BUZZ CHRONICLES > ALL
Saved by @ankitsrihbti
See On Twitter

Twitter Thread by Siddharth Karamcheti

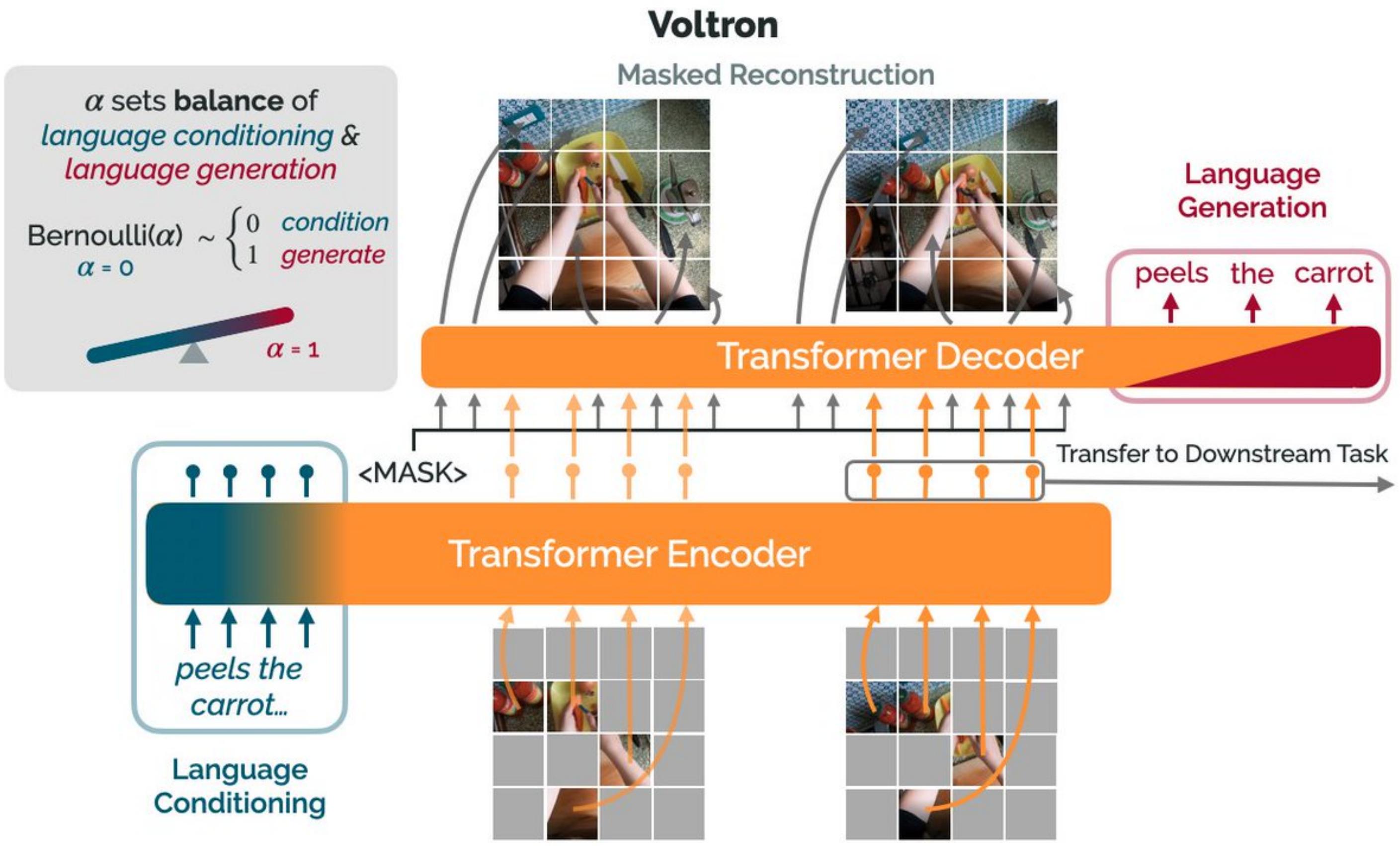
Siddharth Karamcheti
@siddkaramcheti

How can we use language supervision to learn better visual representations for robotics?

