Research Statement

My research is driven by the goal of constructing AI agents that interact with the physical world. My research tackles this through the use of **generative AI** – generating, for instance, the sequence of actions needed to manipulate an object or a video predicting the future states to follow to finish a task. Applying generative AI to decision-making poses fundamental challenges compared to settings where generative AI has excelled, *e.g.* language and images. First, there is a fundamental **lack of data** on how an agent should act, and second there is a need for **generalization beyond the training data**, as an agent is likely to encounter many new situations that are not in training data.

To combat these challenges, my work focuses on the idea of **compositional generative modeling**, where we construct generative models over a complex domain by learning and **combining** several simpler generative models [7, 8, 9]. Each simpler generative model captures a sub-aspect of the domain of interest, *e.g.* we can represent a generative model over complex scenes by combining a set of generative models representing individual objects in the scene. Individual generative models are substantially easier to train than a joint generative model, as they only require datasets of scenes with single object labels, which are substantially easier to gather. Simultaneously, such a composition of models can **generalize to situations outside the training data**, so long as each individual component is locally conditioned on something in its training distribution – for instance, we can generate scenes with a set of 9 objects even if training scenes have 5 objects by combining 9 object-level generative models learned from training scenes.

To combine multiple generative models, my work views each generative model as **parameterizing an energy landscape over datapoints** (referred to as an Energy Based Model)[1]. Sampling from a single generative model corresponds to optimizing for a generation with low energy. A composition of multiple generative models then corresponds to a new energy landscape (*i.e.* by summing the energy values across landscapes from each component generative model), which corresponds to operations such as the product, mixture and inversion of different generative distributions [9]. To enable effective sampling from energy landscapes, in my work, we proposed the use of **Langevin dynamics**, where the gradient of the energy value iteratively refines/denoises a sample initialized with random noise [1]. This proposed sampling procedure led to subsequent development of score-based models [2] and the modern implementation of diffusion models [3].

My work has illustrated the applicability of composing generative models across a variety of domains. First, in the visual setting, in [8], we illustrate how score functions in diffusion models can be composed to generate novel images. Compositions such as conjunction and negative prompts have seen widespread adoption in image generation such as in the Stable Diffusion WebUI. In action prediction, in [16], we illustrate how composing a generative model over trajectories with a model over goals enables the synthesis of trajectory plans generalizing across various goals. Next, by combining multiple foundation models together, in [29], we illustrate how joint sampling over the composition of a LLM, video, and action generative models enables a hierarchical planning procedure that can solve long-horizon tasks by training on Internet data. Finally, I've illustrated how composing generative models can be applied to science in [27, 31], where we design *de-novo* proteins and materials very different than those seen at training.

Compositional Generative Modeling through Energy Based Models

I started working on Energy Based Models (EBM) in mid 2018, when GANs were the dominant generative model. In [1], Igor and I presented an approach to scale EBM training to modern generative tasks. We proposed to train EBMs by drawing samples using Langevin dynamics, where an image sample was initialized from random noise and then generated by iteratively denoising using the gradient of the energy function. The EBM landscape is then trained using the generated samples. This approach inspired follow-up work in score-based models [2], which adopted the Langevin

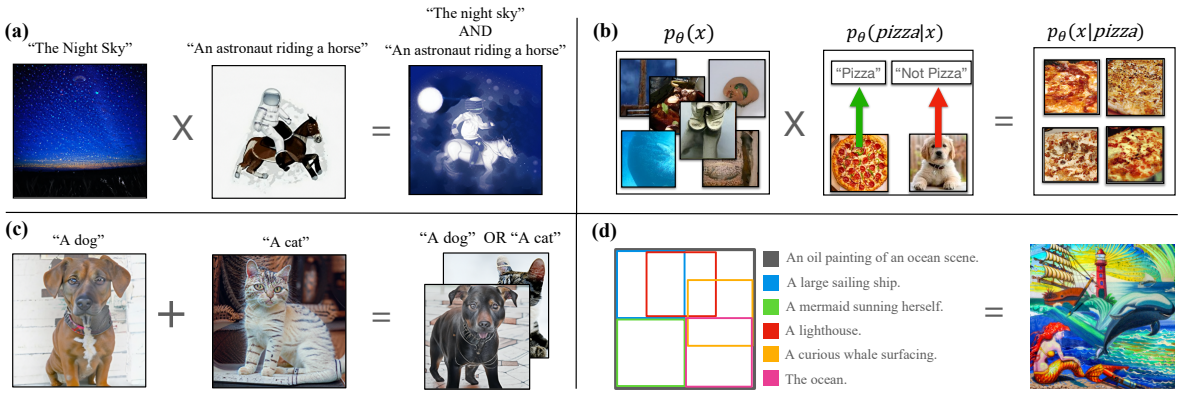

Figure 1: **Creating New Generative Models through Composition.** Simple operators enable generative models to be composed without retraining in settings such **(a)** conjunction of two text descriptions **(b)** conditional image generation, **(c)** mixtures of two text descriptions, **(d)** image tapestries with different content at different locations. All samples are generated through compositions of generative models.

