Structured World Models for Robots

Krishna Murthy Jatavallabhula

Research Objectives

Modern AI technologies have achieved remarkable success in processing language, images, and speech, yet they falter in unpredictable and unintuitive ways when deployed on robotic systems in the physical world. Their abilities to *perceive*, *reason*, and *act* in the real world pale in comparison to humans and other biological entities, raising the critical question: *how do we bridge this gap*?

The success of today's machine learning approaches hinges on the availability of large volumes of high-quality data. In a robotics context, such data must be acquired by interacting with the real-world; which is infeasible considering the diversity of embodiments, environments, and tasks of interest. Developing general-purpose robots capable of autonomous operation across a wide range of environments, undertaking tasks routinely accomplished by humans necessitates advancements along multiple fronts: novel algorithms for sensorimotor control, computational learning frameworks, and cognitive architectures.

My research focuses on designing structured world models that will enable embodied intelligence systems (robots, mixed reality devices, intelligent visual assistants) **to *perceive*, *reason*, and *act* in the real world just as humans are able, and ultimately surpass human-level intelligence.**

Building structured world models will enable more robust operation, and with significantly lesser amounts of data, by drawing on our rich understanding of the physical world. This also helps in the identification of newer computational paradigms (such as differentiable and probabilistic computing) and the implementation of cognitive abstractions (such as analysis-by-synthesis) that are cruicial for robots to understand their environment and accomplish tasks therein.

My research group will focus primarily on the following themes:

- **Spatial and Semantic Understanding**: Developing visual perception systems that effectively represent the spatial structure and semantics of the environment for robots.
- **Physical Understanding**: Devising computational models to understand the physical properties of objects in the robot's environment, facilitating interaction.
- **Multimodal Understanding**: Integrating cues from other modalities such as audio, touch, and language, to enhance and robustify our understanding of the physical world.

I will devise novel ways of combining our vast wealth of prior knowledge of real-world phenomenon with modern learning-based approaches, bringing together the best of both worlds. Doing so will enable robots to accomplish a number real-world tasks that are "*seemingly*" trivial for humans, but currently impossible for AI systems. **My research program complements, extends, and synergizes best with existing robotics, computer vision, machine learning, and graphics research clusters.**