Introducing Voltron: Language-Driven Representation Learning for Robotics!

Paper: https://t.co/gIsRPtSjKz
Models: https://t.co/NOB3cpATYG
Evaluation: https://t.co/aOzQu95J8z

■■(1 / 12)



Videos of humans performing everyday tasks (Something-Something-v2, Ego4D) offer a rich and diverse resource for learning representations for robotic manipulation.

Yet, an underused part of these datasets are the rich, natural language annotations accompanying each video. (2/12)

The Voltron framework offers a simple way to use language supervision to shape representation learning, building off of prior work in representations for robotics like MVP (https://t.co/Pb0mk9hb4i) and R3M (https://t.co/o2Fkc3fP0e).

The secret is *balance* (3/12)

Starting with a masked autoencoder over frames from these video clips, make a choice:

1) Condition on language and improve our ability to reconstruct the scene.

2) Generate language given the visual representation and improve our ability to describe what's happening. (4/12)

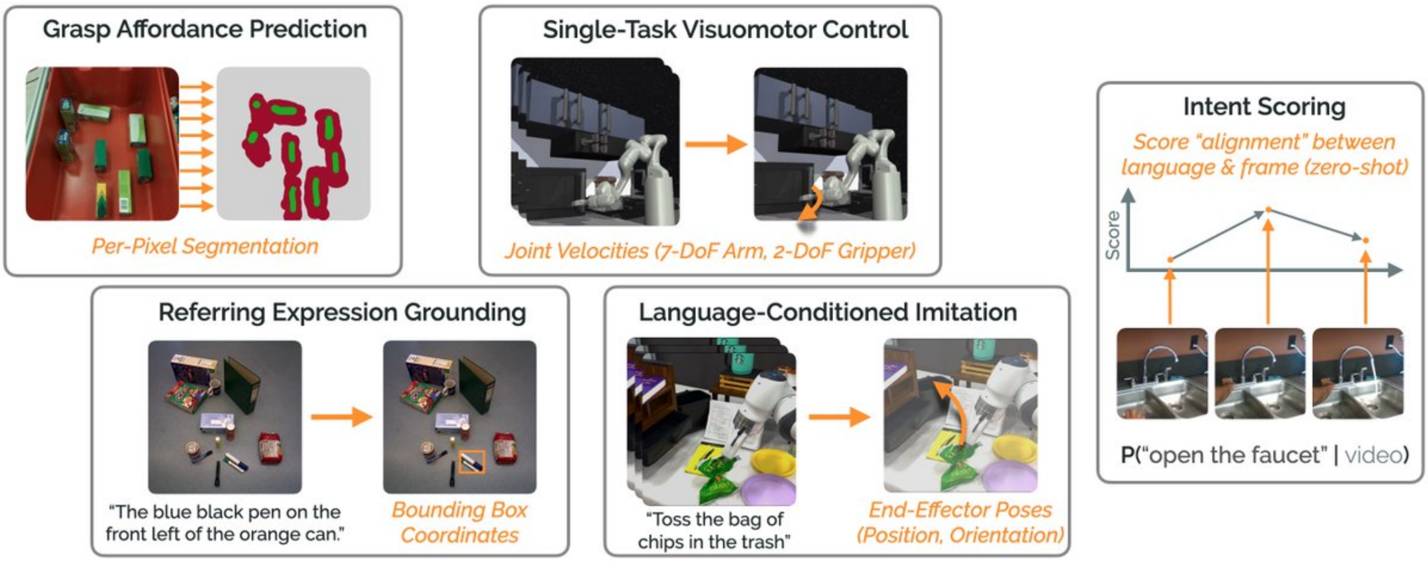
By trading off *conditioning* and *generation* we show that we can learn 1) better representations than prior methods, and 2) explicitly shape the balance of low and high-level features captured.

Why is the ability to shape this balance important? (5/12)

Because robotics isn't a single thing! While prior work focuses on learning for control, there are so many problems we care about – problems that require different features!

