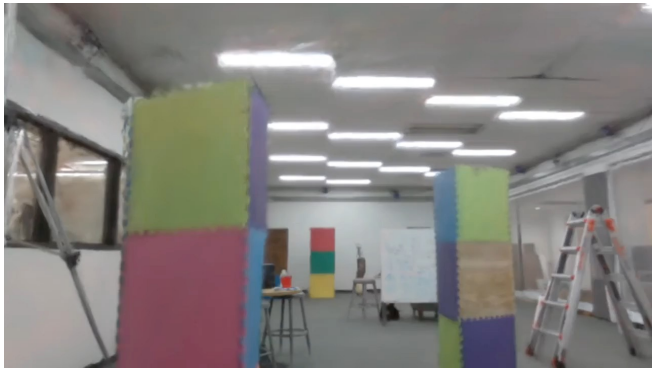
3D World Models for Perception-Rich Robot Autonomy

Mac Schwager
Multi-Robot Systems Lab
Aeronautics and Astronautics Department
Computer Science Department (by courtesy)

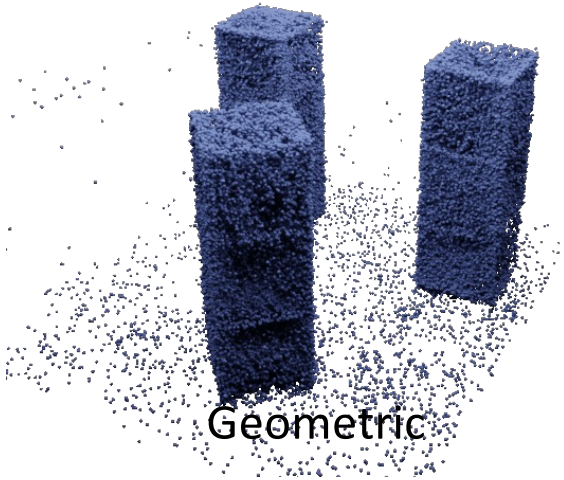
Stanford University

MSL

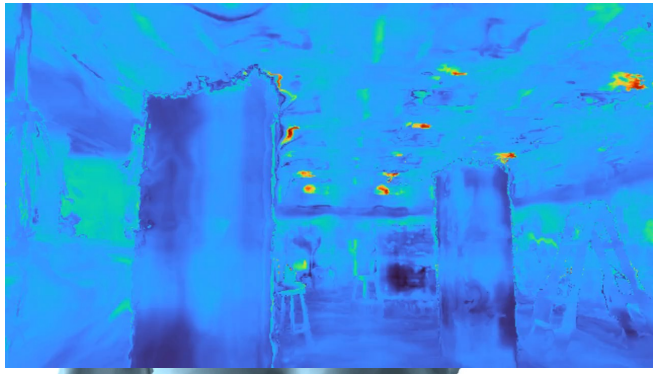
3D World Models that are…



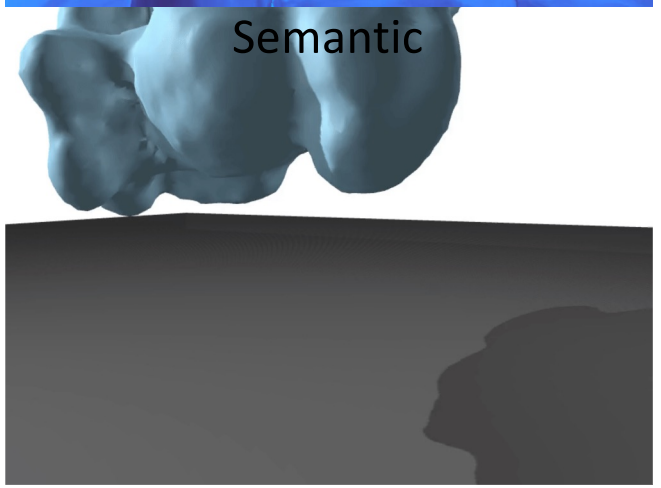
Visual



Geometric

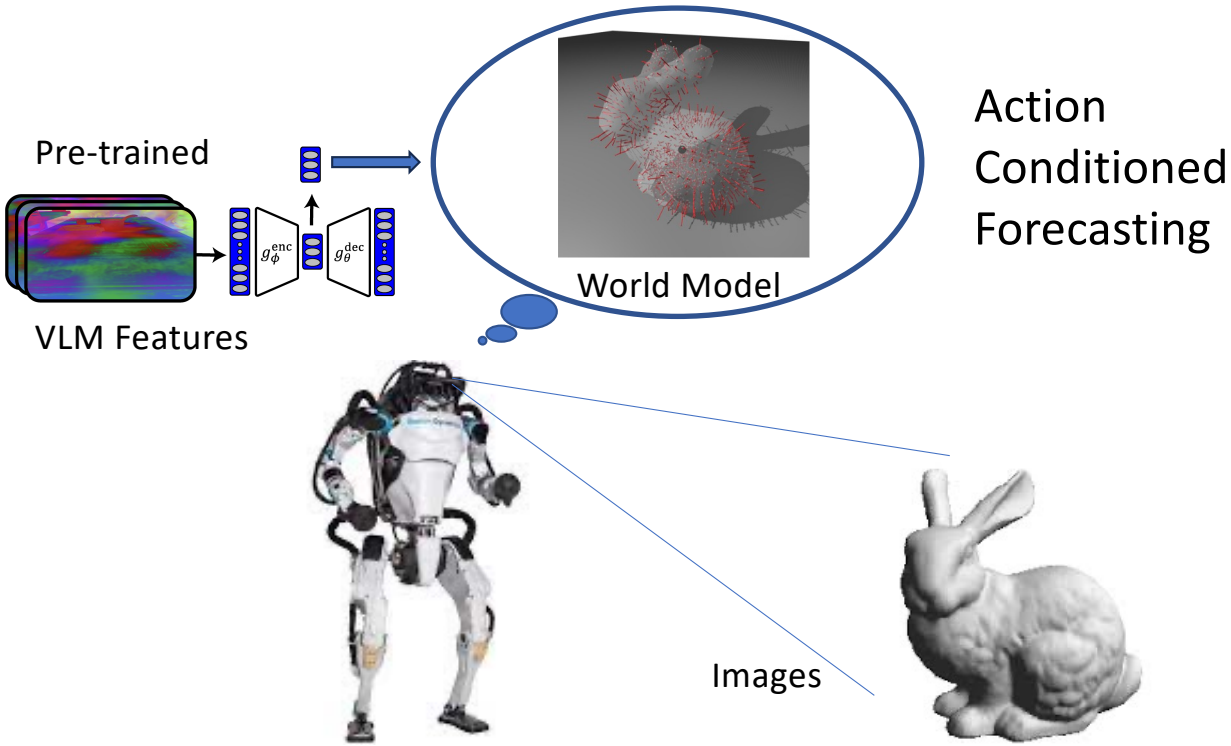


Semantic

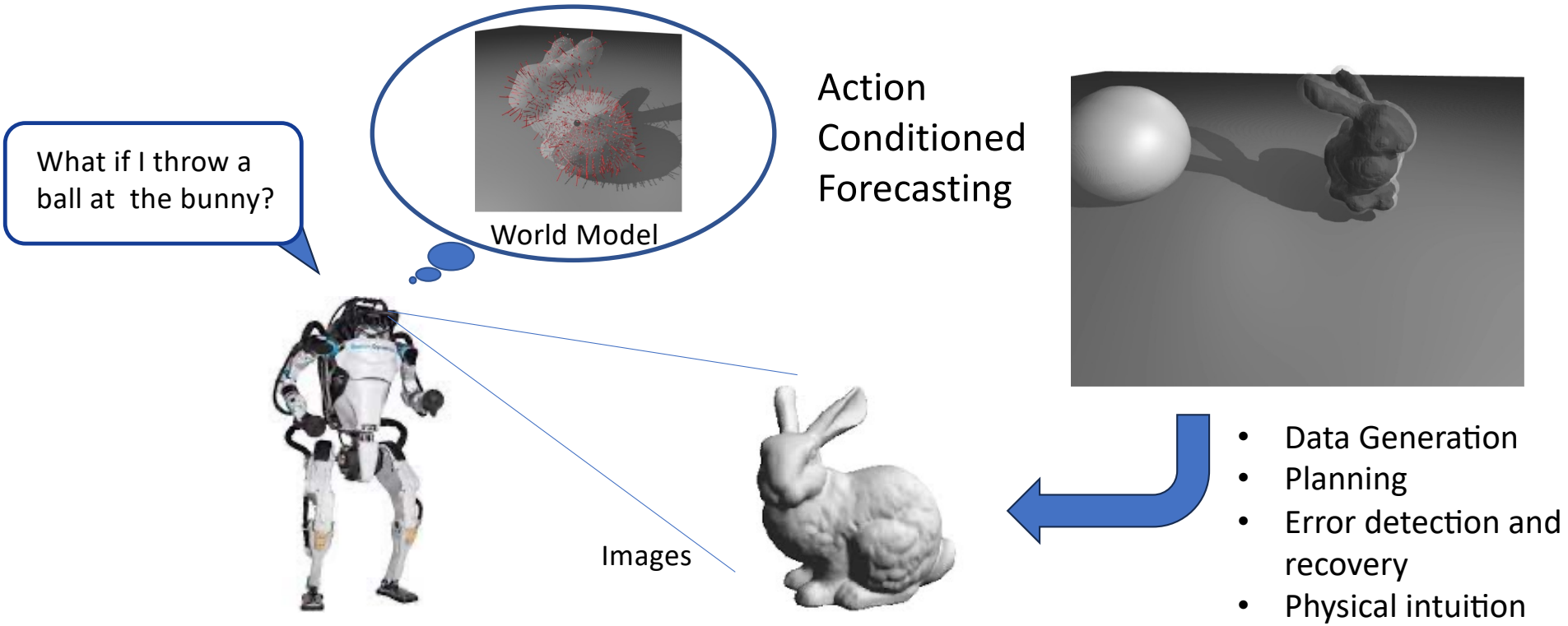


Dynamic

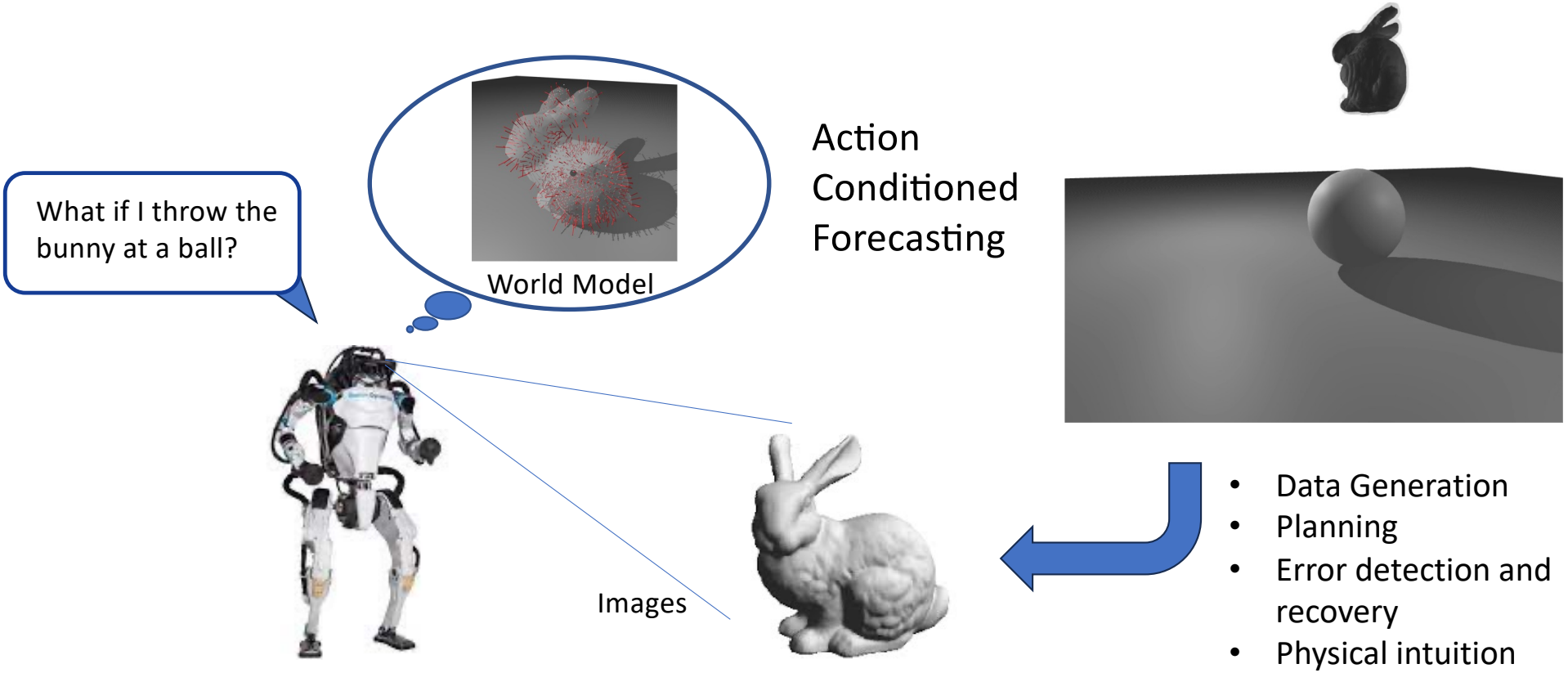
Perception -> world model -> forecasting -> decision making