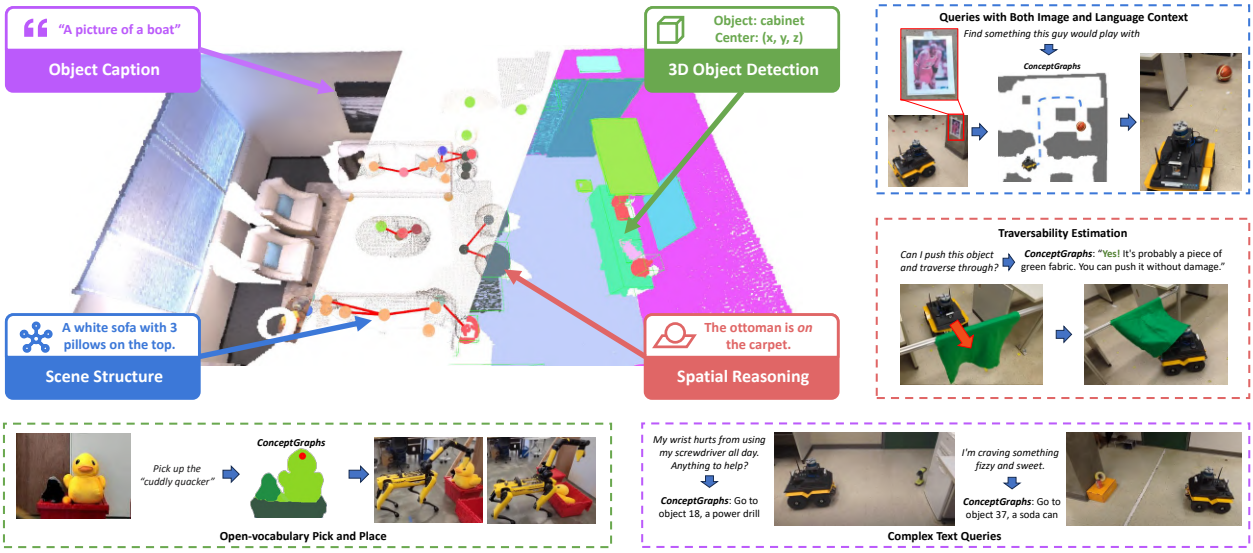

Figure 1: **ConceptGraphs** builds 3D *scene graphs* from RGB-D images and camera poses. The 3D scene graphs comprise nodes (objects in the scene) and edges (spatial/semantic relationships among objects). Different from prior work on 3D scene graphs, ConceptGraphs is *open-vocabulary*, meaning we do not assume a predetermined set of node/edge types, and can represent a large space of concepts that adapt to a downstream task. The object-centric nature of ConceptGraphs allows easy map maintenance and promotes scalability, and the graph structure provides relational information within the scene. Furthermore, our scene graph representations are easily mapped to natural language formats to interface with LLMs, enabling them to answer complex scene queries and granting robots access to useful facts about surrounding objects, such as traversability and utility. We implement and demonstrate ConceptGraphs on a number of real-world robotics tasks across wheeled and legged mobile robot platforms. (Webpage)

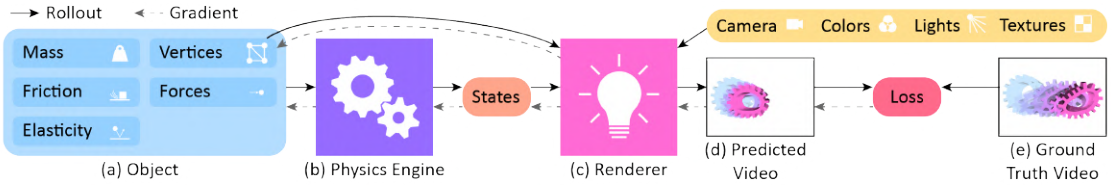

Figure 2: **∇Sim** is a GPU-based differentiable simulator, comprising a differentiable multiphysics engine and a differentiable renderer. The differentiable multiphysics engine takes as input a scene description (objects, states, material properties, etc.) and runs a forward simulation to compute states. These states and additional scene parameters (texture/lighting) are fed to a differentiable rendering engine which renders out a video sequence. This *predicted video* is compared to an observed video to define a discrepancy/loss. Since the entire simulation process is differentiable, we can perform gradient-based optimization to recover the true physical and rendering parameters, for a range of simulations: rigid bodies, deformable solids, thin-shell solids (cloth), and incompressible fluids. As opposed to prior work on differentiable physics engines, which assumes the availability of 3D supervision over object states, ∇Sim unifies physics and rendering engines, enabling physical parameter estimation and visuomotor control from 2D images. (A short video summary)

1 Spatial and Semantic Understanding

Under this theme, my research group will develop algorithms to model the 3D geometric structure and semantics of environments for robotic interaction. This will enable robots to localize themselves and operate safely within the environment. We integrate traditional 3D mapping techniques with ideas from modern machine learning, creating semantic 3D maps that provide detailed object information, facilitating a broader range of tasks.

1.1 Contributions and ongoing initiatives

Differentiable computing for 3D mapping: In ∇SLAM(gradSLAM) [1], I introduced a differentiable computing approach to simultaneous localization and mapping (SLAM). By making each computation in the SLAM pipeline differentiable—introducing reparameterizations for non-differentiable operators where necessary—I enabled gradient computation and propagation through the SLAM system. This allows for backpropagating errors from geometric reconstructions to sensor inputs, enabling SLAM systems to be integrated as layers within neural networks. My work led to the first fully differentiable SLAM system, laying the groundwork for combining learning-based methods with traditional SLAM (as I pursued further in [2, 3, 4]).

