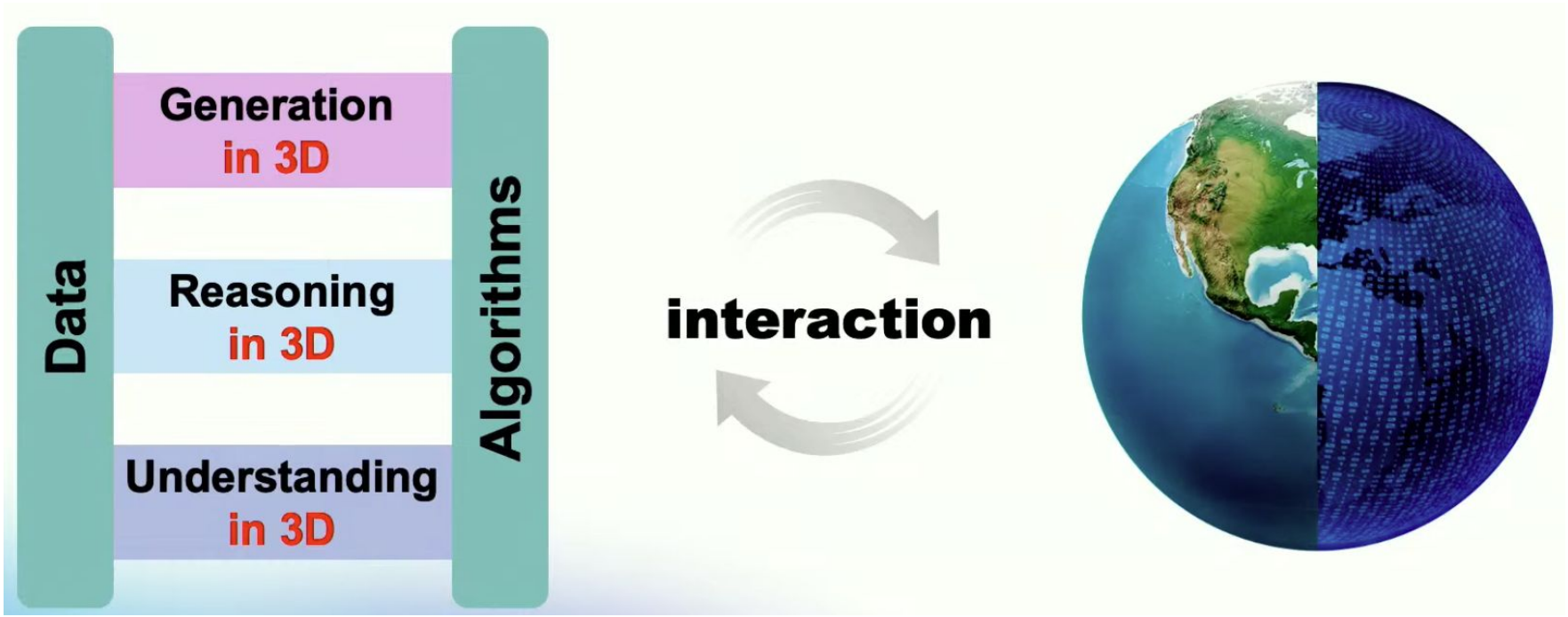
Lack of physics

...

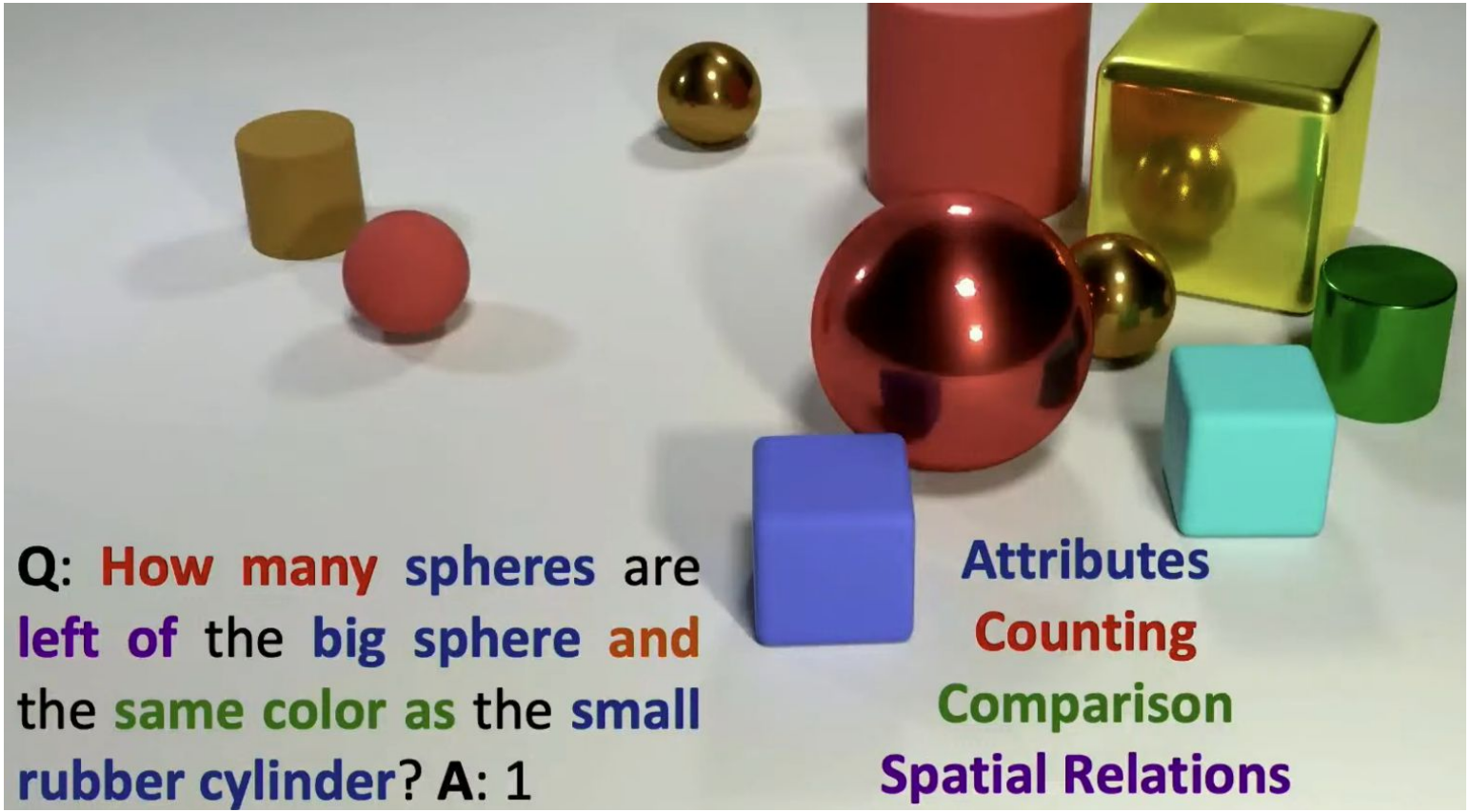
World is not flat!



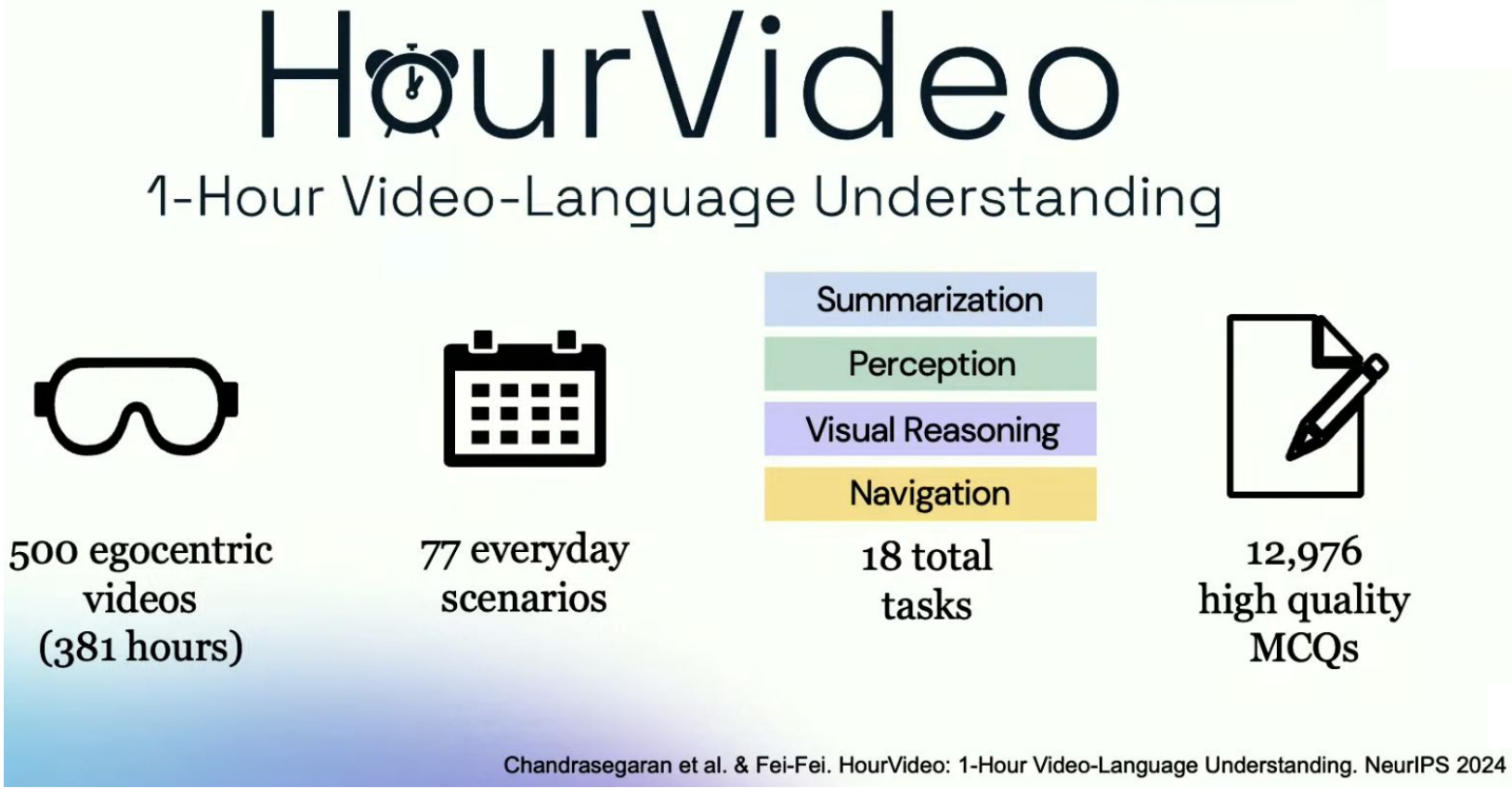
A new ladder to climb: Spatial Intelligence in 3D