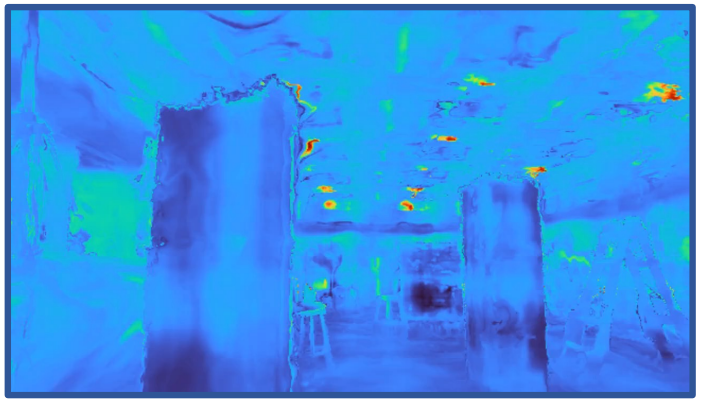
Perception -> world model -> forecasting -> decision making



Perception -> world model -> forecasting -> decision making

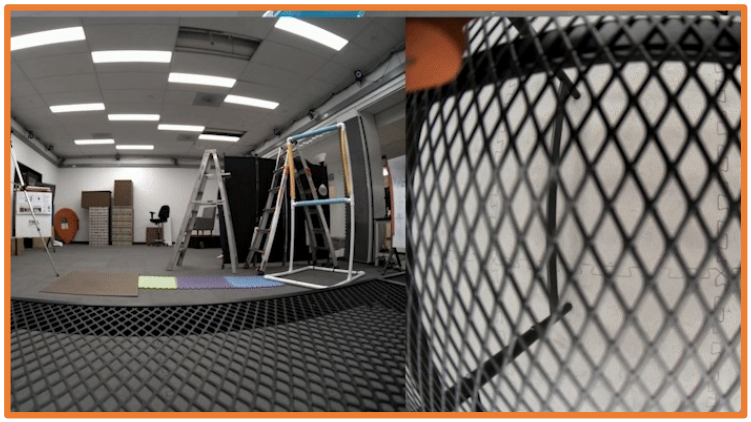


Splat-Nav



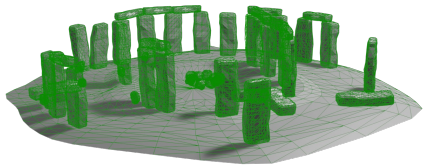
Sous Vide

DroneVLA



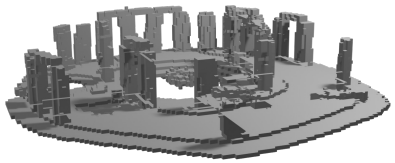
State of the Art

Geometry representations in robotics

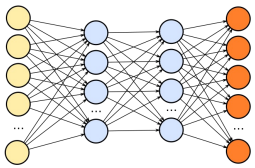


Mesh

Point cloud



Voxel grid

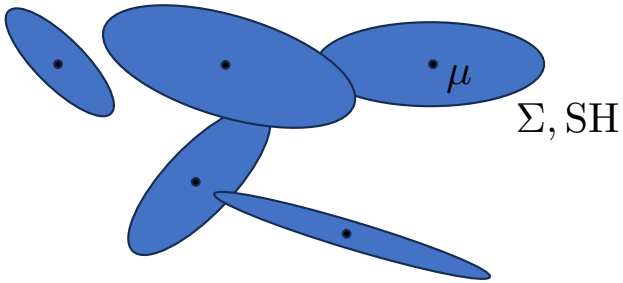


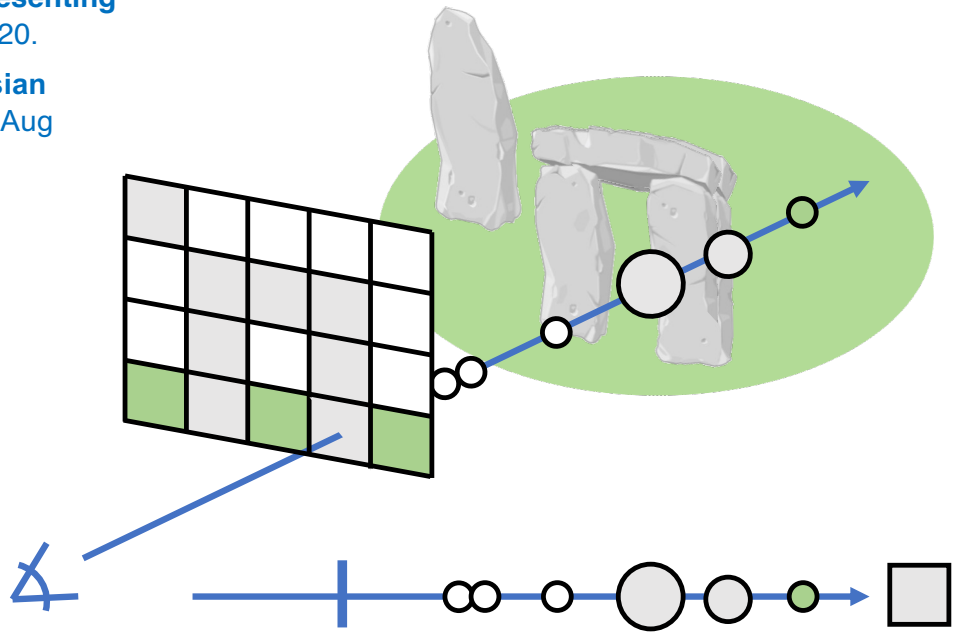
Neural representations

3D Gaussian Splatting (GSplat)

Mildenhall*, Srinivasan*, Tancik*, Barron, Ramamoorthi, Ng. **Representing Scenes as Neural Radiance Fields for View Synthesis**. ECCV '20.

Kerbl, B., G. Kopanas, T. Leimkühler, and G. Drettakis. **3D Gaussian Splatting for Real Tim Radiance Field Rendering**. SIGGRAPH, Aug 2023.



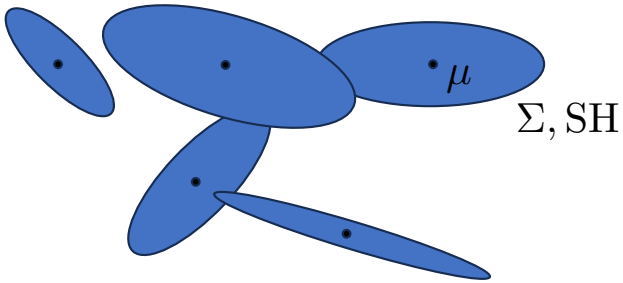


$$(\mathbf{x}, \mathbf{v}) \mapsto \alpha, \mathbf{c}$$

3D point view dir opacity radiance

3D Gaussian Splatting (3DGS)

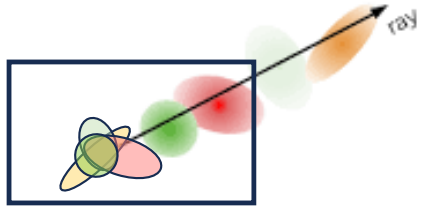
Kerbl, B., G. Kopanas, T. Leimkühler, and G. Drettakis. **3D Gaussian Splatting for Real Tim Radiance Field Rendering**. SIGGRAPH, Aug 2023.



3DGS Rasterized Rendering:

$$(\mathbf{x}, \mathbf{v}) \mapsto \alpha, \mathbf{c}$$

3D point view dir opacity radiance

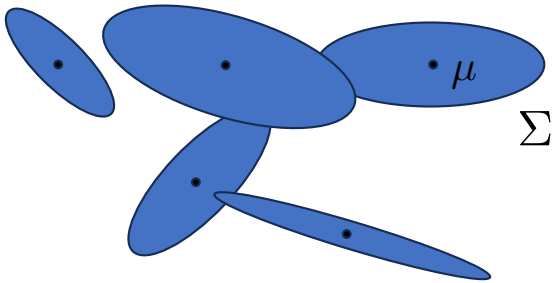


Language Embedded NerF/GSplats

Kobayashi, Matsumoto, and Sitzmann. **Decomposing nerf for editing via feature field distillation**. NeurIPS 2022.

Kerr, Kim, Goldberg, Kanazawa, and Tancik, **LERF: Language embedded radiance fields**, ICCV 2023.

Qin, Li, Zhou, Wang, Pfister, **Lang-Splat: 3D Language Gaussian Splatting**, CVPR 2024.

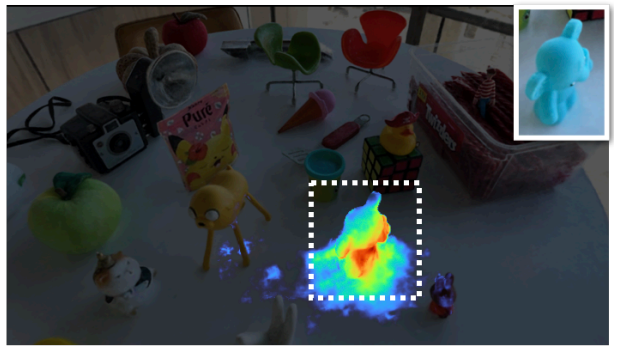


$\Sigma, \mathrm{SH}, \mathbf{f}_{\mathrm{sem}}$

$$(\mathbf{x}, \mathbf{v}) \mapsto \alpha, \mathbf{c}, \mathbf{f}_{\mathrm{sem}}$$

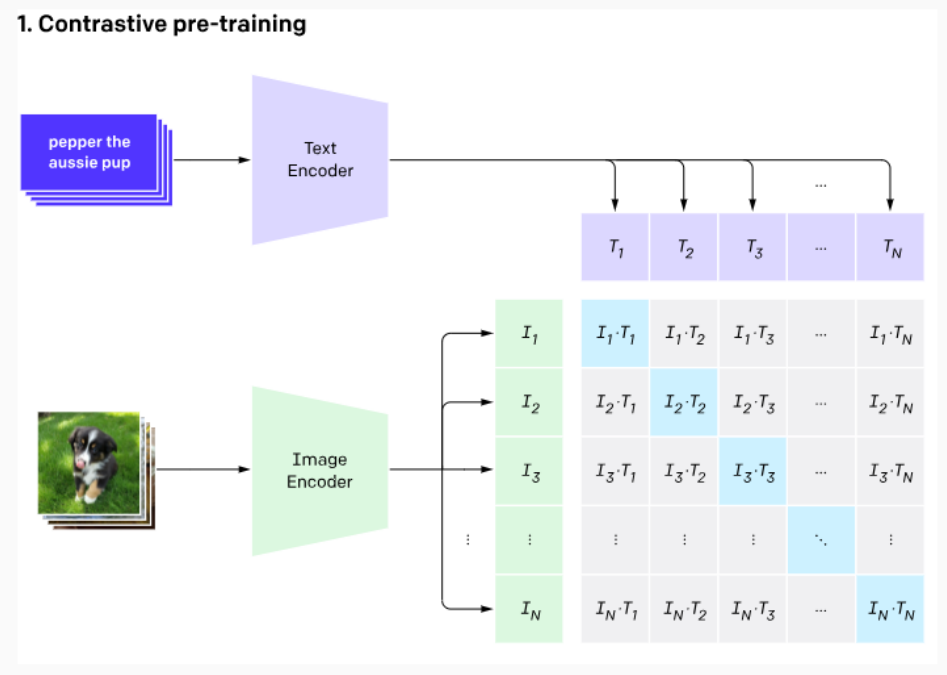
3D point view dir opacticy radiance semantic code (e.g. CLIP)

3D scene: Figurines



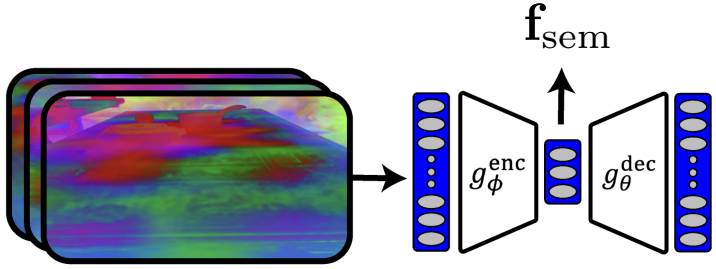
"toy elephant"

CLIP: Contrastive Language-Image Pretraining



Radford et al, **CLIP**, ICML 2021.