dynamics sampling procedure, but substituted the EBM training objective with denoising score matching [4], another standard EBM training objective. This led to the modern implementation of diffusion models [3] which has since become the de-facto generative model for continuous inputs and can be seen as EBMs.

In [1], I provided preliminary experiments showing how an energy landscape perspective on generative models can enable multiple models to be composed, by directly sampling using Langevin dynamics on a combined energy landscape across models. I extended this to a set of probabilistic composition rules corresponding to set operations in [7]. My coauthors and I then showed how such compositions can be applied to diffusion models in [8], where we illustrated how multiple score functions in diffusion models can be directly combined to realize the probabilistic compositions. More recently, I presented techniques to theoretically accurately combine diffusion models in [9]. I illustrate several applications and examples of composing generative models in Figure 1. In the two sections below, I will illustrate how compositions of generative models can be directly applied to plan action trajectories and to construct hierarchical decision-making systems.

Planning Actions by Composing Generative Models

A major challenge in constructing a generative model across all the skills we want an agent to have is the lack of data – it is prohibitively expensive to gather demonstrations of each skill across all possible environments. We construct more data-efficient generative models over actions by factorizing the generative modeling problem into one model that captures the dynamics of the world and a separate model that captures the task we would like to accomplish (*i.e.* a goal or value function) [12, 16, 10, 6]. This decomposition enables the composed model to generalize to **unseen combinations of tasks and dynamics**. In [6], I presented an initial implementation of this idea using EBMs, which we illustrated in diffusion models in [16].

In [16], we explore several instantiations of the above factorization. First, we illustrated how combining a trajectory-level generative model with a value function estimating reward allowed us to use composition to construct policies on the fly that can solve several reinforcement learning tasks, performing competitively with the state-of-the art methods, without any explicit training on the task. Furthermore, we illustrated that by composing a trajectory generative model with a hand-coded model specifying a given start and goal state allows us to construct a goal-conditioned policy that substantially outperforms state-of-the art approaches, despite no explicit training. In follow up work, we illustrate additional compositional operations in [10] and further illustrate how

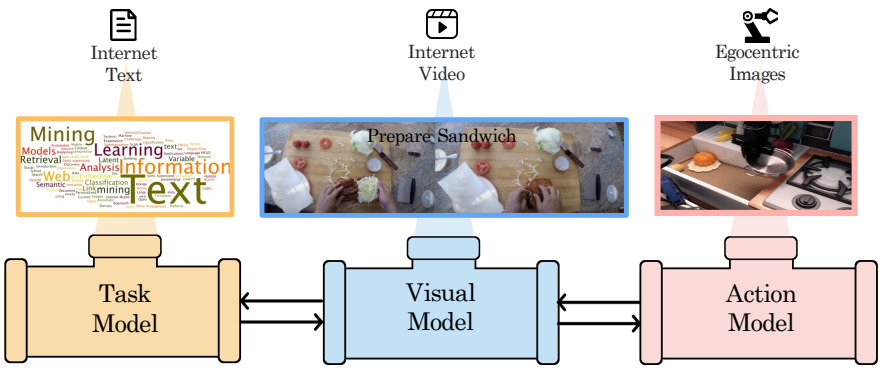

Figure 2: **Composing Foundation Models for Hierarchical Planning.** By combining multiple different generative models defined across several different modalities and at different timescales, iteratively sampling across the composed distribution can serve as planner to get a full long horizon plan.

such an action planning procedure enables dexterous real robot manipulation in [17]. We further illustrate how such a procedure allows us to finally synthesize plans directly in very high-dimensional space such as videos in [5], allowing us to learn a single planner across many different tasks.