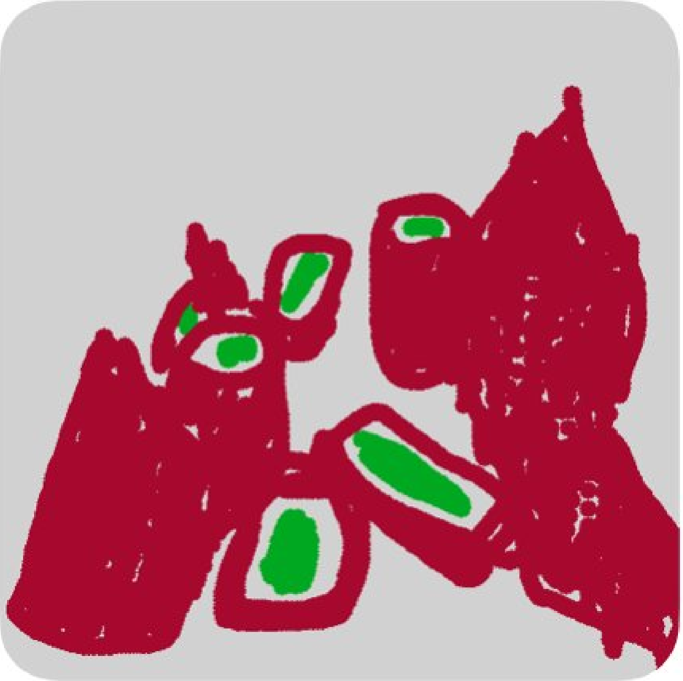
How do we know?

Because we build an evaluation suite of 5 diverse robotics problem domains! (6/12)


Grasp Affordance Prediction
Per-Pixel Segmentation
Single-Task Visuomotor Control
Joint Velocities (7-DoF Arm, 2-DoF Gripper)
Intent Scoring
Score "alignment" between language & frame (zero-shot)
Score
P("open the faucet" | video)
Referring Expression Grounding
"The blue black pen on the front left of the orange can."
Bounding Box Coordinates
Language-Conditioned Imitation
"Toss the bag of chips in the trash"
End-Effector Poses (Position, Orientation)

Problems like grasp affordance prediction (per-pixel segmentation) tend to require more *low-level* spatial features; edges, object boundaries, textures.

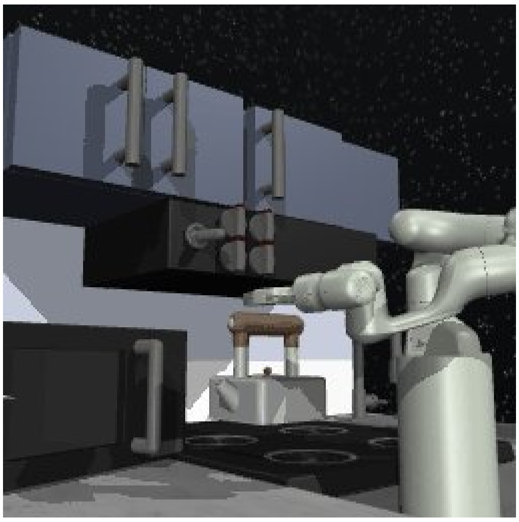
Evaluation: the ARC Grasping dataset (https://t.co/rRI4ya84DL) – CC @andyzengtweets @SongShuran. (7/12)



Learning for control tasks benefit from representations that mix of low and high-level features.

Modeling *multi-frame* contexts (easy with Voltron) is also high-impact!