Firoozi et al, **Foundation Models in Robotics**, IJRR 2024.



Dimension reduction and 3D distillation

Language embedded GSplat: “Microwave”

Language embedded GSplat: “Penguin”

NeRFs/3DGS in Robotics

SLAM

Sucar et al, **iMAP**, ICCV 2021.
Zhu et al, **NICE-SLAM,** CVPR 2022.
Yu et al, **NeRFBridge**, ICRA Workshop, 2023.
Rosinol et al, **NeRF-SLAM**, IROS 2023.
Matsuki et al, **GSplat-SLAM**, CVPR 2024.
Yan et al, **GS-SLAM**, CVPR 2024.
Keetha et al, **Splatam**, CVPR 2024.

iMAP

Navigation, planning, control

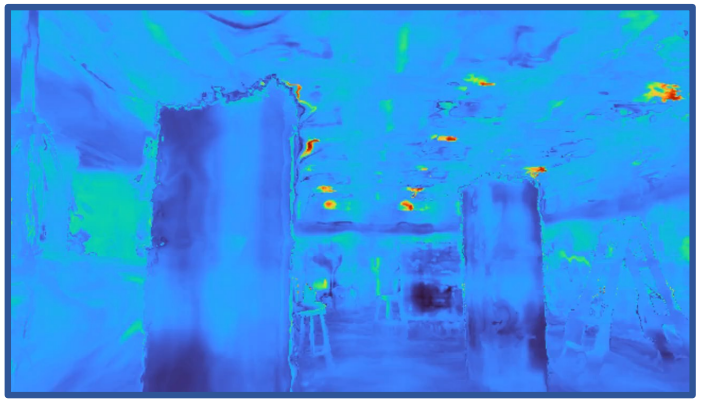
Adamkiewicz et al, **NeRF-Nav**, RA-L 2022.
Tong et al, **NeRF-CBF**, ICRA 2023.
Tao et al, **RT-Guide**, arXiv 2024.
Ong et al, **Atlas Navigator,** arXiv 2025.

NeRF-Nav

Manipulation

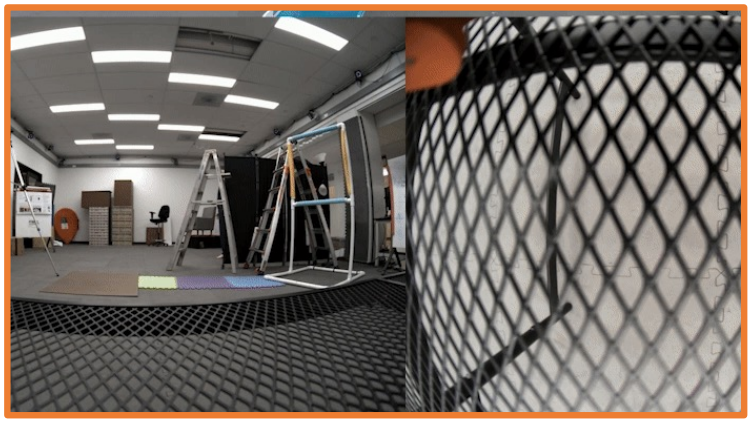
Shen et al, **F3RM**, CoRL 2023. (Best conf paper)
Rashid et al, **LERF-TOGO**, CoRL 2023. (Best conf paper finalist)
Tucker et al, **Splat-MOVER**, CoRL 2024.
Ji et al, **GraspSplats**, CoRL 2024.
Zheng et al, **GaussianGrasper**, CoRL 2024.
Michaux et al, **Let's Make a Splan**, T-RO, 2024.

Splat-Nav

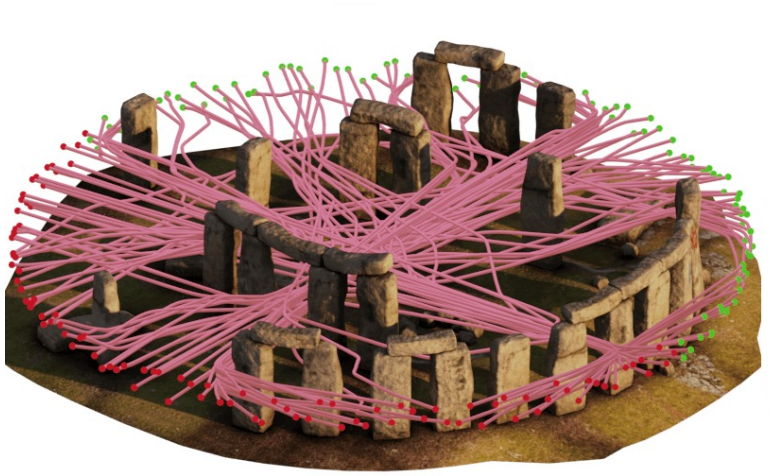


Sous Vide

DroneVLA



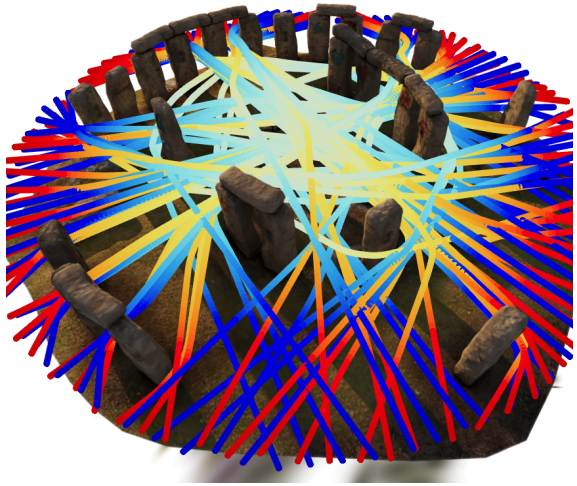
Splat-Nav



Safe planning in a NeRF:
NeRF-Nav, CATNIPS

Adamkiewicz et al, **NeRF-Nav**, RA-L 2022.

Chen et al, **CATNIPS**, T-RO, 2024.

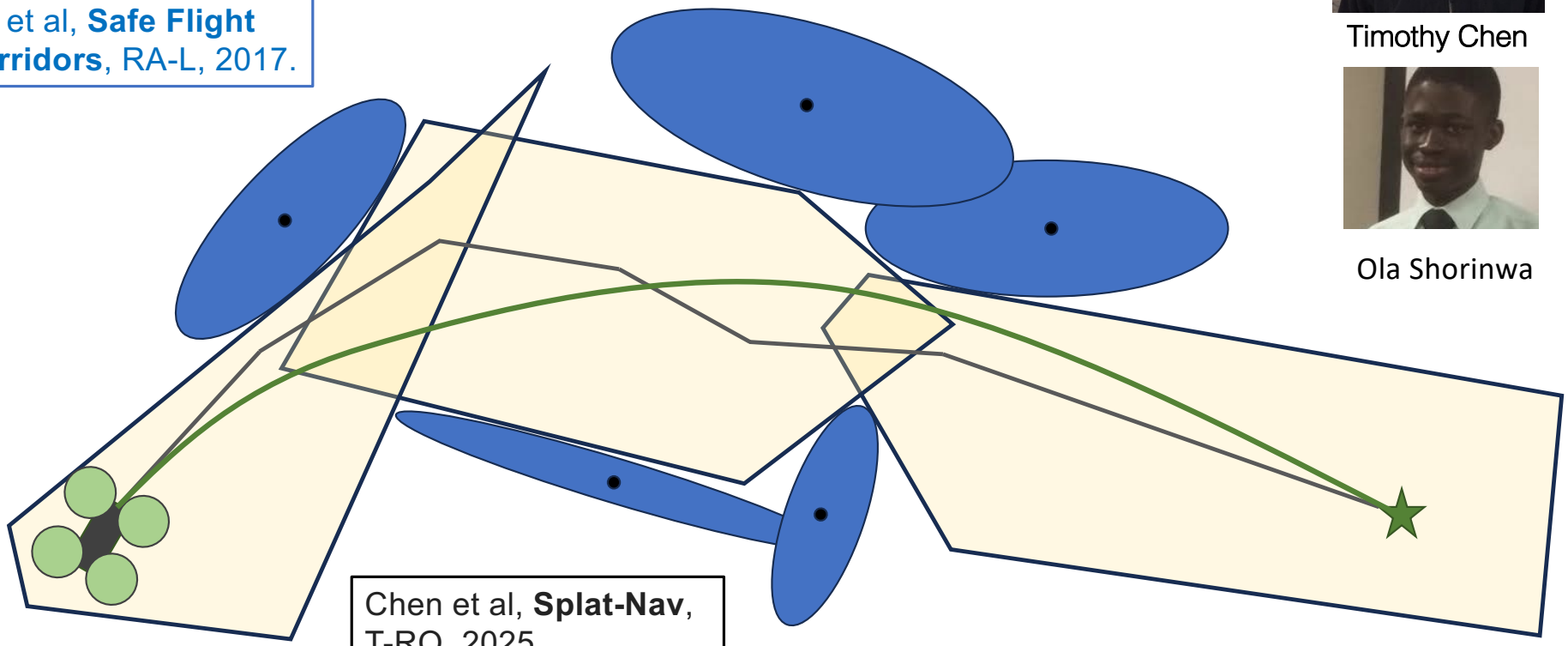


Safe planning in a 3DGS:
Splat-Nav

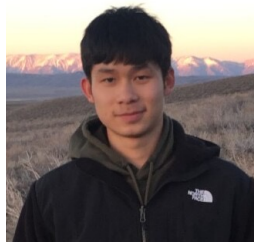
Chen et al, **Splat-Nav**, T-RO, 2025.

Safe trajectory planning in a 3DGS

Liu et al, **Safe Flight Corridors**, RA-L, 2017.



Chen et al, **Splat-Nav**, T-RO, 2025.



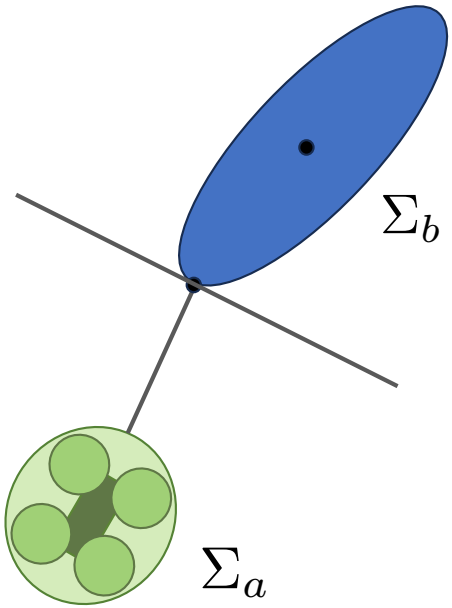
Timothy Chen

Ola Shorinwa

Fast convex ellipse-ellipse nearest point

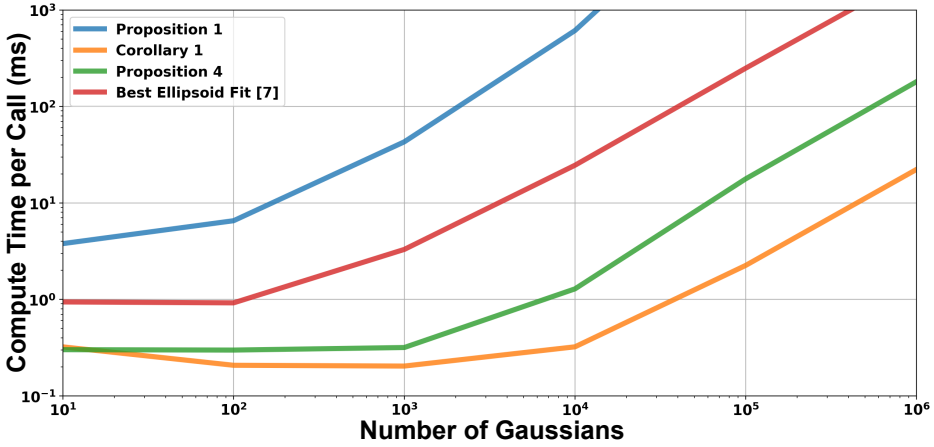
Gilitschenski and Hanebeck, Int. Conf. Info. Fusion, 2012.