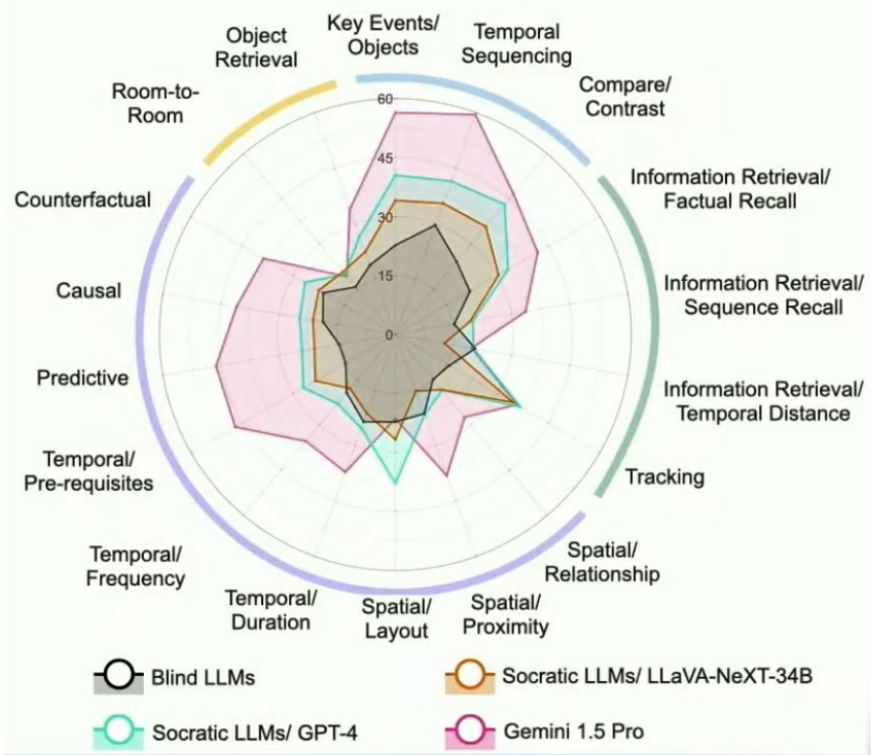
Understanding in 3D (CLEVR Dataset)



Understanding & Reasoning in 3D



How do multimodal models perform on the HourVideo Benchmark?



GPT-4 -> **25.7%**

Gemini 1.5 Pro -> **37.3%**

Human experts -> **85.0%.**

Significant gap in multimodal capabilities!

Chandrasegaran et al. & Fei-Fei. HourVideo: 1-Hour Video-Language Understanding. NeurIPS 2024

A renewed revolution in …

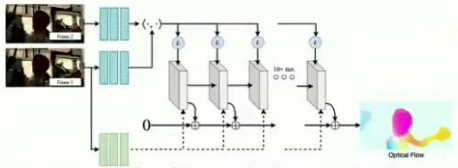
3D reconstruction

Classic Algorithms



Agarwal et al. ICCV 2009,
Curless & Levoy, 1996; Newcombe et al. 2011

Sparse Reconstruction



Teed and Deng, ECCV 2020; NeurIPS 2021

Dense Reconstruction



Mildenhall et al. ECCV 2020; Müller et al. Siggraph 2022;
Kerbl et al. ACM TOG 2023; Hong et al. ICLR 2024

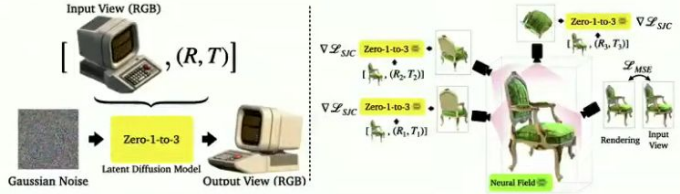
3D reconstruction + generation

DreamFusion



Poole et al. ICLR 2023

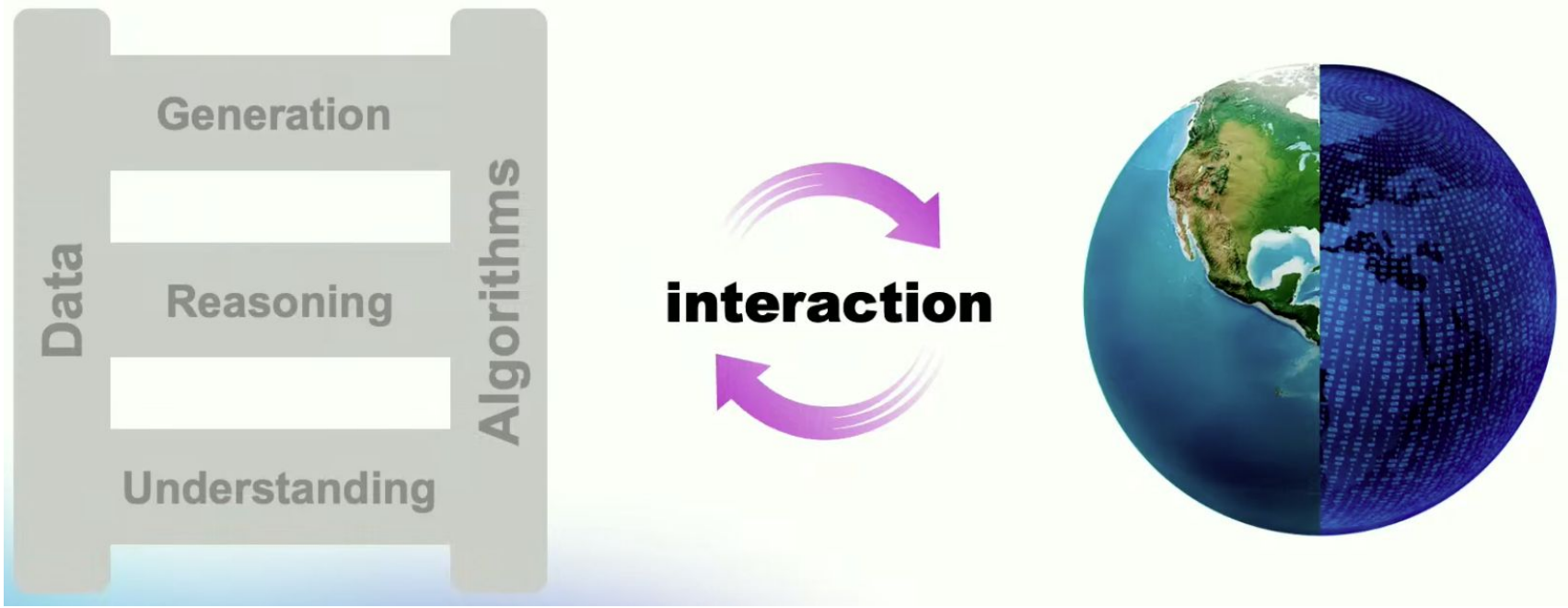
Zero123



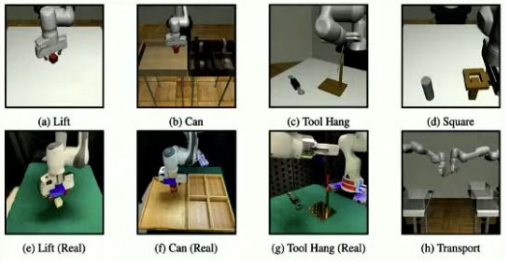
Novel View Synthesis 3D Reconstruction

Liu et al. ICCV 2023; Sitzmann et al. NeurIPS 2019; Chan et al. CVPR 2022

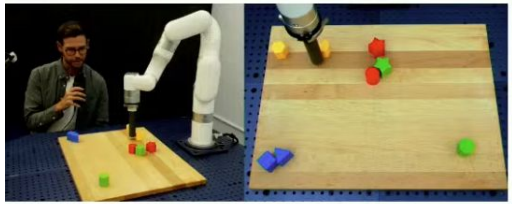
A new ladder to climb: Spatial Intelligence in 3D



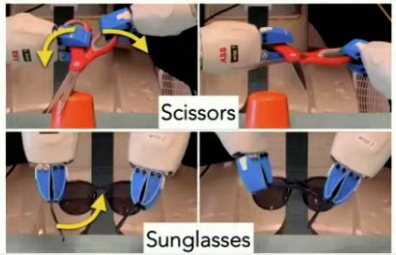
Most robotics research today: Still simple environments and tasks



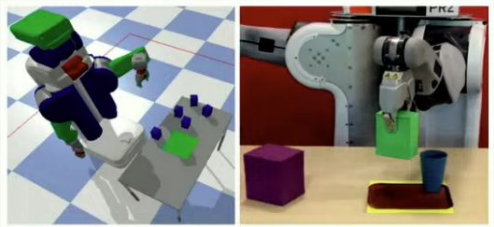
Mandlekar, A., Xu, D., Wong, J., Nasiriany, S., Wang, C., Kulkarni, R., Fei-Fei, L., Savarese, S., Zhu, Y. and Martín-Martín, R., 2021, CoRL



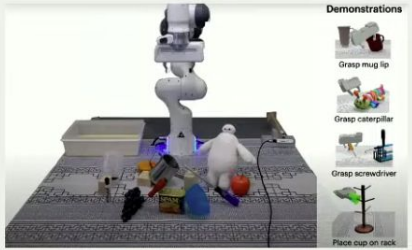
Lynch, C., Wahid, A., Tompson, J., Ding, T., Betker, J., Baruch, R., Armstrong, T. and Florence, P., 2023. RAL



Kerr, J., Kim, C.M., Wu, M., Yi, B., Wang, Q., Goldberg, K. and Kanazawa, A., 2024, CoRL