Decision Making by Combining Foundation Models

Constructing monolithic generative models that can accomplish very long-horizon tasks such as making a cup of tea is especially difficult as it requires gathering demonstrations of an agent executing the full task across all possible configurations in the world. In my work, we propose to instead construct compositional generative models over such tasks by composing a set of three foundation models trained on Internet data. First, we model the task at an abstract level, *i.e.* the high-level steps needed to heat up tea, and use a LLM to capture this semantic information. We next model the task at a visual level, *i.e.* how the agent should visually move to avoid obstacles and use a video model to capture this dynamics information. Finally, we model the task at control level, *i.e* how the agent should actuate joints to lift a cup and use an egocentric action generative model mapping images to feasible actions. Acting is modeled as sampling from the composition of the generative models and corresponds to *optimizing* for a trajectory of actions, images, and language subgoals that are consistent across composed generative models (Figure 2). This system doesn't require any full long-horizon demonstrations to operate and can generalize as long as the abstraction each model operates on is in distribution.

Empirically, we found that such a composition can solve long-horizon tasks such as painting and rearranging blocks or setting a kitchen, substantially outperforming approaches that predict actions using a monolithic generative model, even if such data is available [29]. In [32], I further illustrate how composition, in the form of planning, can synthesize video plans substantially longer horizon than previously possible such as moving blocks to a line as well as enabling corresponding real robot execution. In [33], I also illustrate how composition, in the form of debate, can be applied to multiple instances of language models to enhance both factuality and promote systematic reasoning.

Future Research Plan

In my research, I've focused on how to construct complex generative systems not by constructing increasingly large and monolithic models, but by combining diverse sets of generative models through a complex online inference procedure. In my past research, I've explored a variety of compositions of models [7, 8, 16, 11, 29, 30] and various online inference procedures such as energy-optimization [7], planning [32], and debate [33]. In prior work, the explicit composition between models is fixed and pre-specified, requiring different combinations of models to be constructed per task. In future research, I'm interested in constructing fully decentralized generative systems, where individual

models communicate with each other in an ad-hoc manner, allowing model compositions to adaptively form given the task. I'm also interested in exploring how online inference procedures can aid generalization, more structured generative model training objectives, and the application of my work to engineering and scientific domains.

Decentralized Decision-Making Systems of Foundation Models. I believe we will see the emergence of an ever larger *zoo* of models with differing operating modalities and capabilities. In future research, I am interested in constructing a fully decentralized architecture for decision making, combining many such models, each responsible for independent tasks such as planning, perception, or control, for instance, something similar to a *factor graph of learned models*. I am also interested in exploring the rich set of online inference methods over such architectures, for instance, techniques from probabilistic graphical models (PGMs) such as *loopy belief propagation*. I believe the construction of such a decentralized system will enable us to combine the strengths of multiple institutions and researchers, allowing each researcher to develop a component of a system. Such a system will also be significantly more environmentally friendly and energy efficient than a single monolithic model, allowing individual models to be reused across different tasks, and removing the need at inference time to run a computationally large neural network.

Iterative Prediction Time Inference. I am interested in using online inference to construct neural network systems that generalize better to unseen situations. Neural networks typically perform poorly in unseen situations as they amortize computation and learn a fixed set of layers to make predictions. Online inference at prediction time allows the use of extra computation to adapt predictions to match new situations. As an example, in [13], we train a neural network to explicitly predict a scalar energy value estimating problem completion given a candidate solution and problem statement. We then apply the online inference procedure of energy minimization on this model to find a candidate solution that minimizes the energy estimation from the model. By iteratively refining a solution, our procedure generalizes much better to unseen inputs than feedforward networks. In future work, I plan to explore how other forms of online inference, for instance forward planning with a dynamics model, can improve neural network generalization.

Structured Generative Modeling. Since directly learning generative models across high-dimensional data is often intractable, my previous work uses compositional structure to construct more effective generative models. In future work, I'm interested in using structure to effectively learn generative models. In robotics, one source of structure over images is 3D geometry and my prior work has focused on inferring 3D structure [25, 24], and constructing generative models over such 3D structure [21, 18, 19, 20, 22]. I'm also interested in incorporating structure in the training procedure and moving beyond maximum likelihood estimation which often over-emphasizes irrelevant details in generations such as precise per-pixel frequencies. In [34] we propose instead train diffusion models using only reward functions, enabling direct optimization of generations to human preferences or scene compositionality. In future research, I'm interested in developing additional structured training procedures that prioritize desirable properties in generations, for instance, directly training generative models to learn casual effects or language models to be factually accurate, *e.g.* by explicitly penalizing predictions that lead to factually incorrect answers.