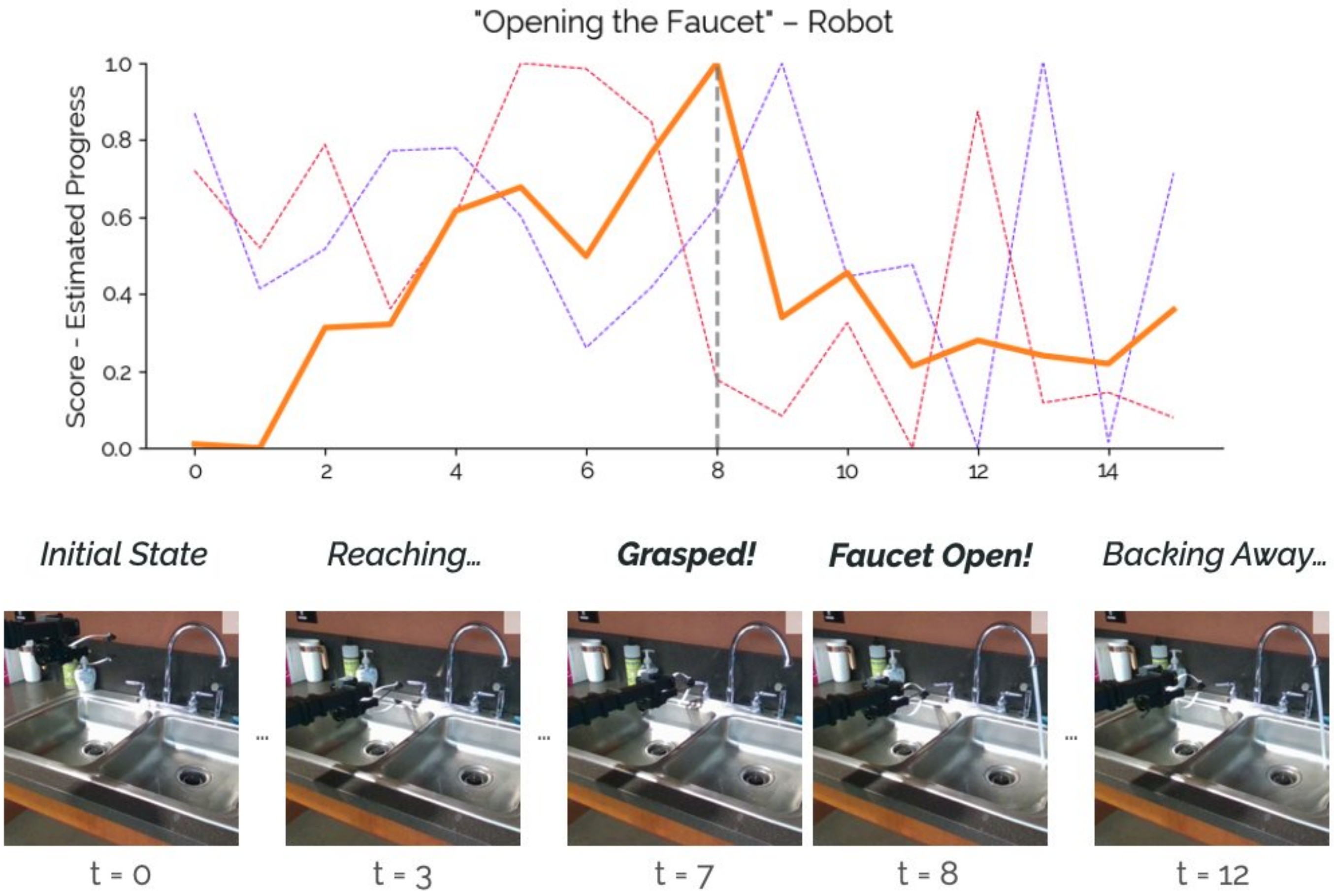
Evaluation: Franka Kitchen & Adroit Manipulation domains from R3M – CC @aravindr93 @Vikashplus. (8/12)



Really cool is how we can use the generative language model zero-shot, with no extra data.

Given a video & language intent, we can score – in real time – how well the behavior in the video captures the intent.

Transfers to *robot data* – no robots during pretraining! (9/12)



But don't take our word for it – try out our representations yourself... or evaluate your own!

Models & Pretraining: https://t.co/NOB3cpATYG
Evaluation Suite: https://t.co/aOzQu95J8z

Use our models: `pip install voltron-robotics` (10/12)

This project was a huge endeavor; one that would not have been possible without amazing collaborators and mentors – @SurajNair_1 @_anniechen_ @tkollar @chelseabfinn @DorsaSadigh and @percyliang.

Further thanks to @ToyotaResearch, @stanfordnlp, and the @StanfordAILab ! (11/12)

I'm really excited to see the impact of language on shaping representations for robotics... but this isn't the end. The hard parts of robotics remain hard.

Voltron is a building block – a tool. I can't wait to see how y'all use it. Thanks folks – and stay tuned ■■! (12/12)