Garrett, C.R., Lozano-Pérez, T. and Kaelbling, L.P., 2020, ICAPS



Shen, W., Yang, G., Yu, A., Wong, J., Kaelbling, L.P. and Isola, P., 2023, CoRL



Wu, J., Antonova, R., Kan, A., Lepert, M., Zeng, A., Song, S., Bohg, J., Rusinkiewicz, S. and Funkhouser, T., 2023. Autonomous Robot

✗ Small scale

✗ Experimenter-picked tasks & setup

✗ Lack of standard benchmark & metrics

But the real world is

Ecologically complex



Dynamic & uncertain



Large variability



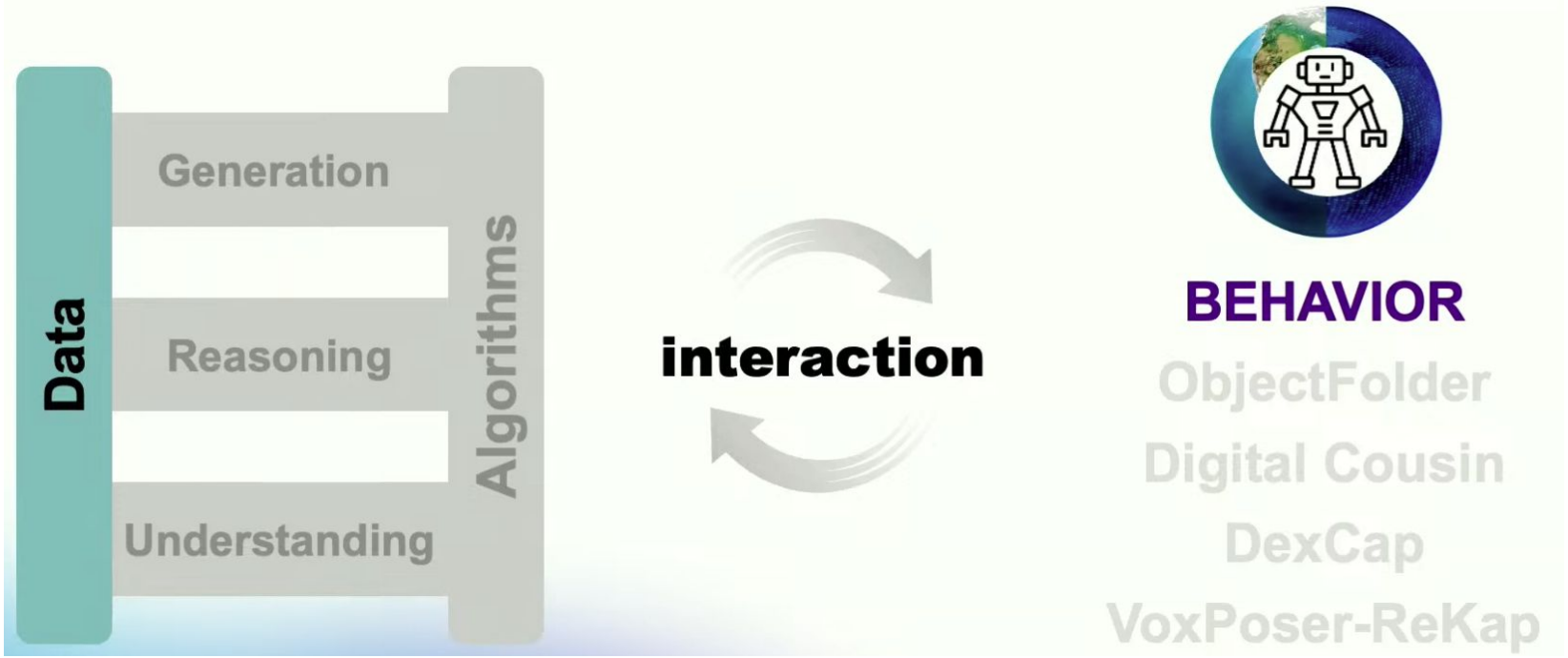
Interactive & social



Multitask



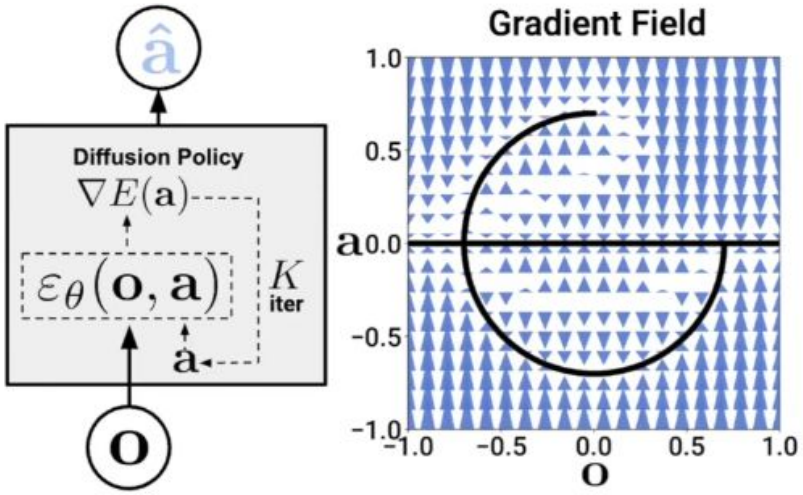
A new ladder to climb: Spatial Intelligence in 3D



Behavior-1K dataset & benchmark

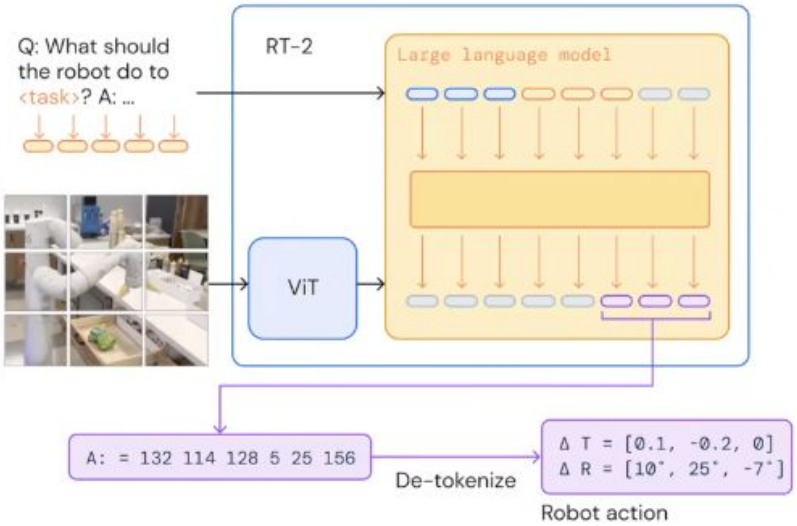


On the quest for open-world robots



Diffusion Policy

[Chi et al., RSS 2023]

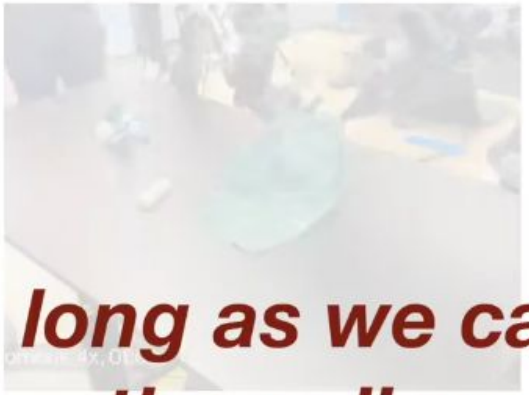


VLA Models

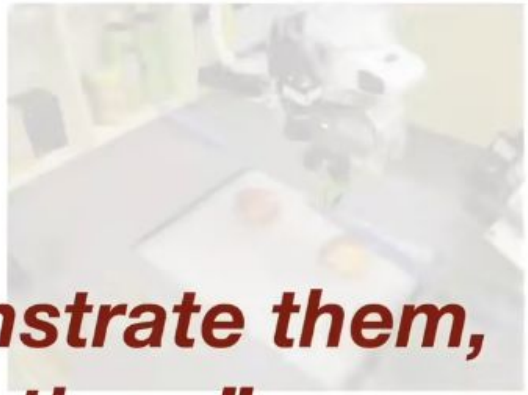
[Brohan et al., CoRL 2023]

On the quest for open-world robots

pi-0.5
Physical Intelligence



LBMs
TRI

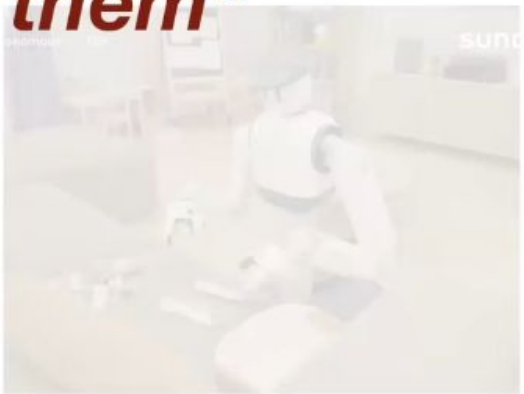


“As long as we can demonstrate them, the policy can learn them”

GEN-0
Generalist AI



ACT-1
Sunday Robotics



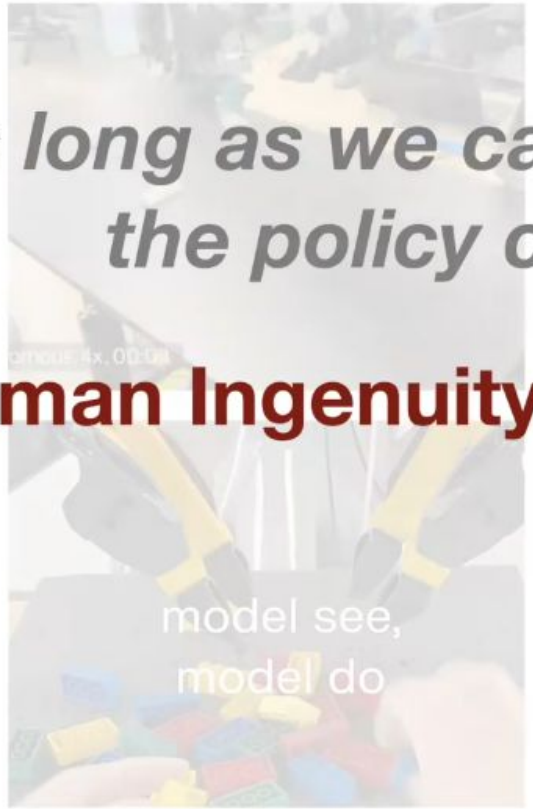
On the quest for open-world robots

pi-0.5
Physical Intelligence

LBMs

"As long as we can demonstrate them, the policy can learn them"