3D scene graphs for perception and planning: I developed approaches to build symbolic abstractions, 3D scene graphs, from real-world RGB-D perception, enabling a rich semantic understanding of the scene [4] (Fig. 1). I also designed algorithms that can efficiently (and provably optimally) produce task plans over large scene graph state spaces by task-conditioned pruning to eliminate extraneous scene graph components, reducing planning complexity [5].

Integrating differentiable 3D vision and graphics: Vision and graphics are inverse problems: while vision deals with lifting images to 3D, graphics deals with rendering 3D scenes to create realistic images. Systematic integration of vision and graphics pipelines will enable the design of self-supervised and weakly-supervised learning techniques that can leverage discrepancies across the two (inverse) processes for gradient-based learning. I leverage this paradigm in [6, 7, 8].

1.2 Future directions

Active, Persistent Perception: Much of the work that exists in the spatial understanding space (including my work listed above) is *episodic*: mapping happens afresh each time a scene changes, and happens so passively. However, this mechanism of acquiring scene representations is in stark contrast with humans and other biological entities. They possess persistent representations of the world that allow them to rapidly infer object state changes over time. My work will shift away from episodic and passive perception, to *active, persistent* perception, where robots actively decide on their subsequent actions, which in-turn affect their subsequent sense-data.

Functional spatiotemporal abstractions for perception and planning: I will also devise *task-centric*, opposed to geometry-centric scene representations; where the downstream task of interest determines the objects that a mapping system must represent in its spatial model. I am particularly interested in building scene representations that enable task planning at varying levels of spatiotemporal and symbolic abstraction.

2 Physical Understanding

Under this research theme, my research group will develop computational models that will enable robots to understand the physical properties of objects and scenes from images. Of particular interest are *structured* computational models that deeply integrate our understanding of physics and graphics, as they are data-efficient, interpretable, and easily reconfigured for new tasks.

2.1 Contributions and ongoing initiatives

Differentiable world models for physical understanding: In a series of works [9, 10, 11, 7], I have developed computational models that perceive physical properties—such as mass, friction, elasticity—of objects in the world solely from videos. While innate to humans, such behaviors have proved extremely hard to replicate in machine learning systems, owing to the vast variability in terms of geometric and physical properties that real-world objects exhibit.

In ∇Sim [9] (Fig. 2), we presented, for the first time, a system that could accurately infer the physical properties of objects without requiring any training data a priori. Key to enabling this is a *differentiable simulation and rendering engine*, which models the phenomenon of video generation, including object physical constraints and dynamics, and image synthesis, in the form of a differentiable computation graph. Since all computations herein are differentiable, we can reliably optimize for accurate physical and material properties of objects by gradient-based inference. In [11], we extend this differentiable simulation framework to compute globally optimal solutions, leveraging Bayesian optimization. In PAC-NeRF [7], we employed neural rendering to simulatenously estimate both the geometric and physical properties of objects.

2.2 Future directions

Differentiable probabilistic programs for physical understanding: While differentiable simulators have enabled great strides in inferring physical properties from pixel observations and in visuomotor control, they are fundamentally unable to capture and reason about uncertainty in their observations and estimates; essential for real-world robotic interaction. To enable this, I will leverage modern probabilistic computing machinery to develop *differentiable probabilistic programs*, which in addition to being amenable to gradient-based inference, are also amenable to automated Bayesian inference. In a first step towards this, in [10], we developed a probabilistic program that infers a posterior distribution over the discrete, graph-structured kinematic chain and continuous-valued physical properties of objects.

Approximate simulators as robot world models: A precise physics simulator, whether differentiable or not, can greatly aid robotic interaction with objects. However, creating an exact world model is often overly complex or impractical. My hypothesis is that high-fidelity simulators are not necessary for all robotic tasks. Instead, we need approximate simulators that can generate sufficiently accurate predictions for specific tasks and quickly adapt to new object and environment configurations with minimal interaction. This approach aligns with the *intuitive physics engine* hypothesis, which proposes an innate, non-verbal, and algorithmic reasoning process. My research group will will design and employ approximate simulators for general-purpose robotics tasks requiring physical reasoning, such as object stacking, reconfiguration, and assembly.