Concave: $$\max_{s\in[0,1]} K(s) = (\mu_b - \mu_a)^T \left[\frac{1}{1-s}\Sigma_a + \frac{1}{s}\Sigma_b\right]^{-1} (\mu_b - \mu_a)$$

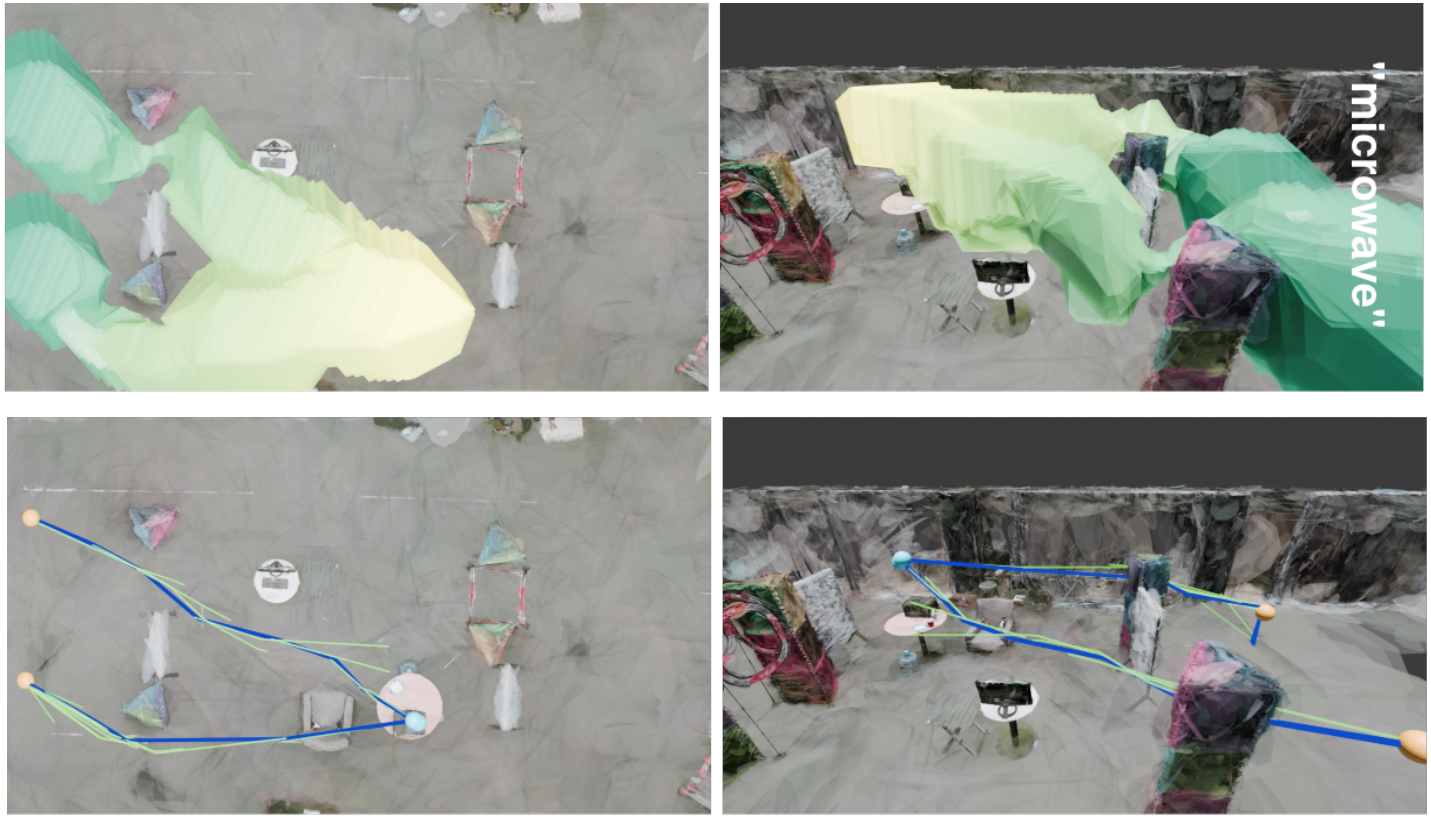


Generalized eigenvalue prob

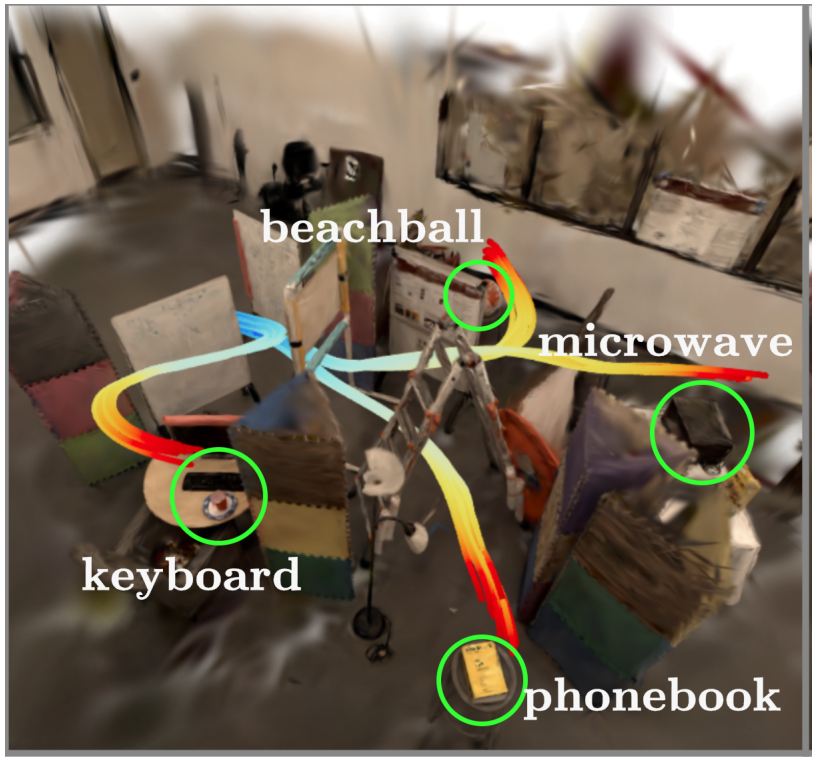
$$\max_{s\in[0,1]} K(s) = v^T \mathbf{diag}\left(\frac{s(1-s)}{1+s(\lambda_i - 1)}\right) v$$