→ **Human Ingenuity from Data Collection**



GEN-0
Generalist AI

ACT-1
Sunday Robotics



Human Ingenuity from Data Collection





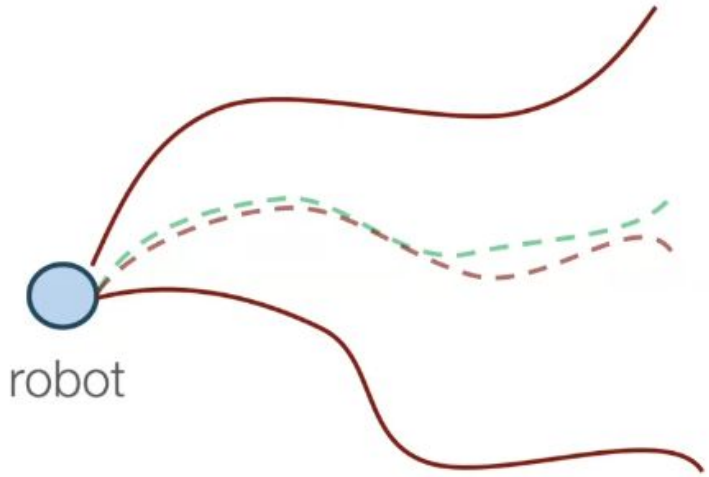
- wrapping sock on gripper to invert it
- mashing plastic bag on table to open it

Human Ingenuity from Data Collection



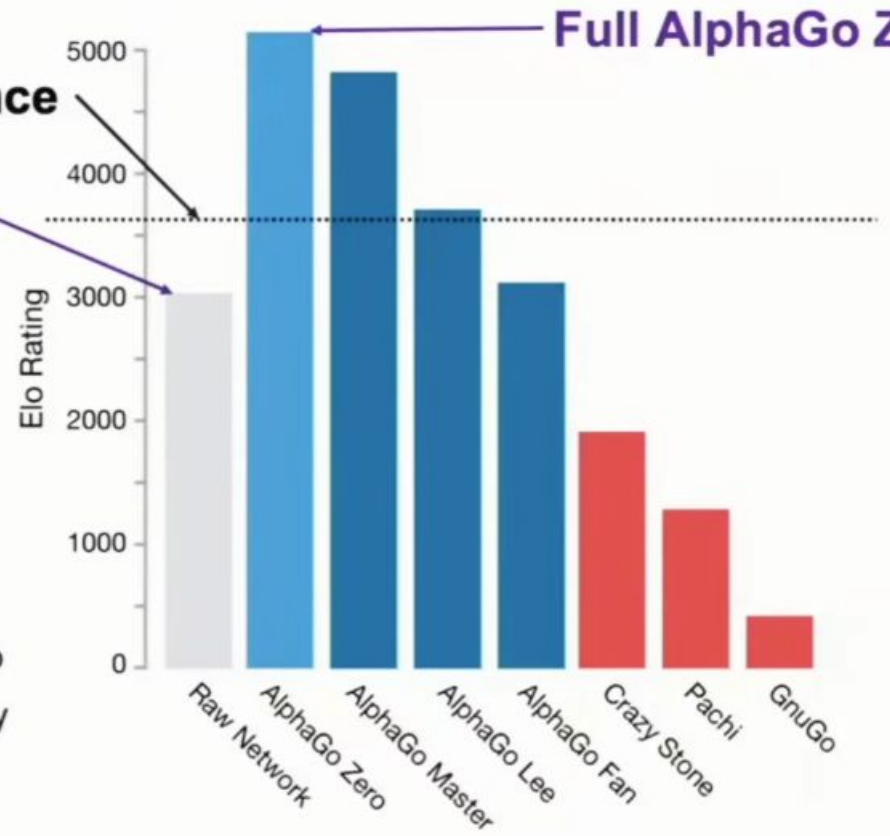
- wrapping sock on gripper to invert it
- mashing plastic bag on table to open it

Conterfactuals



What if I take a **very** different action?

Test-Time Counterfactual Reasoning

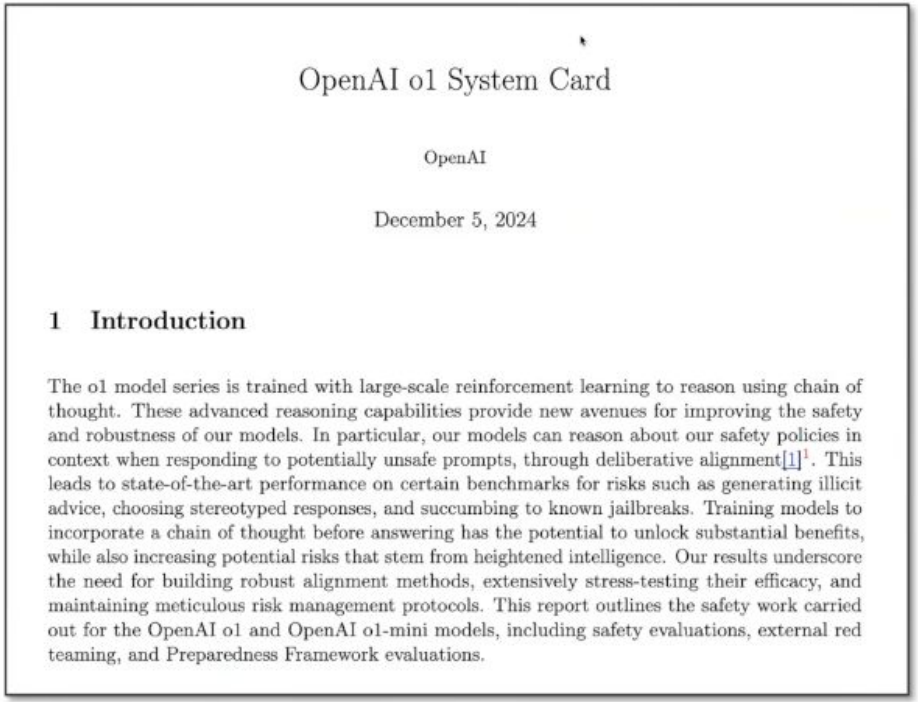


- Increasing Elo by 120 points requires either:
  - ~2x model size and training
  - ~2x test-time search
- To get the raw policy net from 3000 Elo to 5200 Elo, you would need to scale by **~100,000x**



Noam Brown

Test-Time Counterfactual Reasoning



OpenAI o1 System Card

OpenAI

December 5, 2024

1 Introduction

The o1 model series is trained with large-scale reinforcement learning to reason using chain of thought. These advanced reasoning capabilities provide new avenues for improving the safety and robustness of our models. In particular, our models can reason about our safety policies in context when responding to potentially unsafe prompts, through deliberative alignment[1][1]. This leads to state-of-the-art performance on certain benchmarks for risks such as generating illicit advice, choosing stereotyped responses, and succumbing to known jailbreaks. Training models to incorporate a chain of thought before answering has the potential to unlock substantial benefits, while also increasing potential risks that stem from heightened intelligence. Our results underscore the need for building robust alignment methods, extensively stress-testing their efficacy, and maintaining meticulous risk management protocols. This report outlines the safety work carried out for the OpenAI o1 and OpenAI o1-mini models, including safety evaluations, external red teaming, and Preparedness Framework evaluations.

- **scaling pretraining**: understand how the world works
- **scaling test-time compute**: compose that understanding for counterfactual reasoning to generate novel behaviors

Test-Time Counterfactual Reasoning



Robotics is the interdisciplinary study and practice of the design, construction, operation, and use of robots.[1] A roboticist is someone who specializes in robotics.[2] Robotics usually combines four aspects of design work: a power source (e.g. a battery), mechanical construction, a control system (electrical



So what's we needed?

Robotics is the interdisciplinary study and practice of the design, construction, operation, and use of robots.[1] A roboticist is someone who specializes in robotics.[2] Robotics usually combines four aspects of design work: a power source (e.g. a battery), mechanical construction, a control system (electrical circuits)...

Reward Model

code execution
math questions
...

World Model

text as state
text as action
-> text as next state

So what's we needed?



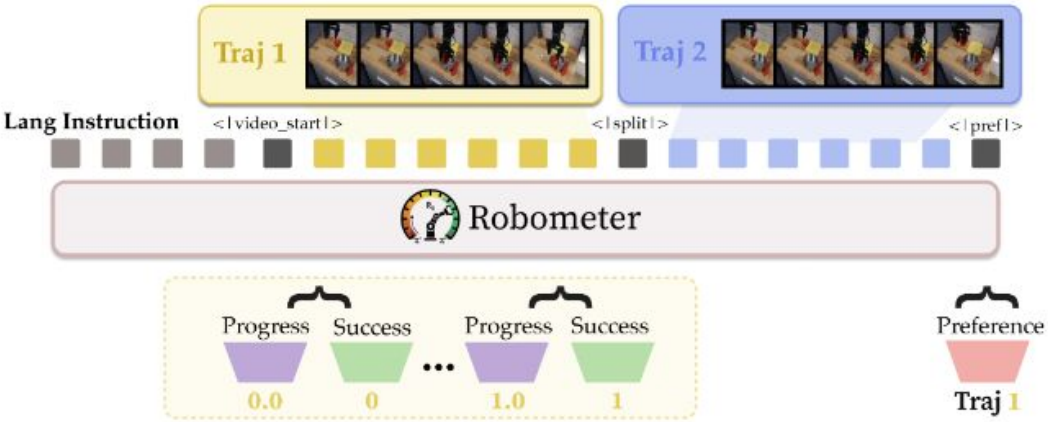
Reward Model

game outcome

World Model

state: stone placement
action: stone placement
-> next state: stone placement

Reward Model for Robotics



Dual-Objective Training

Frame-Level Progress and Success Prediction
Anchors the magnitude of predicted rewards by regressing per-frame progress and success values on expert demonstration data, providing dense supervision at the intra-trajectory level.

Trajectory-Comparison Preference Prediction
Imposes global ordering constraints across pairs of trajectories of the same task, enabling robust learning from suboptimal and failed data via a preference comparison objective.

So what's we needed?



Reward Model

???

World Model

???