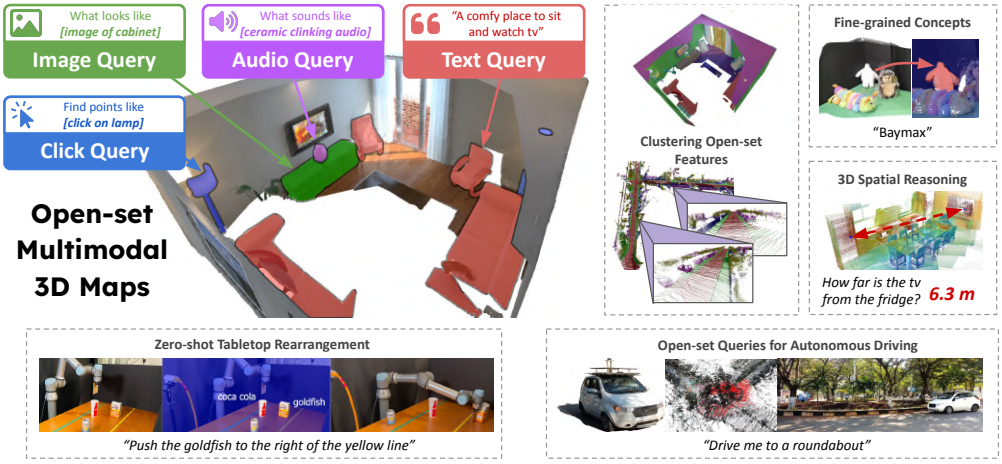

Figure 3: **ConceptFusion** builds dense 3D maps where each point in the map is assigned, in addition to 3D position, orientation, and color, vision-language(-audio) aligned representations extracted from foundation models. We demonstrate a surprisingly simple approach where none of the vision-language(-audio) models require 3D pretraining; we extract features over 2D images and employ traditional RGB-D fusion techniques to compute fused 3D features for each map point. These maps are built online, and can be queried for arbitrary concepts specified as text, images, audio samples, or clicks on the 3D map. The fused features have an implicit understanding of semantic concepts, as evident by visualizing clusters obtained from a K-means algorithm. ConceptFusion features are at retaining fine-grained long-tail concepts, such as the disney character "*Baymax*". We demonstrate ConceptFusion on real-world tabletop manipulation and urban autonomous driving. (Webpage)

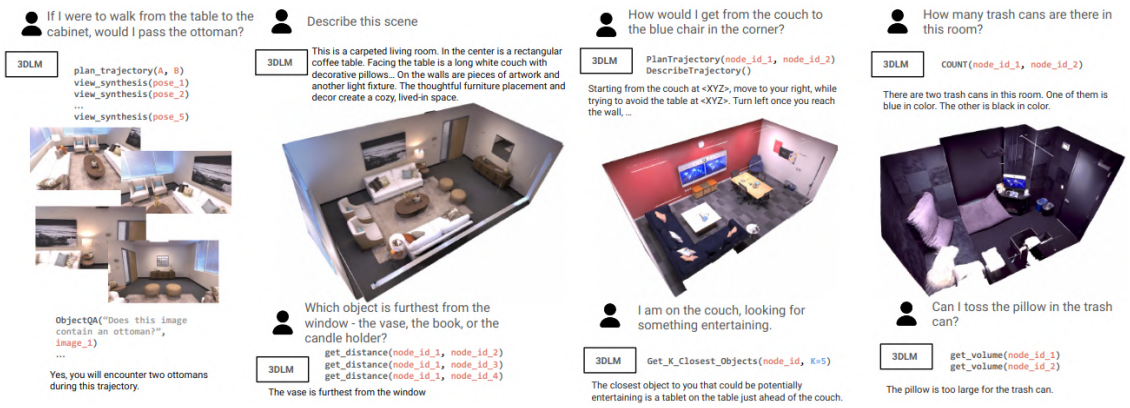

Figure 4: **3DLM** is a neurosymbolic approach to 3D scene understanding. We do not require any 3D pretraining and only assume access to frozen (potentially black-box) off-the-shelf language models (LMs) and image-language models (ILMs). We leverage a symbolic 3D scene representation (a 3D scene graph), a set of neurosymbolic modules operating over the 3D scene, and a language model to transform 3D understanding tasks posed in natural language to a series of instructions that are executed, to provide prescriptive and accurate responses to queries over an input 3D scene. Shown above is a set of representative outputs from 3DLM. Notice how we are able to address complex user queries, including those that require situated reasoning by composing spatial, logical, and neural operators including complex modules like trajectory planning and neural rendering.