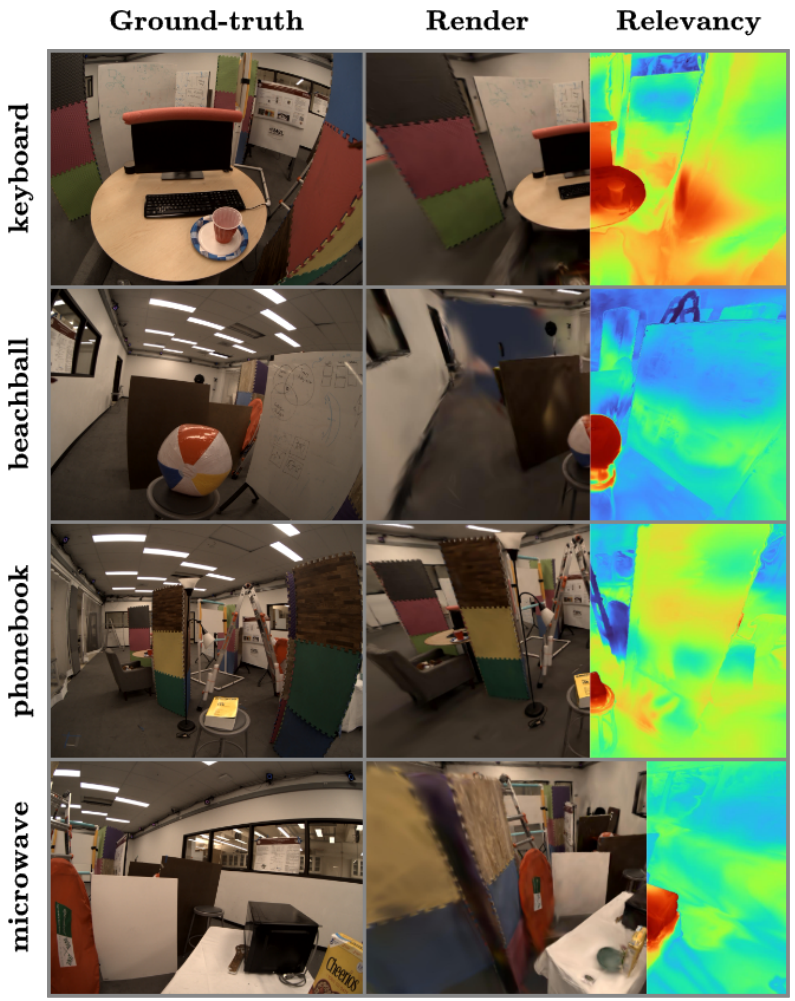


Splat-Nav Corridors and Trajectories

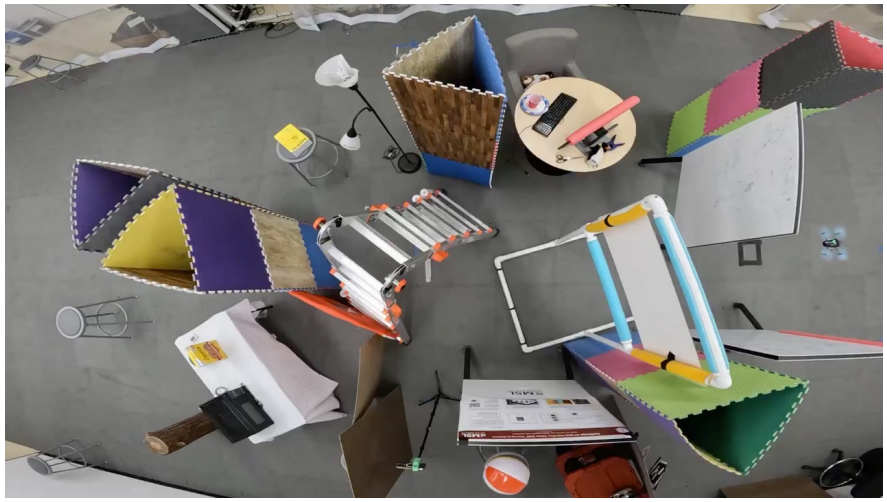


Semantic Splat-Nav

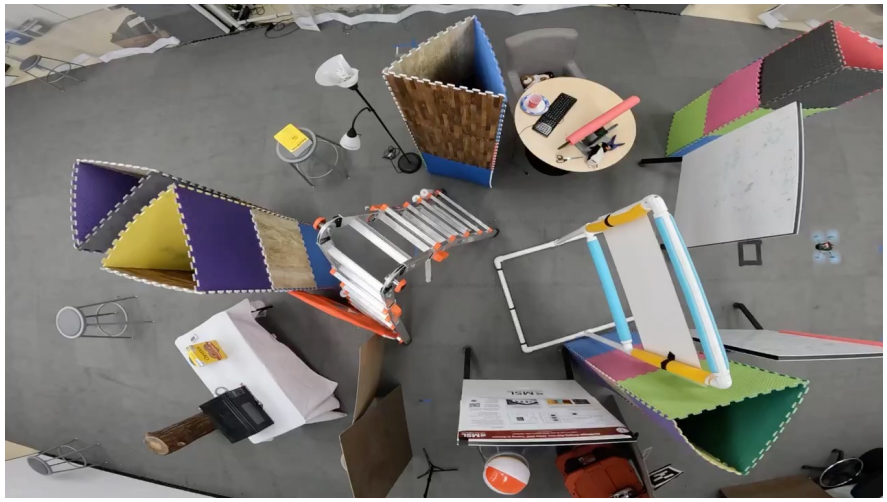




Splat-Nav Experiments

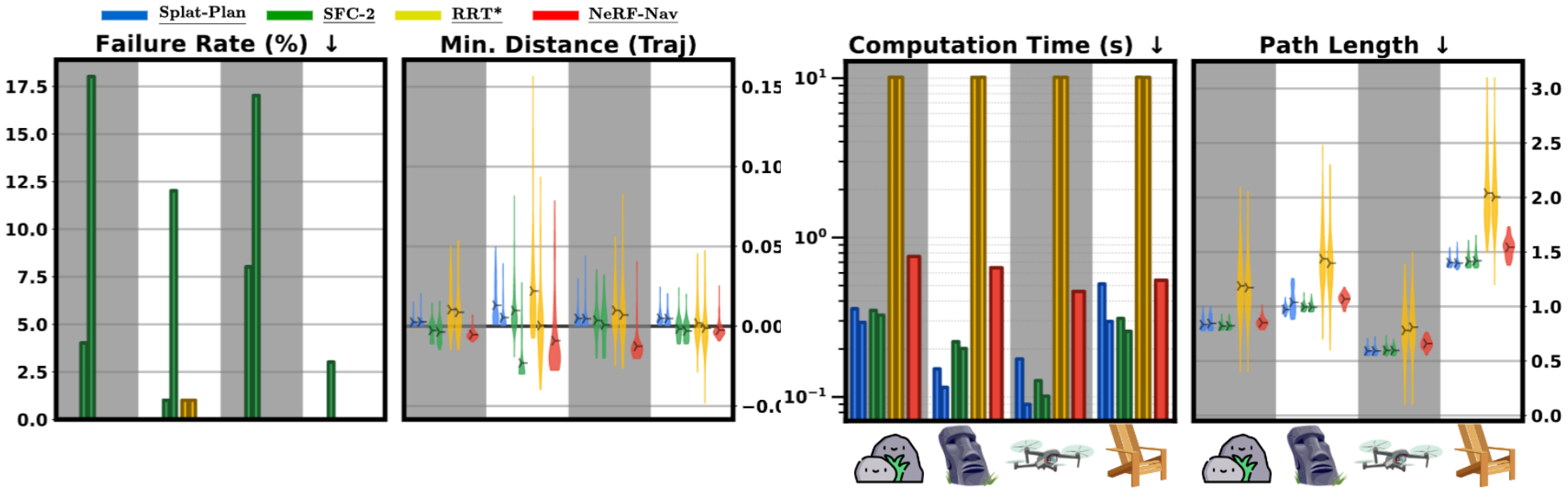


“phonebook”

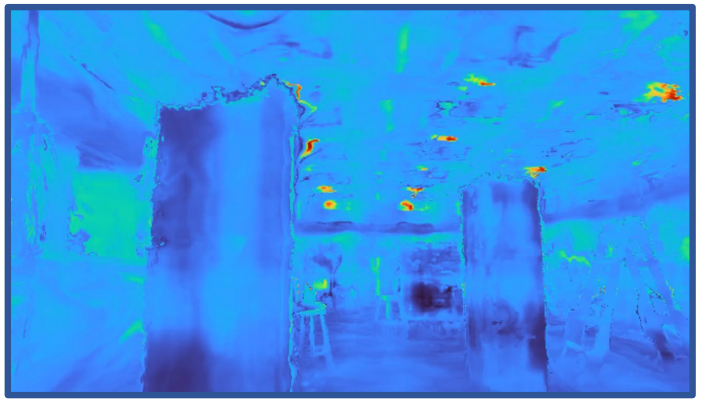


“microwave”

Performance comparison

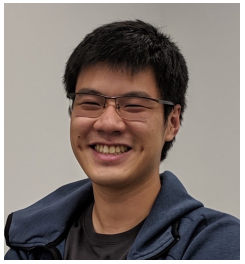


Splat-Nav

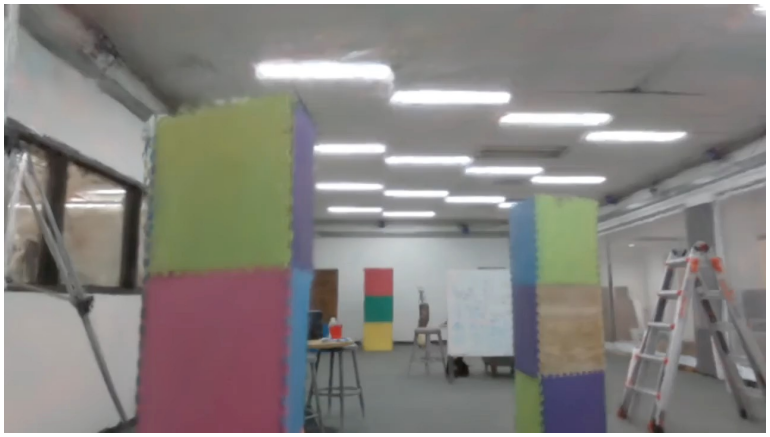


Sous Vide

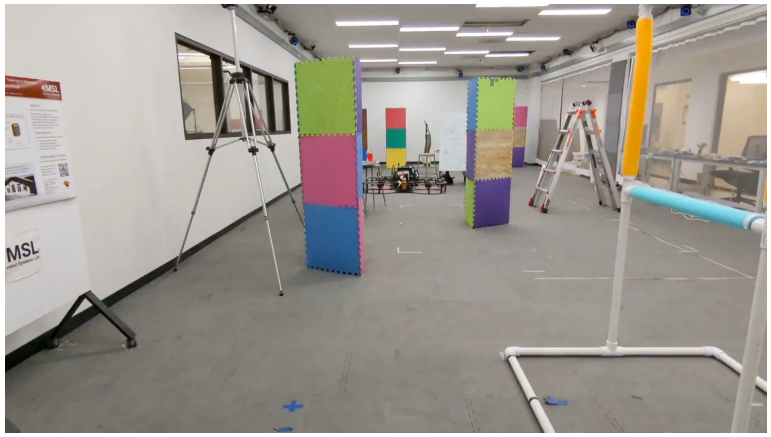
Sous Vide



JunEn Low



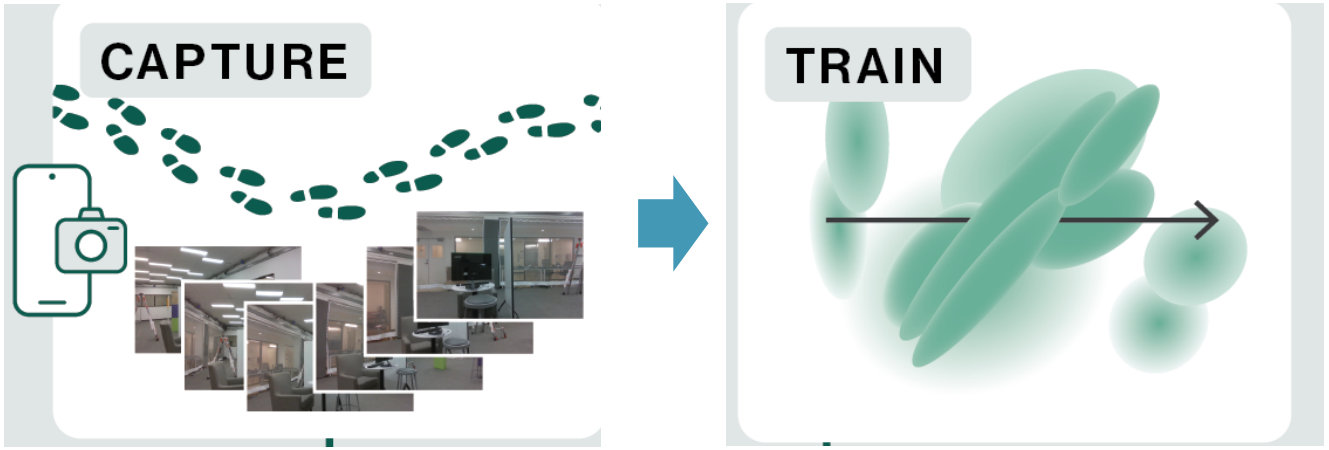
Train Policy in 3DGS World Model

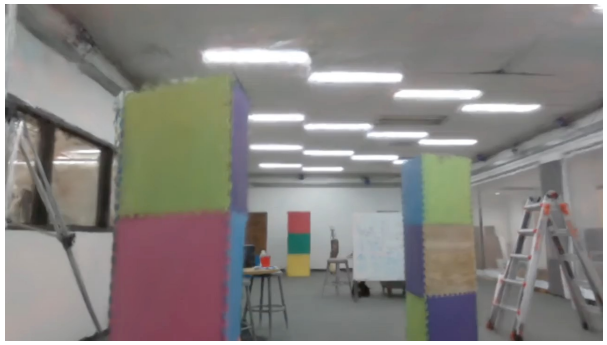


Zero-Shot Sim-to-Real Transfer

Low et al, **Sous Vide**, RA-L 2025.

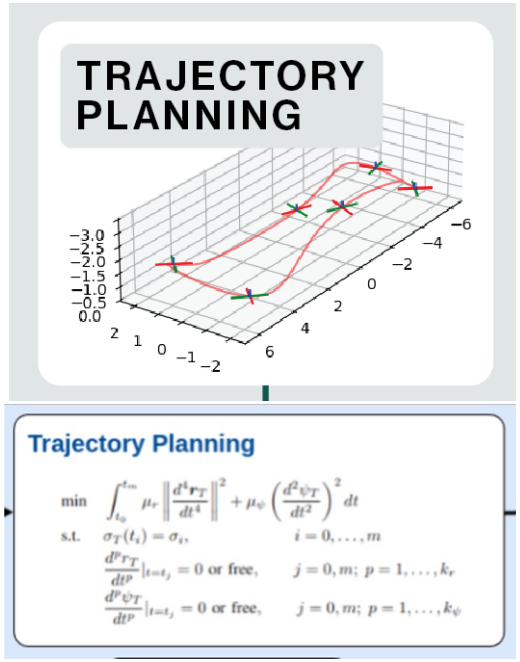
Flying in Gaussian Splats (FiGS)

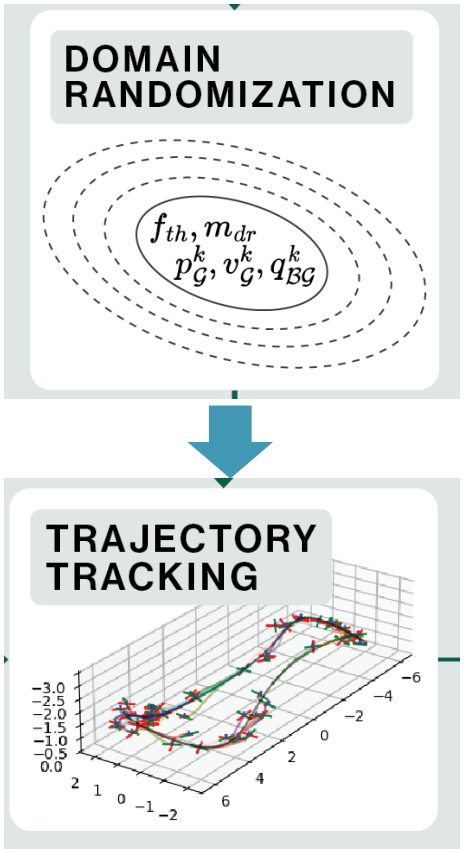




FPV Visual Drone Simulator

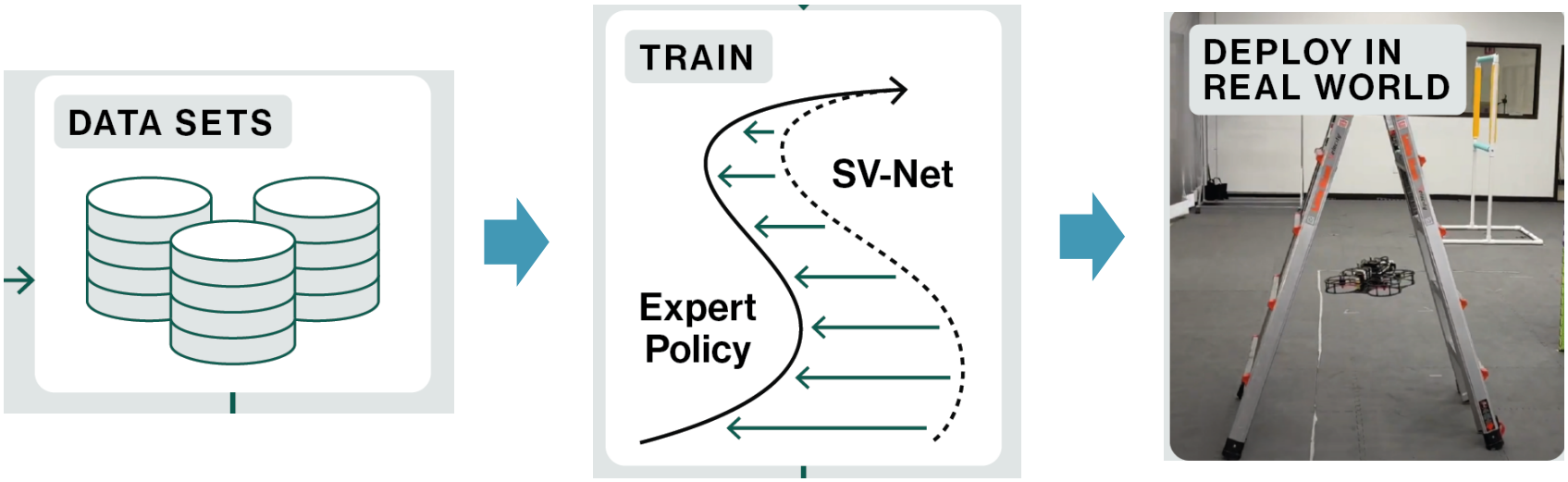
Data Generation from Nonlinear MPC Expert



