Caltech
Student-Faculty Programs

Session J Abstracts

How do multimodal language models reason about space and time?
Hongqiao Chen
Mentors: Pietro Perona, Georgia Gkioxari, and Raphaela H. Kang

Spatiotemporal reasoning about visual input in the brain proceeds through a multi-step, structured process. With VLMs we have the unique opportunity to decompose latents and uncover exact mathematical mechanisms by which such reasoning can take place. In this work, we investigate whether spatiotemporal structure emerges in the textual activations of autoregressive vision-language models (VLMs), both with and without reinforcement learning-based reasoning strategies, and if they can be mechanistically disentangled and manipulated in the latent space. We show that VLMs without explicit supervision emergently employ a multi-hop mechanism for spatial and temporal queries, wherein intermediate spatiotemporal IDs, which are ubiquitous across the model, are encoded as linearly decomposable latents in textual activations. Perturbing these IDs in key modality alignment layers significantly steers the model's output, revealing their causal role in reasoning. We show that spatiotemporal IDs can serve as a diagnostic tool for pinpointing limiting reagents in the VLM pipeline, or as an internal learning signal that encourages structured reasoning. Monitoring these IDs also reveals that explicit reasoning traces enhance robustness to incorrect internal beliefs. By identifying and analyzing spatiotemporal IDs, we offer new insights into the internal reasoning mechanisms of VLMs, with implications for interpretability and the principled design of more aligned and capable models.

Rigorous investigation of human baselines towards more comprehensive VQA benchmarks and robust VLMs
Olivia Y. Wang
Mentors: Pietro Perona, Katelyn Haly, and Raphaela H. Kang

The rapid growth of large Vision Language Model (VLM) capacity is largely owed to the practice of benchmarking. Vision Question Answering (VQA) benchmarks are image-text question sets which allow for a controlled environment to assess how well models can understand and reason about images and text together. Despite these intentions, the construction of these benchmarks lacks standardization and these benchmarks are frequently misapplied. Human baseline evaluation across different benchmarks are evaluated in undercontrolled varied settings, making actual human-experienced difficulty on each task unclear. While the benchmarks offer a great way to get a general sense of model performance in VQA, they do not offer any insights into how specific model architectures may be affecting the final performance. In addressing these points, we collect human evaluations on ten VQA benchmarks across 100 subjects. Evaluating the effects of question presentation order, timing, and image resolution on human evaluation compared to model evaluation, we offer a paradigm for vision scientists to follow in the future when creating new benchmarks or model architectures.

Knowledge graph-informed predictions of multigene transcriptional profiles
Annika S. Viswesh
Mentors: Pradeep Ravikumar, Frederick D. Eberhardt, and Chandler Squires

Predicting how gene perturbations alter cellular transcriptional profiles is fundamental for understanding genetic interactions, designing combinatorial therapies, and engineering new cell states. Recent foundation models incorporate graph-based representations of genes and perturbations to improve prediction for both single and multigene cases. However, these models are limited to narrow knowledge graphs, where gene relationships are largely defined by co-expression rather than functional interactions within transcriptome pathways. Furthermore, genetic perturbations are not

treated as probabilistic priors, which limits interpretability into how different perturbations interact to shape transcriptional response. The contributions of our work are threefold. First, to more accurately represent gene interactions within molecular networks, we incorporate biomedical knowledge graphs to extract information on shared reactions, pathway membership, and molecular function, and encode these relationships through pre-trained language models. Secondly, we use a mixture-of-experts approach combined with permutation-invariant perturbation embeddings to form post-perturbation representations that distinguish between different genetic interaction patterns, making it possible to relate predicted transcriptional changes to the underlying type of interaction. Finally, we introduce a classifier-free guidance diffusion framework in place of linear cross gene decoders, allowing us to obtain predicted transcriptional profiles conditioned on post-perturbation information. We demonstrate that the diffusion component of the framework improves generalizability for both seen and unseen multigene perturbations. Evaluation of the constructed domain knowledge embeddings and mixture-of-experts modules is ongoing.

Gibbs conditioning principles for Markovian discrete-time interacting particle systems on regular graphs
Ritvik S. Teegavarapu
Mentors: Kavita Ramanan, Franca Hoffmann, I-Hsun Chen, and Sarath Yasodharan

This project investigates Gibbs conditioning principles for the empirical neighborhood measure of discrete-time Markovian interacting particle systems evolving on regular graphs. The dynamics of each vertex state are governed by local Markov transition kernels acting on the vertex neighborhood, leading to dependence structures that preclude direct application of classical tools. We express the Gibbs conditioning principle dynamic LDP for the sequence of empirical measures, identifying a rate function expressed as a relative entropy correction by the log-likelihood ratio of interacting and non-interacting systems. This rate function characterizes rare fluctuations away from the law of large numbers limit, known as the local field equation (LFE). For the one-step case, we formulate and solve a variational problem subject to co-variation constraints, yielding an explicit form of the Gibbs conditioning principle. The resulting optimizer exhibits a modified Markov kernel with exponential tilts determined by Lagrange multipliers enforcing marginal consistency. We further analyze the case of 2-regular graphs, where the limiting graph is the integer lattice, and the LFE admits tractable structure, providing insights into the most likely system evolution under rare event conditioning.