3 Multimodal Understanding

Under this research theme, I will integrate multiple cues such as vision, audio, touch, and language to construct multimodal generative world models that offer far greater adaptability, robustness, and a richer understanding compared to visual understanding alone. This will help understand properties of the world that simply cannot be perceived via vision (weight, stability, roughness, temperature, etc.), and additionally robustify learning by developing models that can simultaneously process and explain multiple sensory stimuli.

3.1 Contributions and ongoing initiatives

Open-vocabulary Multimodal 3D Representations: While 3D maps are central to robot navigation, planning, and interaction, existing approaches that integrate semantic concepts with 3D maps have two key limitations. First, they are *closed-set*, meaning they can only reason about a finite set of concepts, pre-defined at training time. Second, they are *unimodal*, aggregating data solely from vision sensing. In ConceptFusion [3] (Fig. 3), I developed the first open-set and multimodal 3D mapping approach that integrated cues from images, language, and audio. This enables querying maps for arbitrary concepts, including those unseen during training, using diverse inputs such as text, sound, and images. We also developed ConceptGraphs [4] (Fig. 1), interfacing these compact 3D representations with large language models to handle complex user queries. For instance, ConceptGraphs can direct a robot to a duct tape roll in response to a query like "*find something to temporarily secure a broken zipper*". We also deploy this for outdoor navigation on a full-scale autonomous driving platform [12] and aerial vehicles [13].

Neurosymbolic approaches to 3D Understanding: Despite their impressive linguistic abilities, large (image-)language models have been demonstrably brittle, particularly for tasks involving (2D) spatial reasoning, localization, and symbol grounding; which is further exacerbated when designing such models for 3D scenes. In 3DLM [14] (Fig. 4), we developed a language model that interfaces with 3D scenes, enabling users to specify a diverse range of queries about the 3D scene in natural language. Opposed to relying solely on the ability of large transformer models to reason about spatial and logical attributes, we leverage an explicit symbolic scene structure, a 3D scene graph, in conjunction with neural and symbolic operators that enable prescriptive 3D reasoning capabilities. These queries enable a multitude of 3D understanding tasks including question answering and dialog, scene captioning, and referring object detection. We also explore the impact of such neurosymbolic operators in [3, 15].

3.2 Future directions

Multisensory digital twins of real-world environments: A significant challenge in combining multisensory understanding with machine learning is the lack of high-quality, real-world-aligned multisensory data. Existing simulation environments, while modeling aspects like physics, rendering, and increasingly, sound or touch, face an irreducible *reality gap*. This prohibits the deployment of approaches developed in such simulators into real-world scenarios. To this end, I aspire to develop multisensory simulation environments by *digitizing the real world* environments at high fidelity. This will drive the next wave of advancements in multisensory learning.

References

* indicates equal first-authorship † indicates equal advising

[1] **Krishna Murthy Jatavallabhula**, Ganesh Iyer, and Liam Paull. gradslam: Dense slam meets automatic differentiation. In *ICRA*, 2020. 3

[2] Dominik Muhle, Lukas Koestler, **Krishna Murthy Jatavallabhula**, and Daniel Cremers. Learning correspondence uncertainty via differentiable nonlinear least squares. In *CVPR*, 2023. 3

[3] **Krishna Murthy Jatavallabhula**, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, Ayush Tewari, Joshua B. Tenenbaum, Celso Miguel de Melo, Madhava Krishna, Liam Paull, Florian Shkurti, and Antonio Torralba. Conceptfusion: Open-set multimodal 3d mapping. In *RSS*, 2023. 3, 6