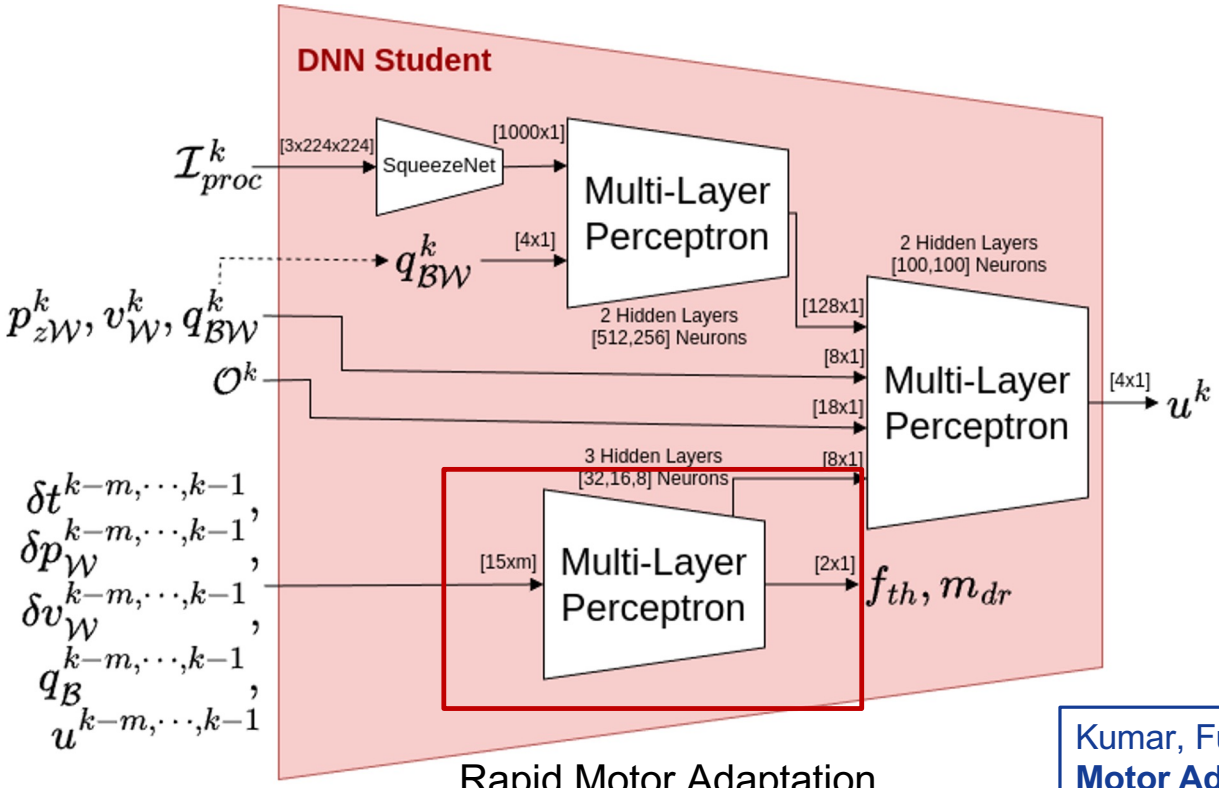


- Expert trajectories with trajectory optimization
- Augmentation data to recover from perturbations
- Randomize drone dynamics parameters +/- 30%
- 300k Image-action pairs

Teacher-Student Policy Training



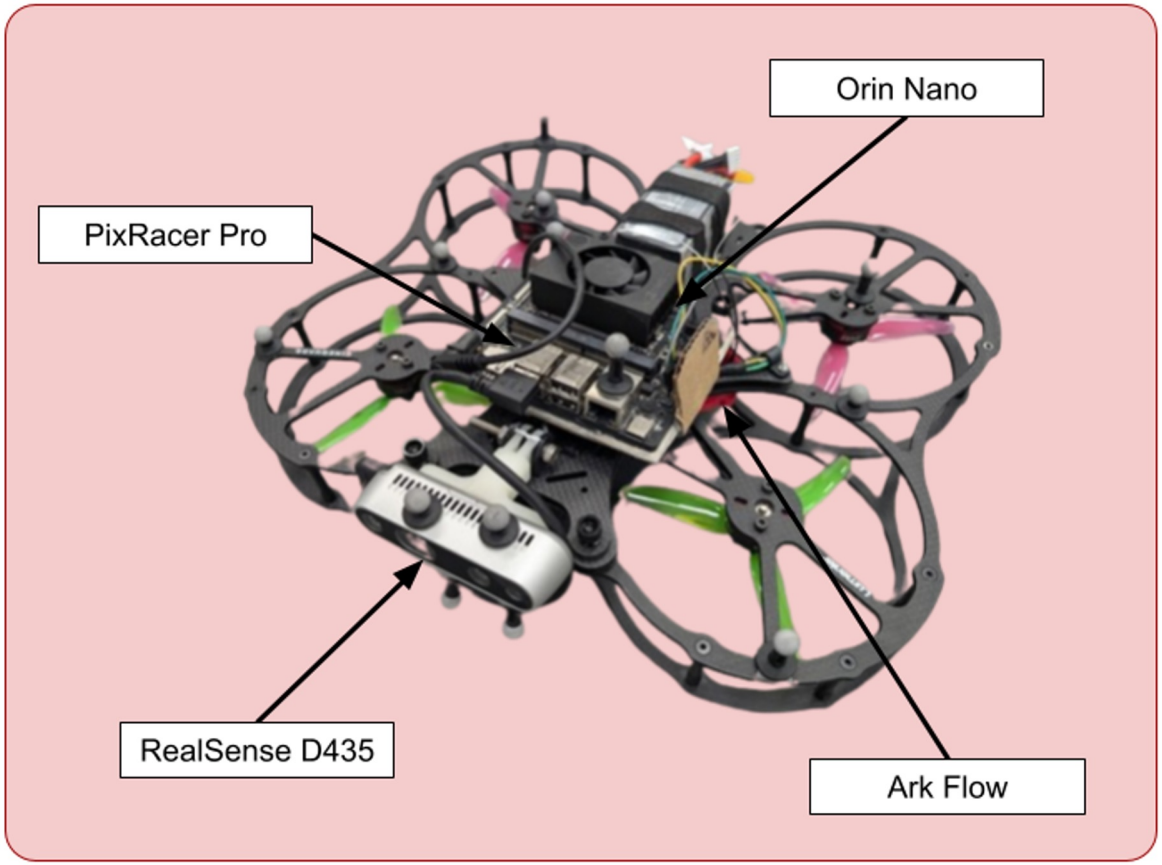
SV-Net Policy Architecture



Rapid Motor Adaptation

Kumar, Fu, Pathak, Malik, **Rapid Motor Adaptation**, RSS 2021.

Drone Hardware

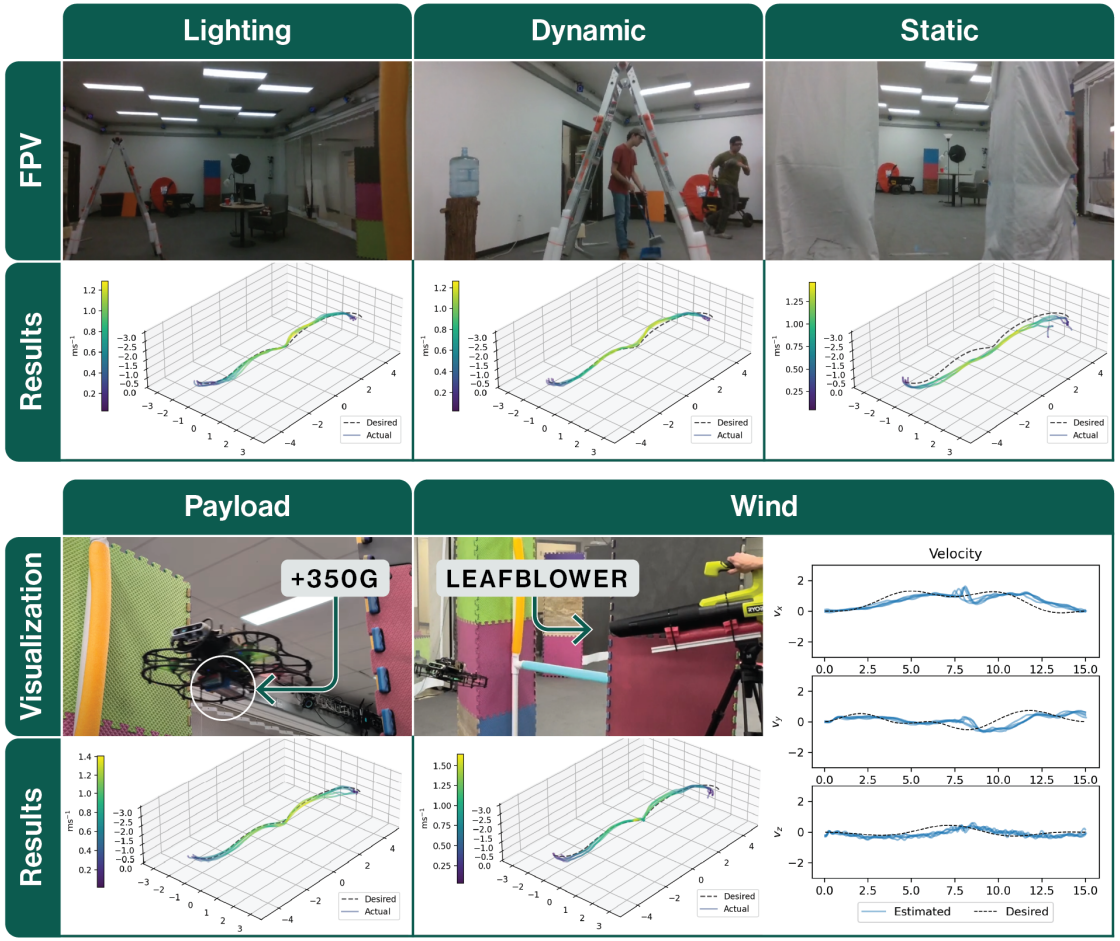


Runs at 30Hz Orin Nano onboard drone

Thrust and body rate commands for greater agility.