World Models for Robotics ?

Text as Action



Sora 2

Navigation as Action



Genie 3

World Models for Robotics ?



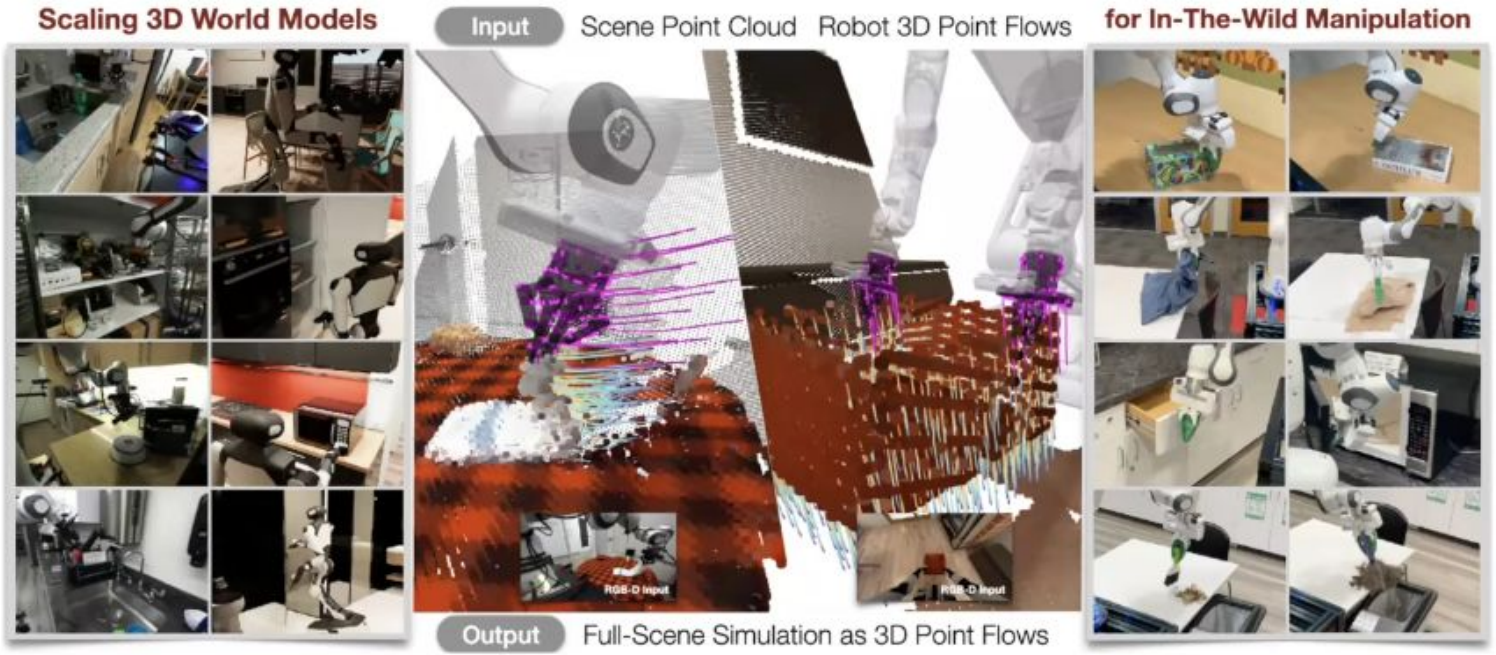
which angle?



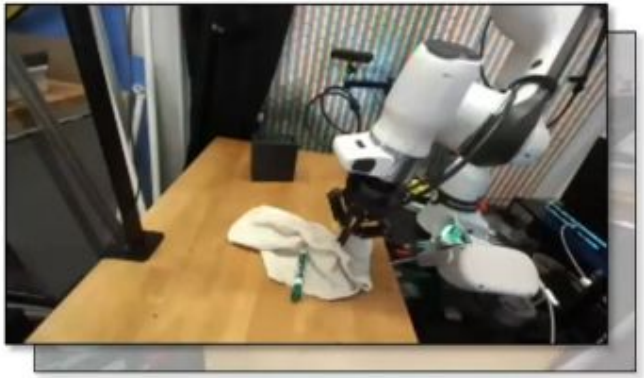
how much force?

PointWorld: Scaling 3D World Models for In-The-Wild Robotic Manipulation

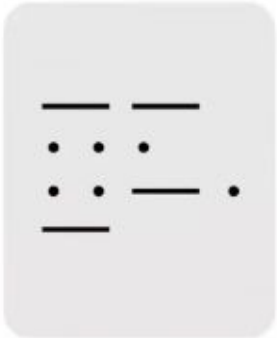
Wenlong Huang, Yu-Wei Chao, Arsalan Mousavian,
Ming-Yu Liu, Dieter Fox, Kaichun Mo*, Li Fei-Fei* (*Equal advising)