[4] Qiao Gu*, Alihusein Kuwajerwala*, Sacha Morin*, **Krishna Murthy Jatavallabhula***, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *arXiv*, 2023. 3, 6

[5] Chris Agia, **Krishna Murthy Jatavallabhula**, Mohamed Khodeir, Ondrej Miksik, Vibhav Vineet, Mustafa Mukadam, Liam Paull, and Florian Shkurti. Taskography: Evaluating robot task planning over large 3d scene graphs. In *CoRL*, 2021. 3

[6] Nikhil Keetha, Jay Karhade, **Krishna Murthy Jatavallabhula**, Gengshan Yang, Sebastian Scherer, Deva Ramanan, and Jonathon Luiten. Splatam: Splat, track, and map 3d gaussians for dense rgb-d slam. *arXiv preprint*, 2023. 3

[7] Xuan Li, Yi-Ling Qiao, Peter Yichen Chen, **Krishna Murthy Jatavallabhula**, Ming Lin, Chenfanfu Jiang, and Chuang Gan. Pac-nerf: Physics augmented continuum neural radiance fields for geometry-agnostic system identification. In *ICLR*, 2023. 3, 4

[8] Andrew Spielberg, Cengiz Oztireli, Derek Nowrouzezahrai, Fangcheng Zhong, Konstantinos Rematas, **Krishna Murthy Jatavallabhula**, and Tzu-Mao Li. Differentiable visual computing for inverse problems and machine learning. In *Nature Machine Intelligence*, 2023. 3

[9] **Krishna Murth**y **Jatavallabhula**, Miles Macklin, Florian Golemo, Vikram Voleti, Linda Petrini, Martin Weiss, Breandan Considine, Jerome Parent-Levesque, Kevin Xie, Kenny Erleben, Liam Paull, Florian Shkurti, Derek Nowrouzezahrai, and Sanja Fidler. gradsim: Differentiable simulation for system identification and visuomotor control. In *ICLR*, 2021. 4

[10] **Krishna Murthy Jatavallabhula**, Miles Macklin, Dieter Fox, Animesh Garg, and Fabio Ramos. Bayesian object models for robotic interaction with differentiable probabilistic programming. In *CoRL*, 2022. 4

[11] Rika Antonova, Jingyun Yang, **Krishna Murthy Jatavallabhula**, and Jeannette Bohg. Rethinking optimization with differentiable simulation from a global perspective. In *CoRL*, 2022. 4

[12] Mohd Omama, Pranav Inani, Pranjal Paul, Sarat Chandra Yellapragada, **Krishna Murthy Jatavallabhula**†, Sandeep Chinchali†, and Madhava Krishna†. Alt-pilot: Autonomous navigation with language augmented topometric maps. In *preprint*, 2023. 6

[13] Alaa Maalouf, Ninad Jadhav, **Krishna Murthy Jatavallabhula**, Makram Chahine, Daniel M Vogt, Robert J Wood, Antonio Torralba, and Daniela Rus. Follow anything: Open-set detection, tracking, and following in real-time. *arXiv preprint arXiv:2308.05737*, 2023. 6

[14] Shivam Chandhok*, **Krishna Murthy Jatavallabhula***, Chaitanya Devaguptapu, Qiao Gu, Deepti Hegde, George Tang, Connie Jiang, Sarah Schwettmann, Joshua B. Tenenbaum, Vibhav Vineet, Ondrej Miksik, Vineeth Balasubramanian, Leonid Sigal, and Antonio Torralba. Neurosymbolic language models for 3d understanding. In *arXiv*, 2023. 6

[15] Vikrant Dewangan, Tushar Choudhary, Shivam Chandhok, Shubham Priyadarshan, Anushka Jain, Arun Singh, Siddharth Srivastava, **Krishna Murthy Jatavallabhula**†, and Madhava Krishna†. Talk2bev: Language-enhanced bird's-eye view maps for autonomous driving. In *arXiv*, 2023. 6