Zero-Shot Sim-to-Real Transfer

Robustness in Hardware Flights

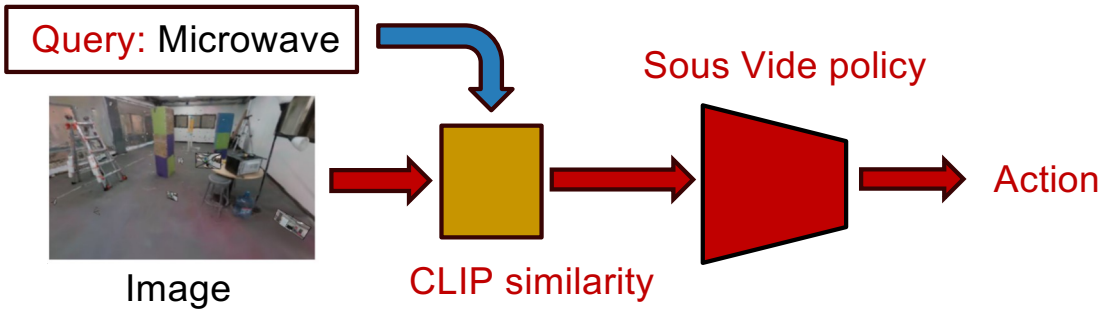


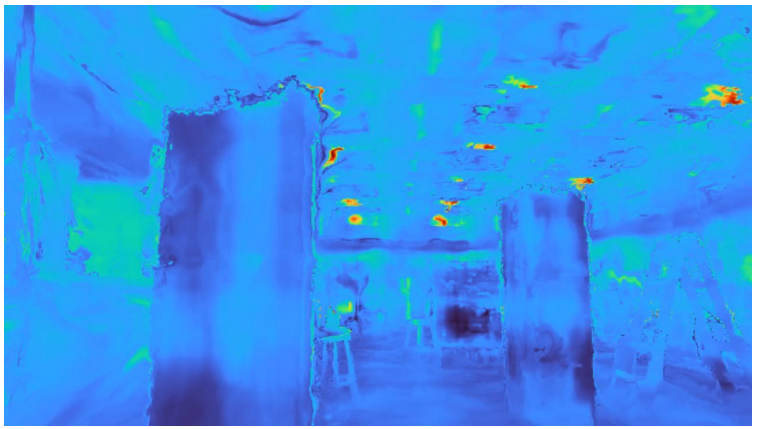
Robust to:

- 40 m/s wind gust from leaf blower
- +30% mass
- Lights dimmed by over 50%
- Furniture moved
- Dynamic distractors