What we have as input



RGB-D Capture

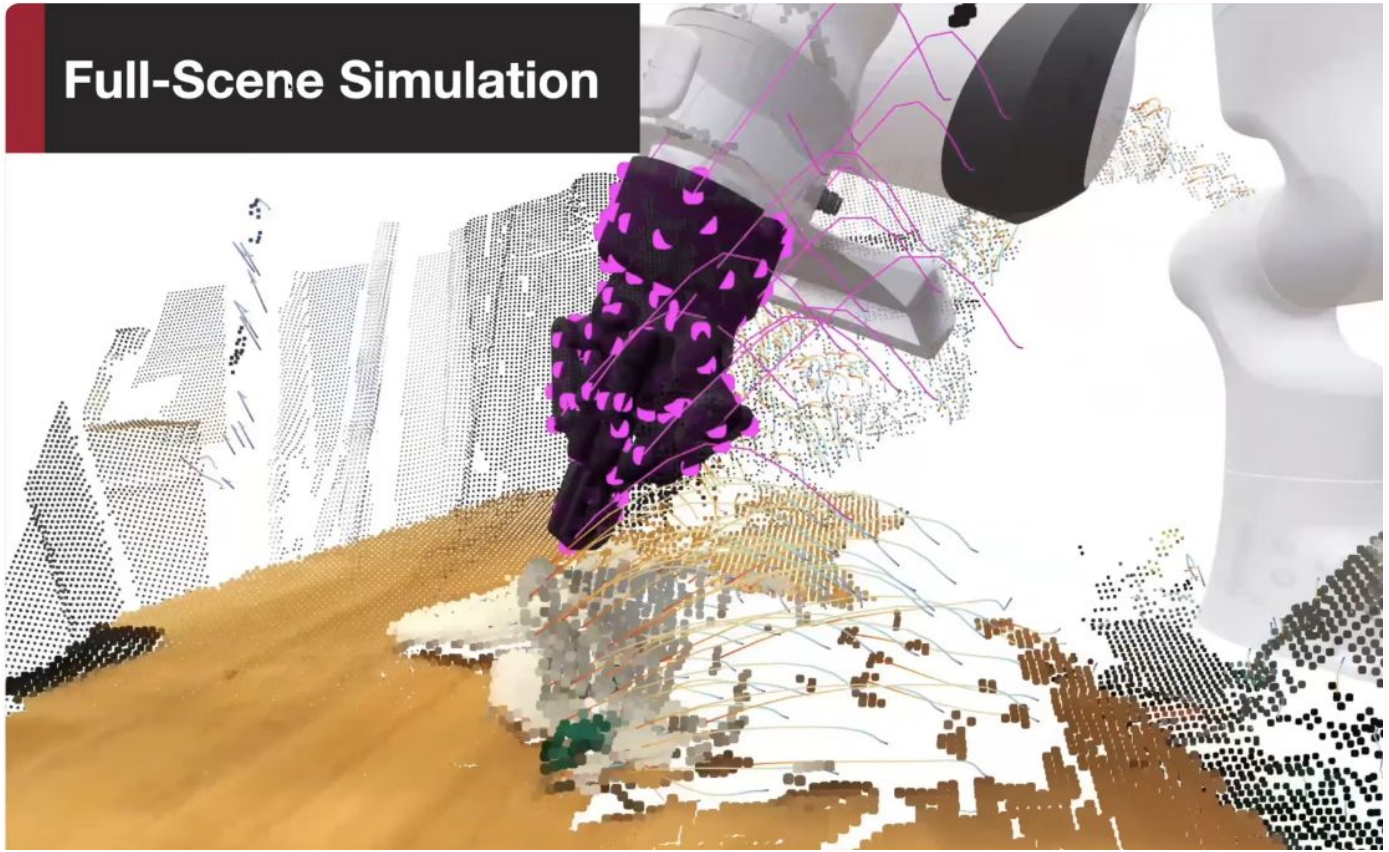


Robot URDF

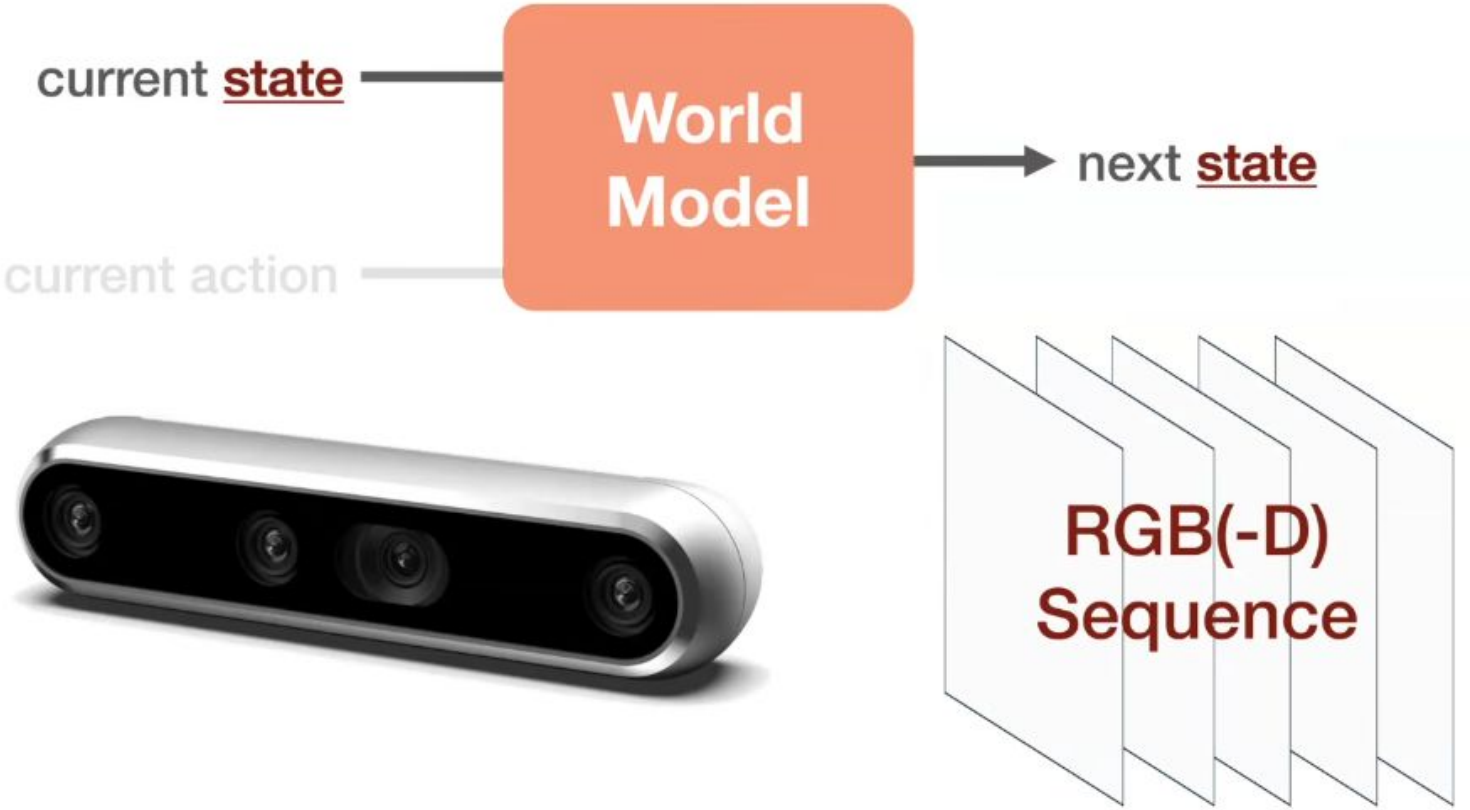


Robot Actions

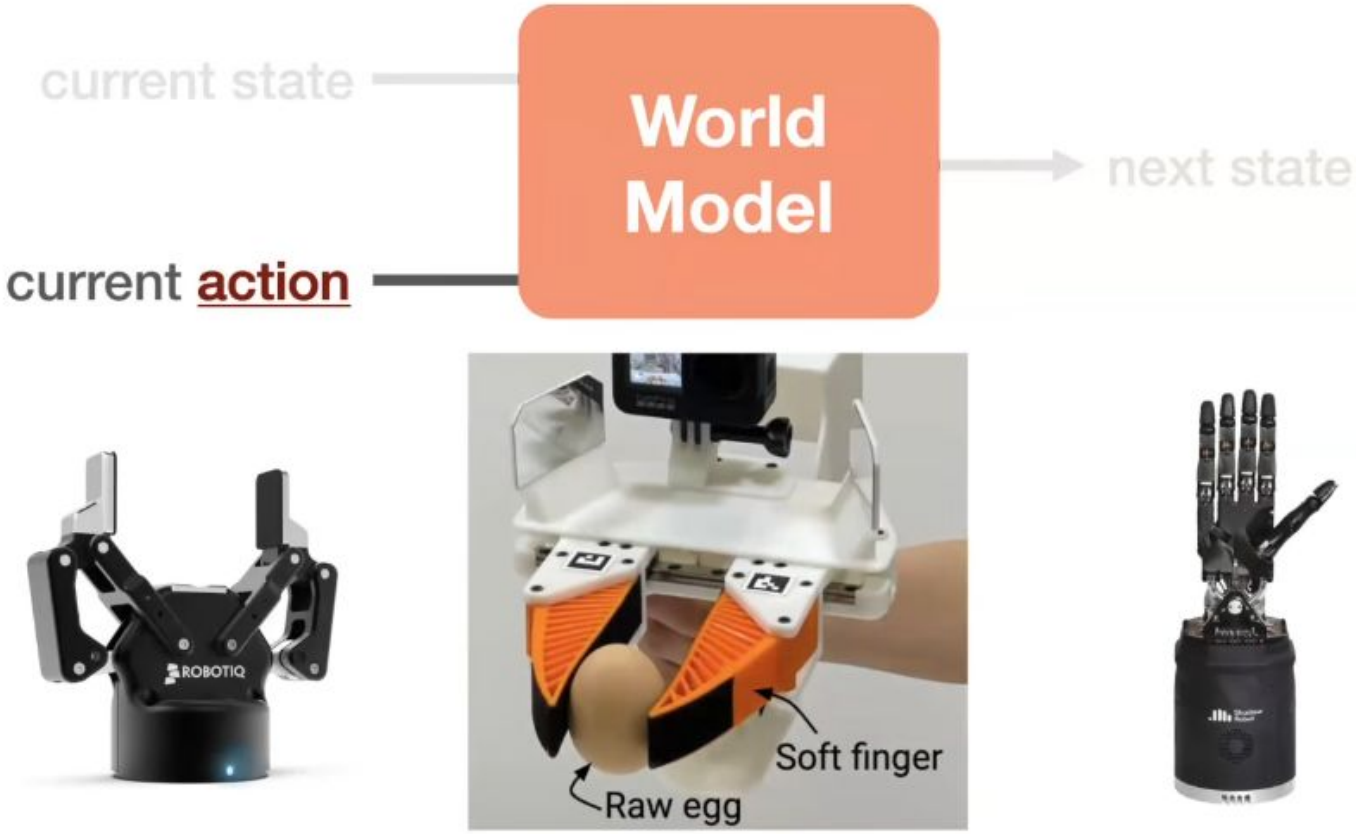
What we want as output



Why 3D world model?



Why 3D world model?



The “All-or-Nothing Embodiment Hypothesis”

> Robot learning *capability breakthroughs* will eventually *only* come from *multi-embodiment* models.

Psuedoproof:

i. Do you ever want to change your robot hardware?
Even just to tweak it?
Or: will you be forced to? (supply chain, manufacturing...)

Yes.

ii. When you do change hardware, do you want to throw away all your previous data?

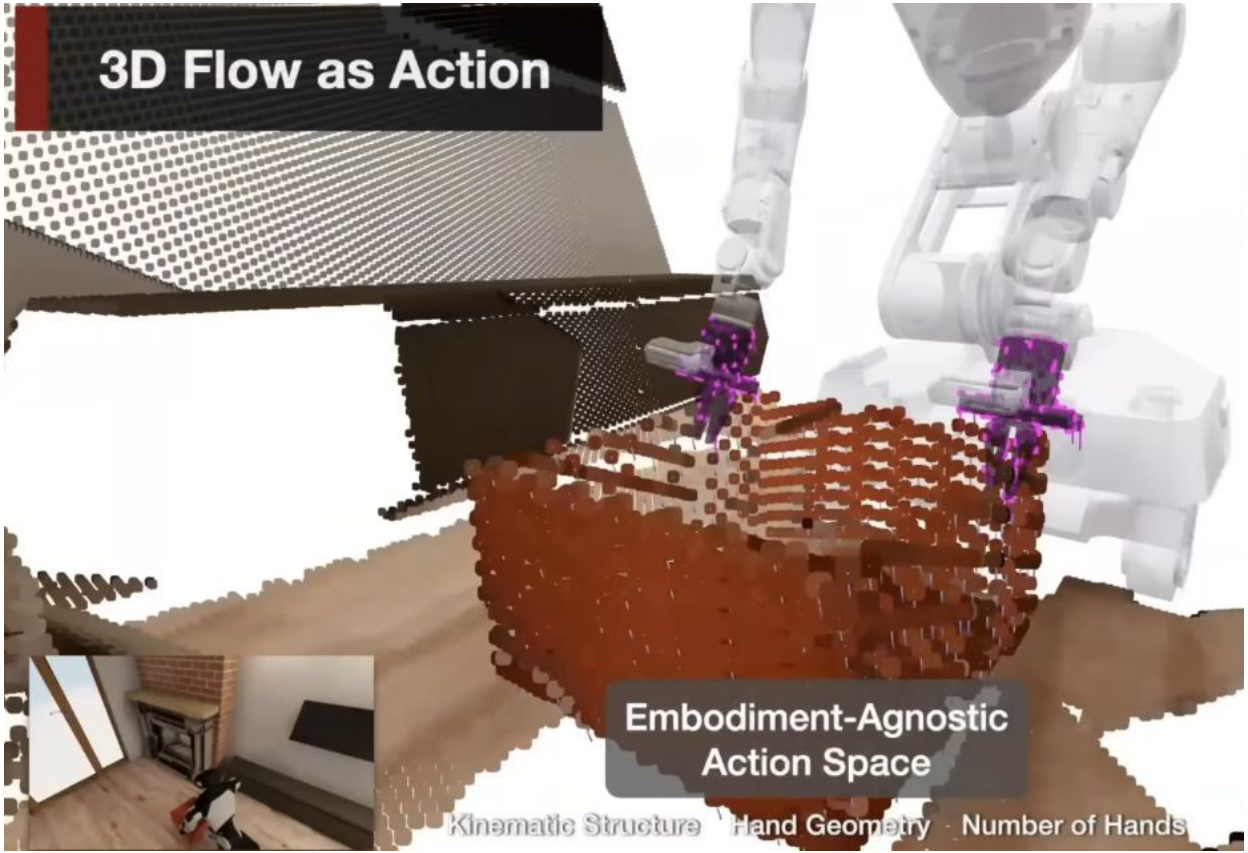
No.

iii. The above 2 imply the hypothesis may be true.