Safe offline reinforcement learning in noisy environments with digital twin-supported medical decision making
Aaron M. Dumas
Mentors: Rose Yu, Eric V. Mazumdar, and Aysin Tumay

Mean aortic pressure (MAP) is a primary factor in systemic hemodynamics and organ perfusion. Developing safe strategies for weaning cardiogenic shock patients from mechanical circulatory support (MCS) flow devices requires sequential decision-making under uncertainty. Offline reinforcement learning (RL) offers a promising framework to learn such strategies with retrospective data, avoiding the safety risks of online exploration. However, offline RL faces challenges such as distribution shift, where learned policies recommend actions not represented in the clinical dataset, and noisy signals from sensor error, medication effects, and changing patient states. A major barrier to clinical implementation is the lack of a safe online environment to test learned policies. To address these issues, we pair offline RL with a Transformer-based digital twin (TDT) of the circulatory system to simulate the environment. We show that our TDT outperforms baseline models on predictive accuracy and uncertainty. We evaluate offline RL policies with medically grounded metrics: physiological reward based on MAP, heart rate, and pulsatility; Action Change Penalty (ACP) to discourage abrupt pump-level changes; and Weaning Score (WS) to capture appropriate flow reduction. This work highlights the current limitations of offline RL and provides a practical methodology for rigorously testing policies prior to deployment, addressing safety and reliability challenges.

Task-aware mesh decomposition for Sim2Real robotics learning
Justin Luo
Mentors: Hao Su and Gunter Niemeyer

Sim2Real robotics training depends on a realistic and efficient simulation process; however, accurate physical simulation is a major computational bottleneck. As such, mesh decomposition algorithms such as CoACD have emerged, allowing for simulation agents to be simplified into convex meshes for simulation. The goal of this project is task-aware mesh decomposition: for many tasks, certain sections require more detail than others (e.g. when training a robot to pick up a mug, we can apply greater detail to the handle than the rest of the object). By optimizing our convex meshes to suit our tasks, we can save huge amounts of computational complexity while maintaining a realistic interface between all interacting agents. With fewer meshes to consider training can be accelerated massively, allowing robotics researchers to iterate models at a faster pace.

Implementing finite element modeling in the MS-Human-700 musculoskeletal model
Ellie J. Wang
Mentors: Yanan Sui and Joel W. Burdick

The MS-Human-700 is a whole-body musculoskeletal model trained to move with reinforcement learning. This project seeks to increase the anatomical fidelity of the MS-Human-700 by applying finite element modeling (FEM), a modeling technique that simulates objects as collections of small, deformable parts. We adapt soft-body modeling strategies from the Toyota Total Human Model for Safety (a leading FEM model used for injury simulation) into the MS-Human-700 physics engine and training algorithm. This novel implementation enables body components such as extremities to move more realistically under trained muscle control while maintaining computational efficiency. By increasing the realism of movement dynamics, this approach may introduce sensorimotor experience in simulation and provide richer training data for future motor learning research.

Modular motion tokenization for high-dimensional motor control: A case study on MS-Human-700
Erica Wang
Mentors: Yanan Sui and Adam C. Wierman

Tokenization is a foundational tool in natural language and vision, enabling composable representations that scale effectively in large models, such as GPT. In current motion control literature, discrete tokenization remains largely unexplored. In this work, we propose a novel framework for the tokenization of complex motor trajectories in musculoskeletal systems, with a focus on MS-Human-700, a highly-detailed humanoid model featuring over 700 muscle actuators. Our approach leverages vector quantization techniques and byte-pair encoding to form compositional representations of full body motion. These tokenizations capture reusable structure across the control space, allowing for efficient modeling, planning, reinforcement learning in high-dimensional motor tasks.

Adaptive modeling and exogenous support for neuromotor-impaired musculoskeletal models
Yingyin Tan
Mentors: Yanan Sui and Gunter Niemeyer

We present an adaptive, simulation-driven framework for designing and optimizing gait-assistive exoskeletons for individuals with neuromotor impairments. Stroke-related gait deficits are modeled in the MS-Human-700 musculoskeletal framework by scaling muscle activations to reflect neural signal degradation and validated against clinical gait data. Four one-dimensional actuators are placed at key body points, and the Dynamical Synergistic Representation with Soft Actor-Critic (DynSyn-SAC) algorithm identifies optimal magnitudes and directions of assistance forces across the gait cycle. The optimized force patterns aid the redesign of the single-motor "Exohip" exoskeleton developed by the

Ren Group. Using the MPC² model-based control framework, we test proposed mechanical changes in near real time, ensuring new designs deliver the simulated assistance benefits while respecting mechanical constraints. This approach supports the creation of more efficient and personalized exoskeletons for diverse neuromusculoskeletal impairments without requiring structural redesign.

Learning affine maps for an amortized noise-agnostic filter
Gautham Kappaganthula
Mentors: Andrew M. Stuart and Bohan Chen

Classical data assimilation (DA) methods typically assume access to the observation operator H and the observation-noise covariance Γ. In many practical settings, neither is known: only noisy measurements are available, or H and Γ can be inferred only crudely, which degrades the performance of Kalman-type filters. We propose a learning-based analysis step that learns an affine mean-field map acting on ensemble mean and anomalies, which recovers the Ensemble Kalman Filter and Ensemble Square Root Filter updates under specific parameterizations. However, our method is noise- and operator-agnostic and does not require explicit H or Γ. To amortize across different noise levels, ensemble sizes, or even observation operators, we train with a reweighted loss that balances optimization across noise and ensemble size, and can mix different observation operators during training. Experiments on Lorenz '63, Lorenz '96, and the Kuramoto–Sivashinsky systems show that our learned affine map performs comparably to classical methods with access to true H and Γ and significantly outperforms classical baselines that rely on approximate or unknown noise information.