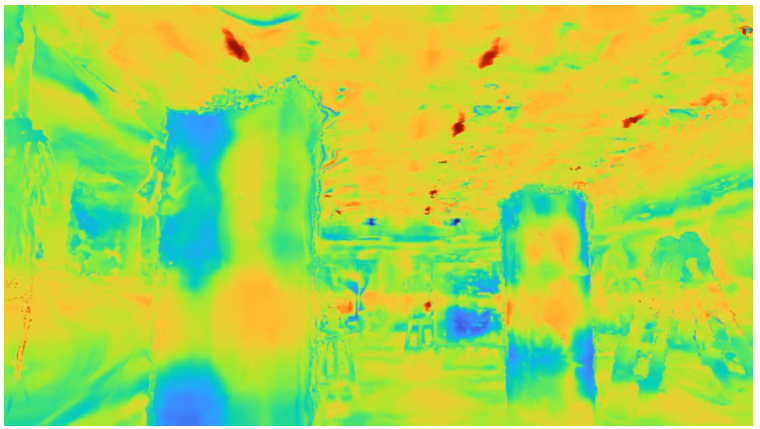
Language Commanded Sous Vide

Max Adang



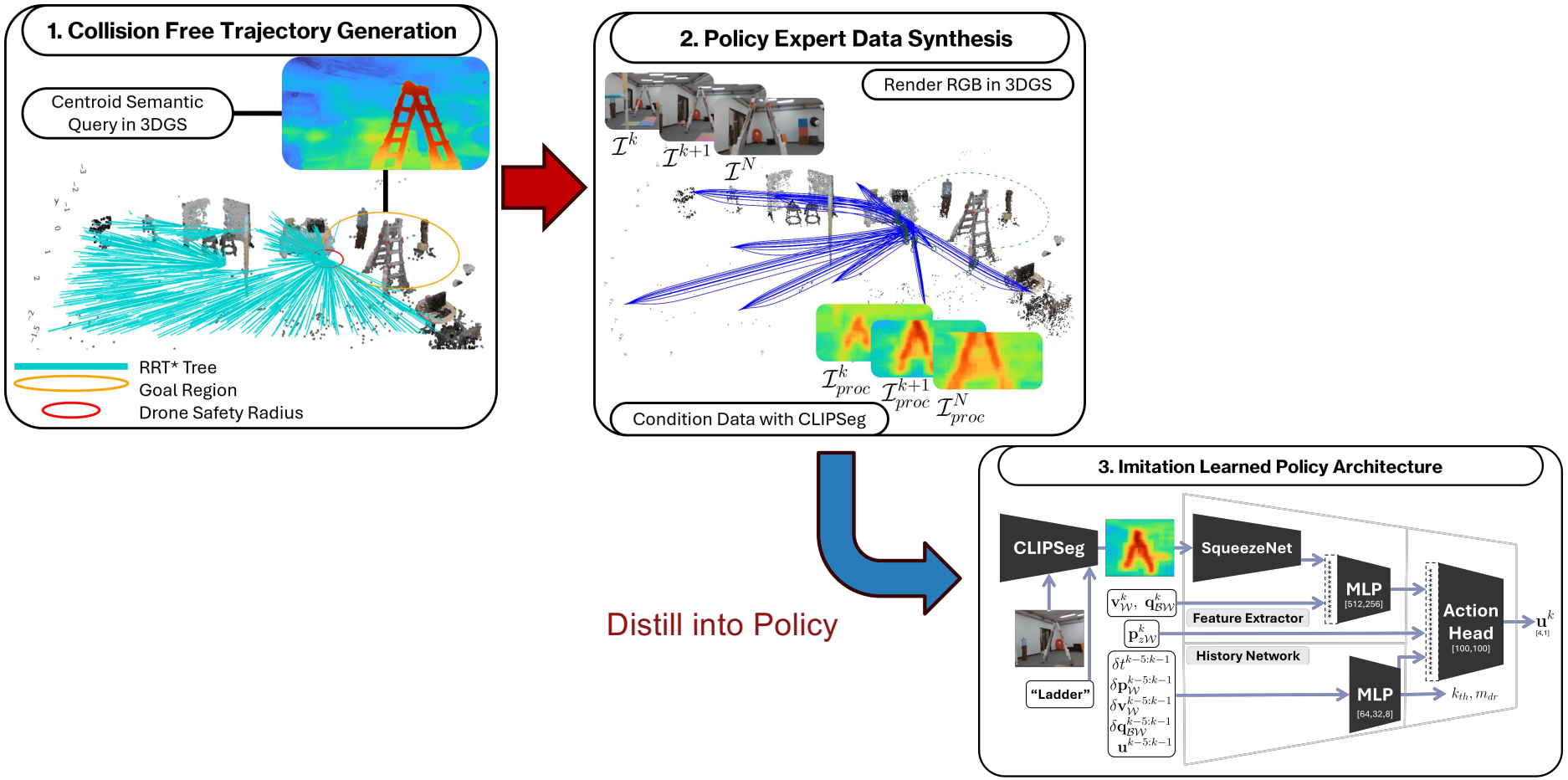


Query: Microwave



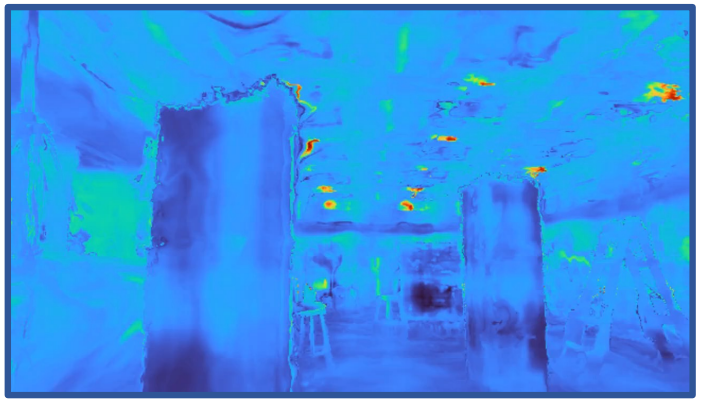
Query: Computer

Object and Scene Generalization



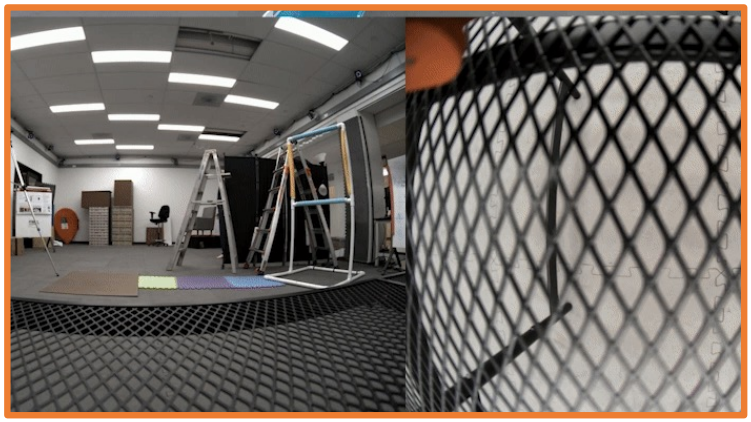
Language Commanded Sous Vide

Splat-Nav

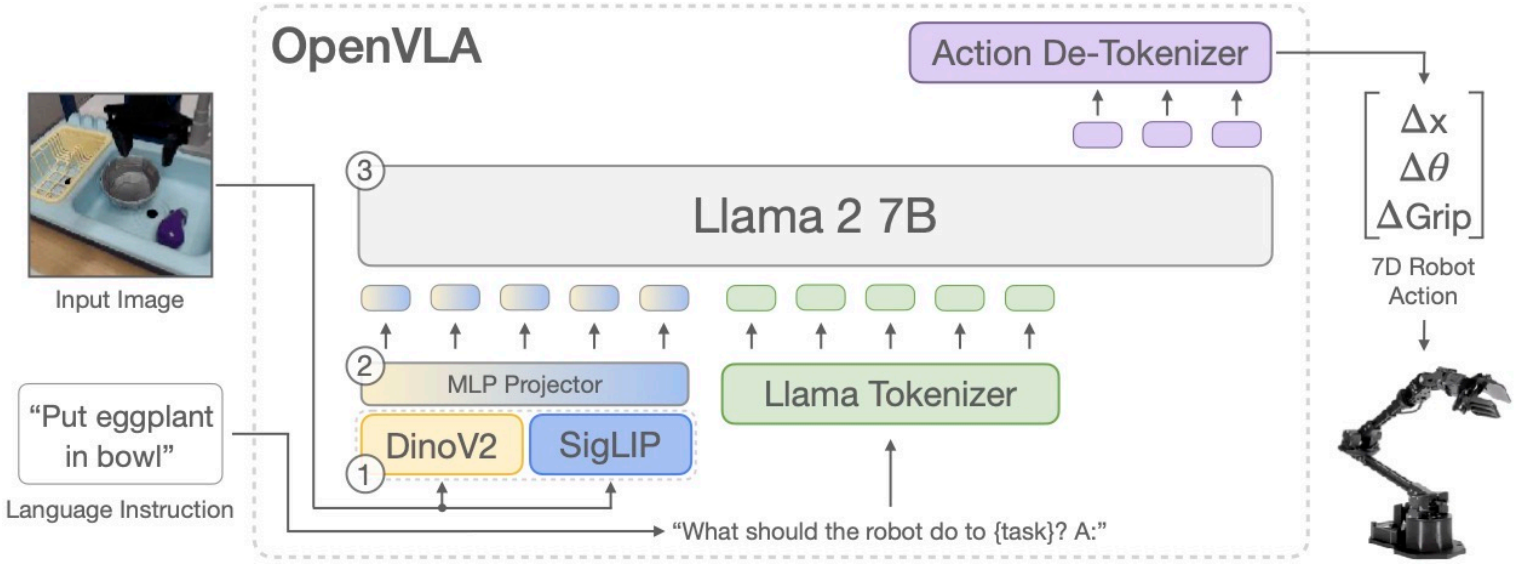


Sous Vide

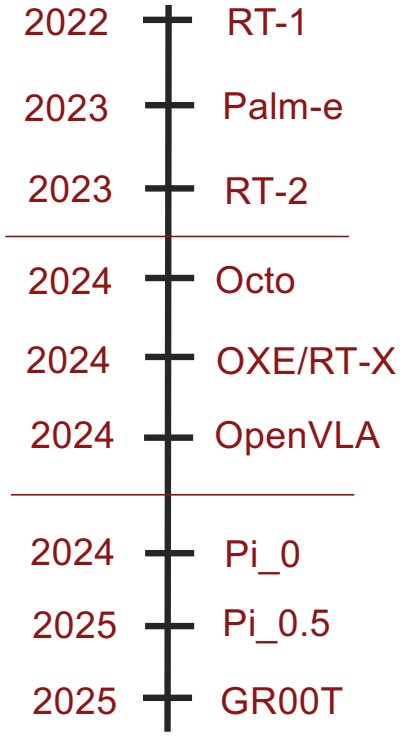
DroneVLA



The Vision Language Action (VLA) Paradigm



From OpenVLA project page: https://openvla.github.io/

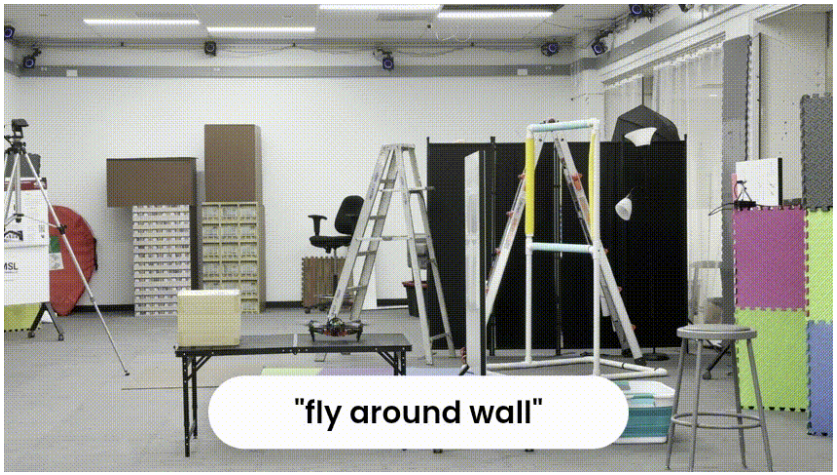


Drone VLA UMI Gripper

Drone VLA Camera Feeds

Human-Piloted Manipulation Demonstrations

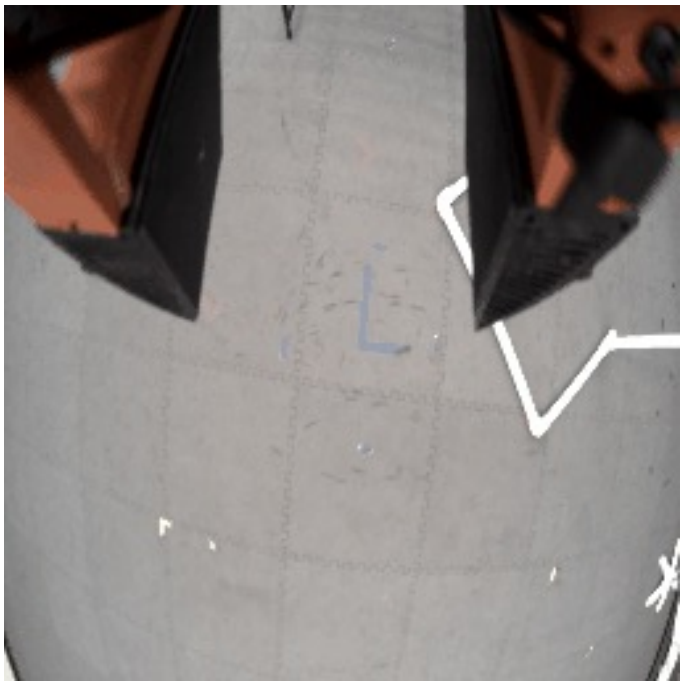
Human-Piloted Demos

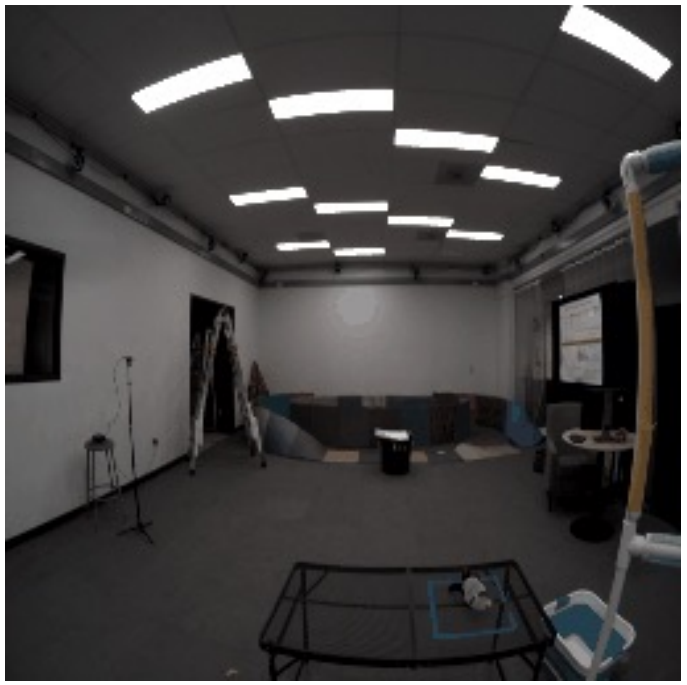


- 500 total human demos
- Penguin Grasp: 120
- Chip Grasp: 120
- Gate navigation in Figs: 100

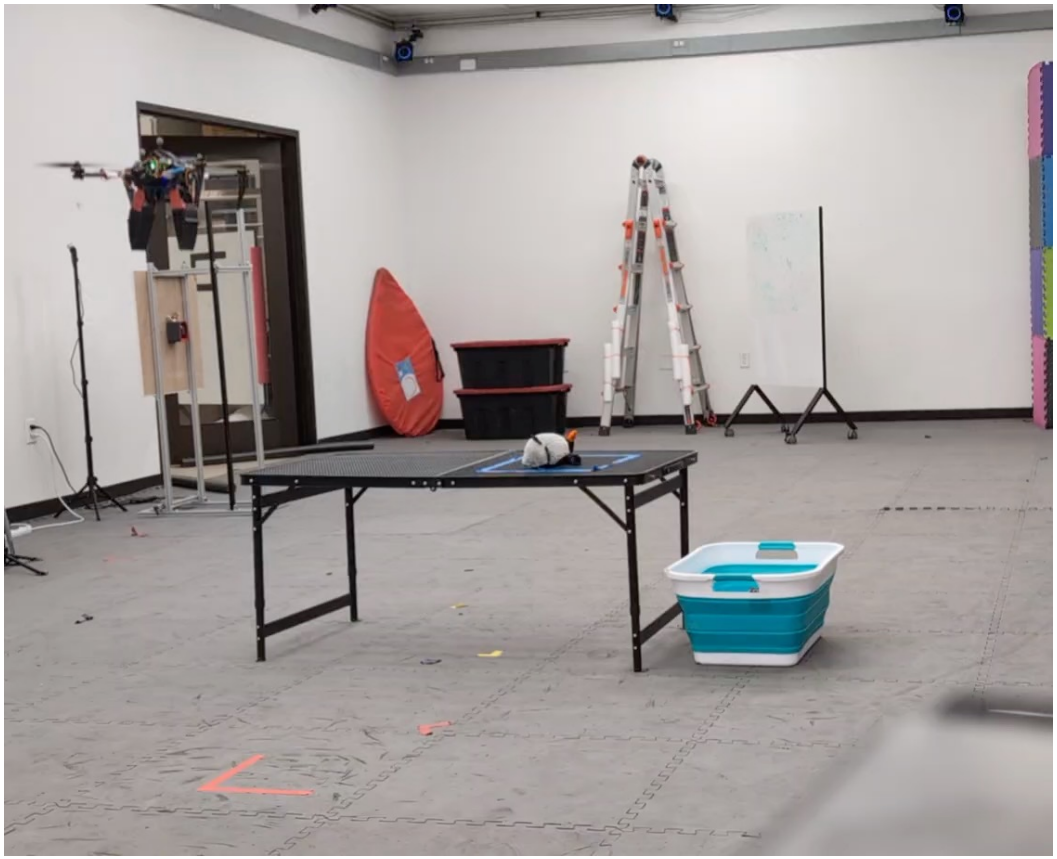
100 Synthetic FiGS Expert Demos

Trained policy with the real drone





Trained policy with the real drone



Does it always work?

Big Picture Outlook for Robot Learning

Current paradigm: We need 100Ms of high-quality demonstrations (we're not even close)

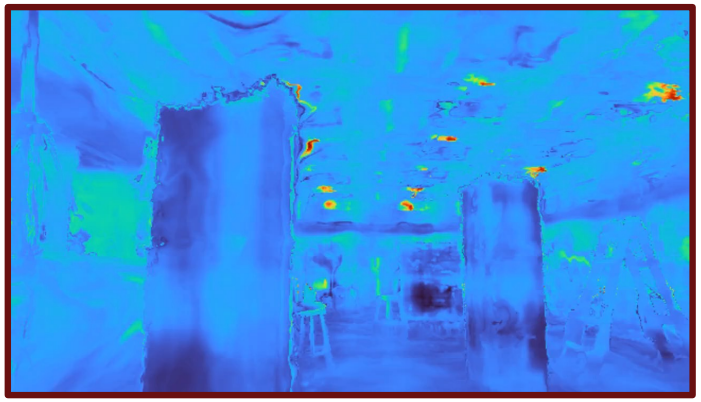
- Better human operators? More operators?
- Better tele-op interfaces?

Better world models

Broad-context self-evaluation models

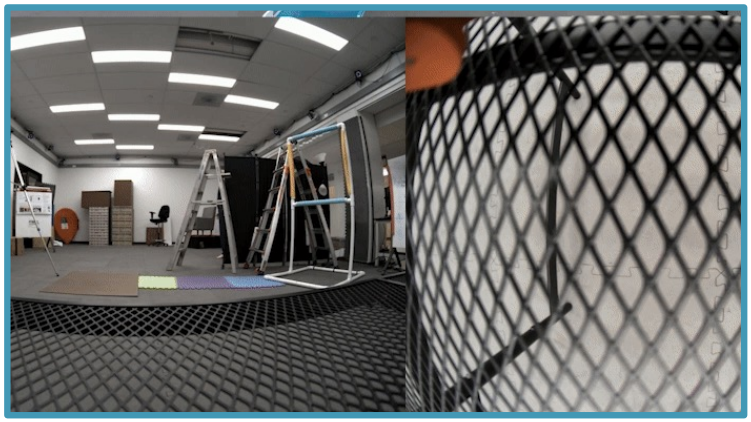
Policies that learn from their own experience

Splat-Nav



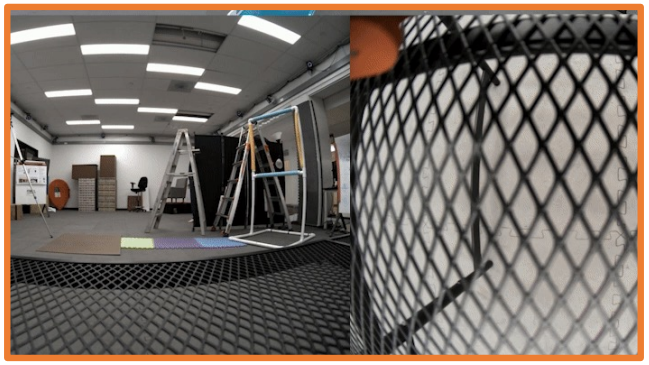
Sous Vide

DroneVLA



TL;DR

3D world models will transform robot autonomy

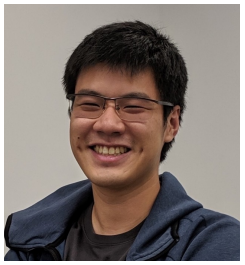


Thank you!

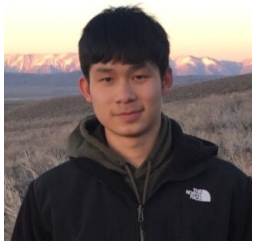
Ola Shorinwa

John Tucker

Javier Yu



JunEn Low



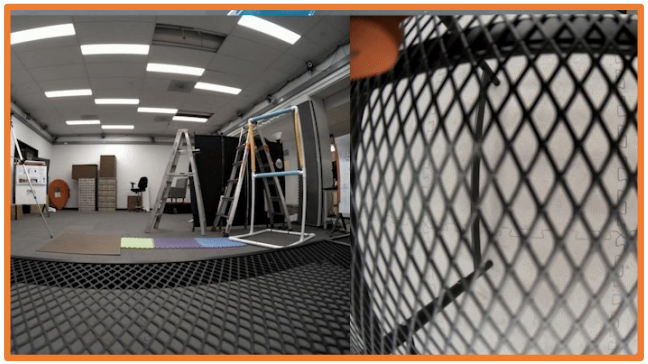
Tim Chen

Keiko Nagami

Max Adang

Prof. Philip Dames

Questions?



Stanford University

MSL

https://msl.stanford.edu/