Pete Florence

Solution of 3D world model

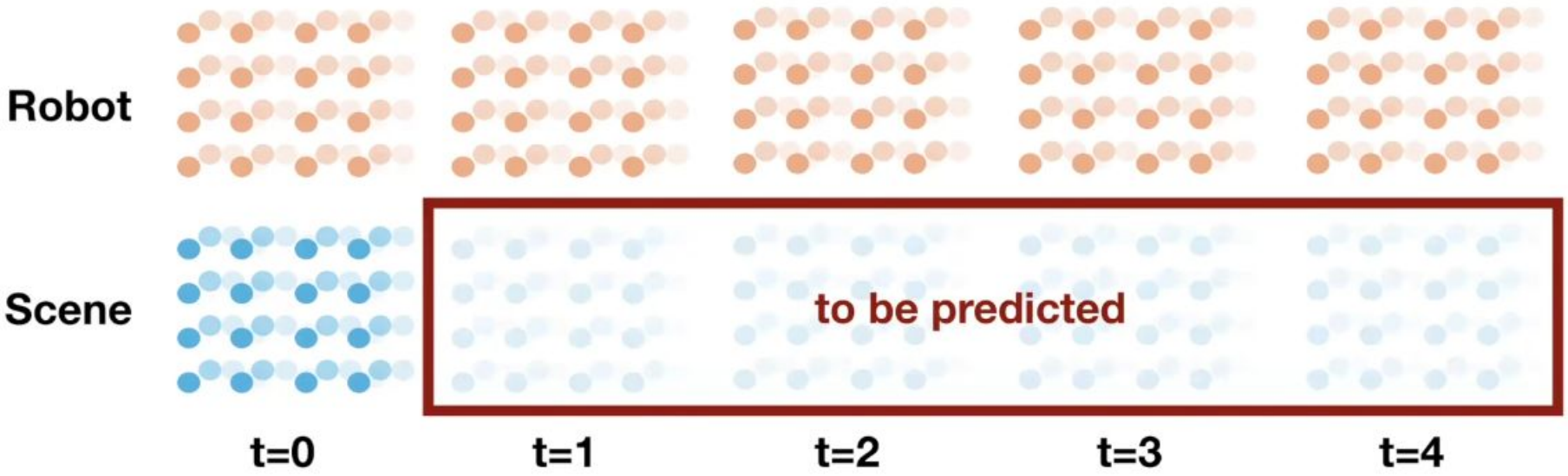


3D Point Flows from Robot URDF

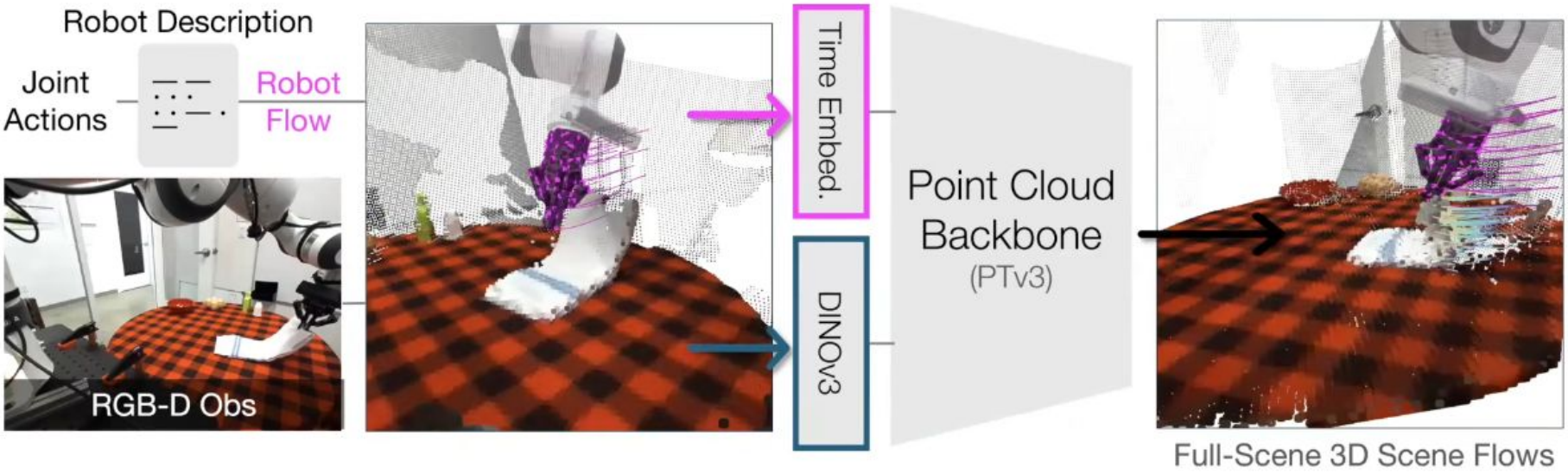


1. Sample N points on each link
2. Compute forward kinematics
3. Transform the points for each link

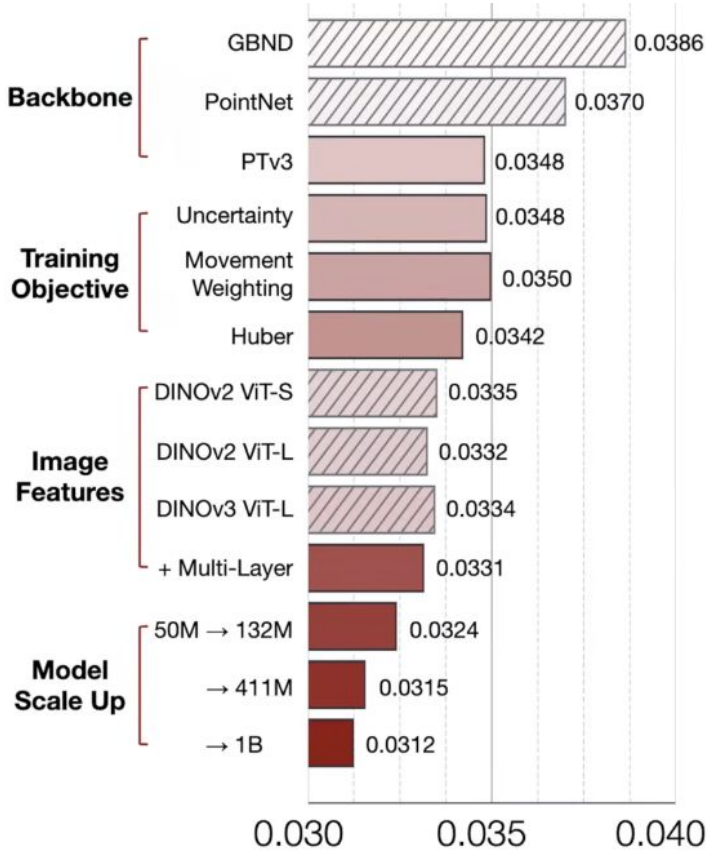
PointWorld Input-Output Formulation



PointWorld-1B (a zero-shot real-world simulator for robots)



PointWorld-1B (a zero-shot real-world simulator for robots)

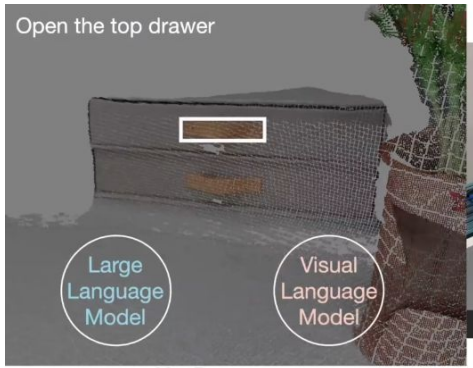


The Roadmap for Scaling 3D World Models

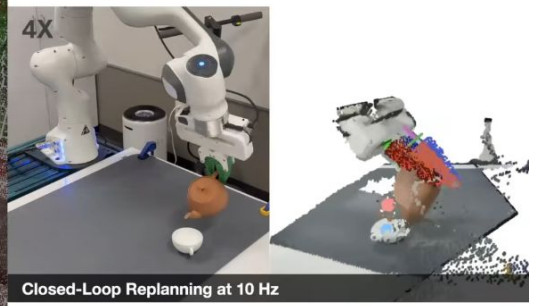
How can counterfactual reasoning enable robot control?

Training RL Policies Inside Imagination (Dreamer 4)

Planning through Imagination



VoxPoser
[Huang et al., 2023]



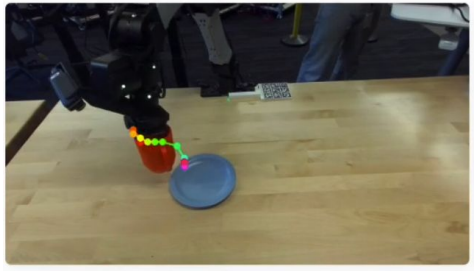
ReKep
[Huang et al., 2023]

MPC with PointWorld



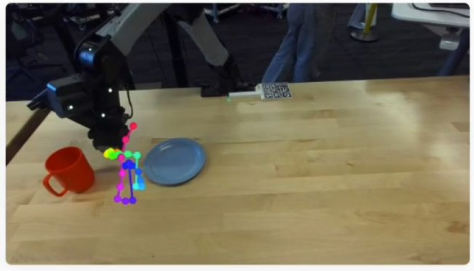
Deformable Object Manipulation

Robot policies are spatially dumb



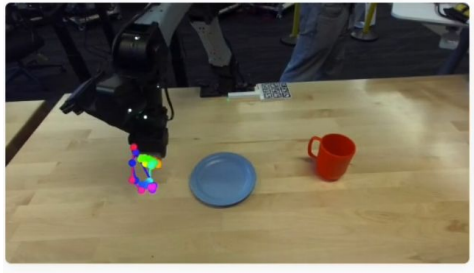
TASK **Simple Task**

Pick up the cup, place it on the plate.



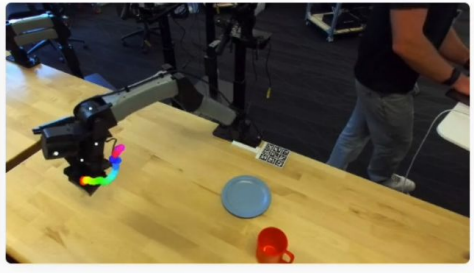
FAIL **Fails even at 70 demos.**

ACT trained on 70 episodes of the pick-and-place task — directionally correct but not precise.



FAIL **Object OOD Pos**

Train object on left of table, test on right. Policy doesn't approach the object.



FAIL **Camera OOD Viewpoint**

New camera viewpoint — policy doesn't reach towards the object.

Data-inefficient and spatially fragile — hundreds of demos for a simple pick-and-place task; fails when the camera is bumped or the object is OOD.

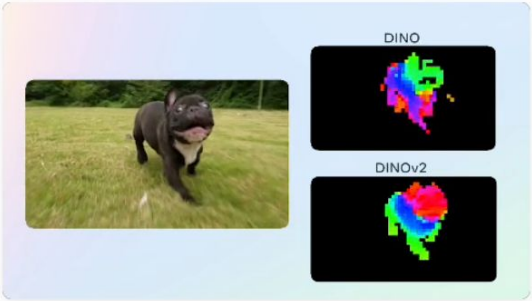
Why this sucks:

- Keeping the camera fixed is really hard — people bump into it often.
- "Just cover the distribution over all object positions and viewpoints": collecting that much data is hard; teleop is slow.

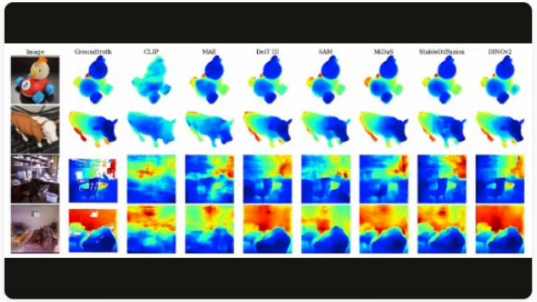
The weird part - image features are already spatially smart



DINO features PCA on our pick-and-place scene — semantically consistent across positions + viewpoints.



Classic DINO dog — strong semantic correspondence across viewpoints.



Probing frozen ViTs — depth, normals, camera pose recoverable from features never trained on them.

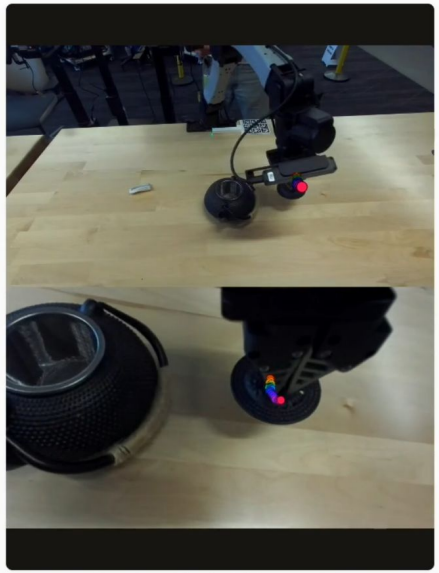
Off-the-shelf modern image features are semantically rich and geometrically robust.

Not just semantic — probing studies show these features contain strong geometry (depth, normals, camera pose) as well.

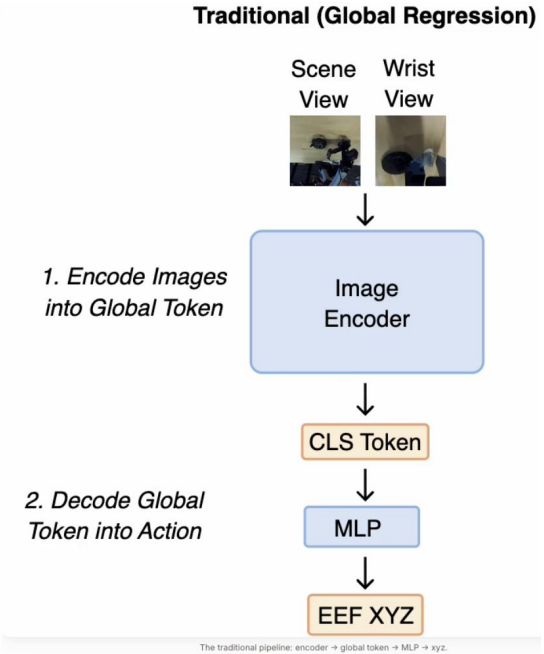
What's the gap?

$$\text{pixel}(xyz_{\text{robot}}) = \pi(K, T_{\text{robot}\to\text{cam}}, xyz_{\text{robot}})$$

The 3D-to-2D projection an oracle action head implicitly would have to learn.



Teapot task demonstration with end-effector keypoints overlaid.



The traditional pipeline: encoder → global token → MLP → xyz.

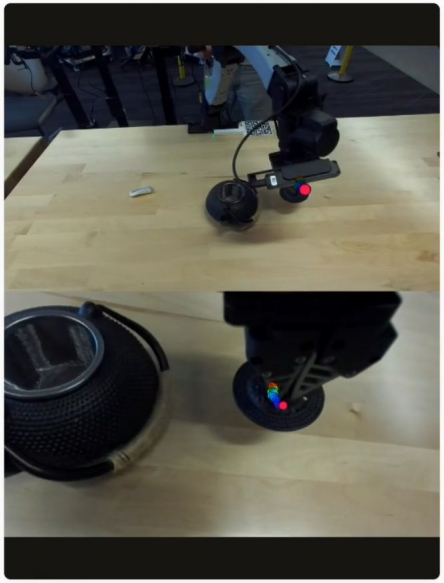
If DINO patches implicitly encode per-pixel semantics + xyz geometry, why is "DINO patches → robot EEF XYZ" so hard?

One intuition for why: Policy weights have to learn the relationship between *EEF XYZ targets* and *corresponding image patches/features*.

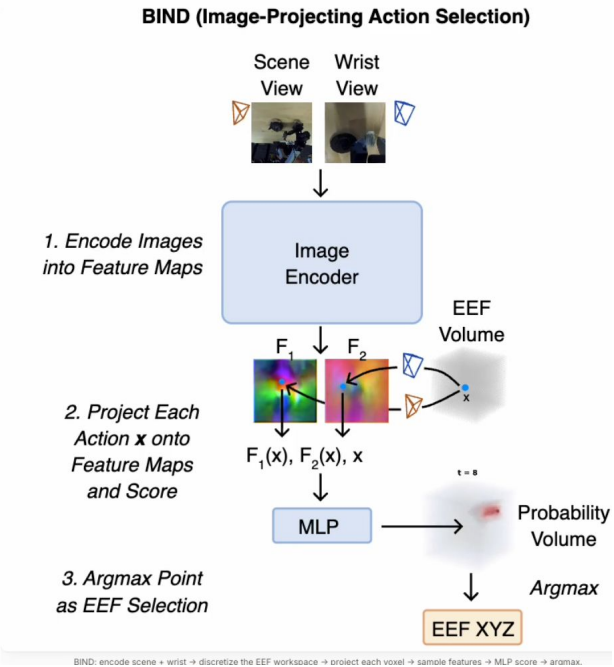
How to bridge this gap - bring the 3D robot actions to their corresponding 2D image patches

$$\text{pixel}(xyz_{\text{robot}}) = \pi(K, T_{\text{robot}\to\text{cam}}, xyz_{\text{robot}})$$

The 3D-to-2D projection an oracle action head implicitly would have to learn.



Teapot task demonstration with end-effector keypoints overlaid.



BIND: encode scene + wrist → discretize the EEF workspace → project each voxel → sample features → MLP score → argmax.

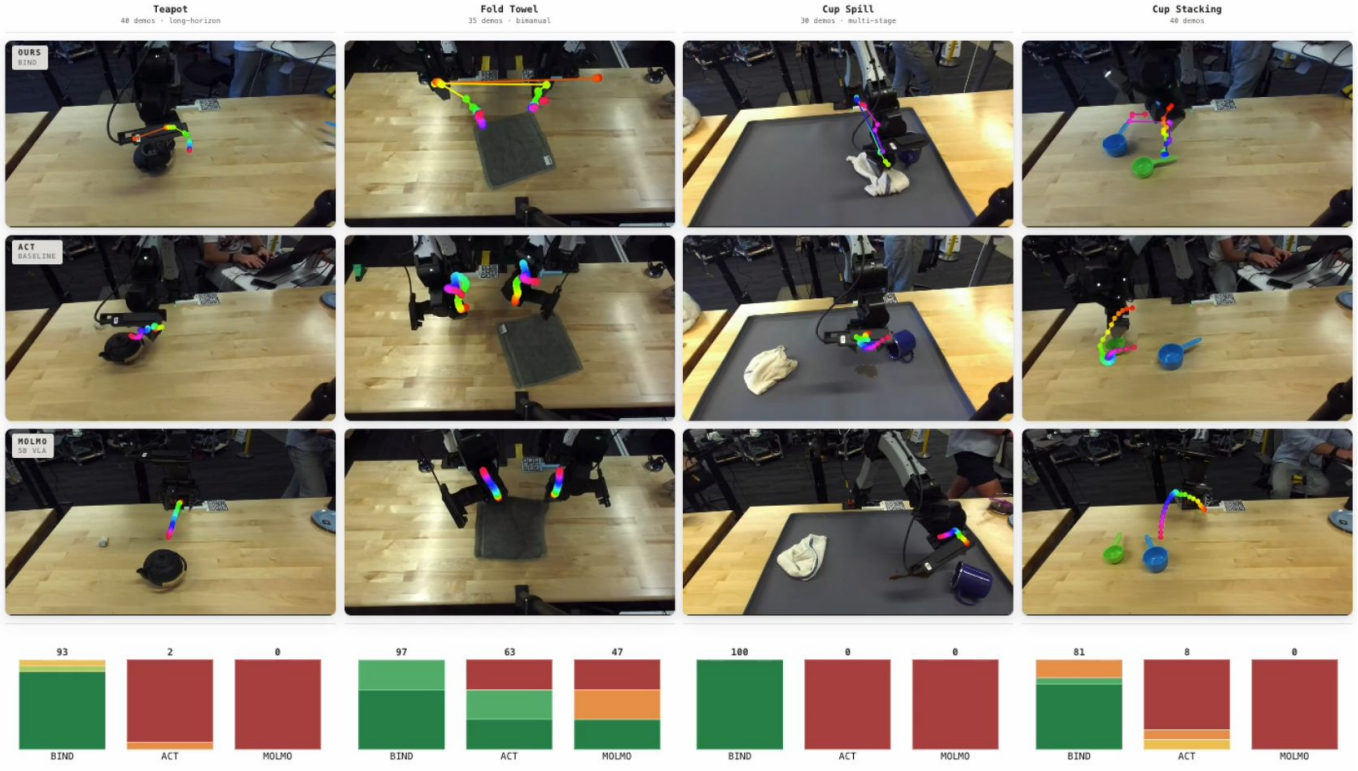
The 3D robot actions → 2D image patches projection is one line of computer vision.

- 'choosing this robot action would move the robot EEF to this image feature'

Step 1: Discretize the action space. Turn regression into classification.

Step 2: Project each candidate onto the image features before scoring.

BIND: Binding 3D Robot Actions to 2D Image Features: Results



BIND: Binding 3D Robot Actions to 2D Image Features: Results