Broader Applications to Sciences and Engineering. My research focuses on constructing complex generative models, which we don't have sufficient data to train, from simpler generative models which we do have data to train. These composite models can be used to synthesize de-novo datapoints in engineering and scientific settings such as proteins in biology [26, 27], new robotic design [28], or material and surface design [31]. Generative models can also be used outside of data generation to construct generative classifiers, as a means to detect spurious or anomalous data, or as a means to invert and understand data.

References

[1] **Yilun Du**, Igor Mordatch. "Implicit Generation and Generalization with Energy Based Models." *Neural Information Processing Systems*, 2019.

[2] Yang Song, Stefano Ermon. "Generative Modeling by Estimating Gradients of the Data Distribution." *Neural Information Processing Systems*, 2019.

[3] Jonathon Ho, Pieter Abbeel. "Denoising Diffusion Probabilistic Models." *Neural Information Processing Systems*, 2020.

[4] Pascal Vincent. "A connection between score matching and denoising autoencoders." *MIT Press*, 2011.

[5] **Yilun Du***, Mengjiao Yang*, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, Pieter Abbeel. "Learning Universal Policies through Text-Conditioned Video Generation." *Neural information Processing Systems*, 2023.

[6] **Yilun Du**, Toru Lin, Igor Mordatch. "Model Based Planning with Energy Based Models." *Conference on Robot Learning*, 2019.

[7] **Yilun Du**, Shuang Li, Igor Mordatch. "Compositional Visual Generation with Energy Based Models." *Neural Information Processing Systems*, 2020.

[8] Nan Liu*, Shuang Li*, **Yilun Du***, Antonio Torralba, Joshua B. Tenenbaum. "Compositional Visual Generation with Composable Diffusion Models." *European Conference on Computer Vision*, 2022.

[9] **Yilun Du**, Conor Durkan, Robin Strudel, Joshua B. Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, Will Grathwohl. "Reduce, Reuse, Recycle: Compositional Generation with Energy-Based Diffusion Models and MCMC." *International Conference in Machine Learning*, 2023.

[10] Anurag Ajay*, **Yilun Du***, Ahbi Gupta*, Joshua B. Tenenbaum, Tommi S. Jaakkola, Pulkit Agrawal. "Is Conditional Generative Modeling all You Need for Decision-Making?" *International Conference on Learning Representations*, 2023.

[11] Shuang Li*, **Yilun Du***, Joshua B. Tenenbaum, Antonio Torralba, Igor Mordatch. "Composing Ensembles of Pre-trained Models via Iterative Consensus" *International Conference on Learning Representations*, 2023.

[12] Hongyi Chen*, **Yilun Du***, Yiyi Chen*, Joshua B. Tenenbaum, Patricio Vela. "Planning with Sequence Models through Iterative Energy Minimization" *International Conference on Learning Representations*, 2023.

[13] **Yilun Du**, Shuang Li, Joshua B. Tenenbaum, Igor Mordatch. "Learning Iterative Reasoning through Energy Minimization" *International Conference in Machine Learning*, 2022.

[14] Nan Liu*, Shuang Li*, **Yilun Du***, Joshua B. Tenenbaum, Antonio Torralba. "Learning to Compose Visual Relations" *Neural Information Processing Systems*, 2021.

[15] **Yilun Du**, Shuang Li, Yash Sharma, Joshua B. Tenenbaum, Igor Mordatch. "Unsupervised Learning of Compositional Energy Concepts" *Neural Information Processing Systems*, 2021.

[16] Michael Janner*, **Yilun Du***, Joshua B. Tenenbaum, Sergey Levine. "Planning with Diffusion for Flexible Behavior Synthesis." *International Conference in Machine Learning*, 2022.

[17] Cheng Chi, Siyuan Feng, **Yilun Du**, Zhengjia Xu, Eric Cousineau, Benjamin Burchfiel, Shuran Song. "Diffusion Policy: Visuomotor Policy Learning via Action Diffusion." *Robotics, Science, and Systems*, 2023.

[18] Anthony Simeonov*, **Yilun Du***, Andrea Tagliasacchi, Joshua B. Tenenbaum, Alberto Rodriguez, Pulkit Agrawal, Vincent Sitzmann. "Neural Descriptor Fields: SE(3)-Equivariant Object Representations for Manipulation." *International Conference in Robotics and Automation*, 2022.

[19] Anthony Simeonov*, **Yilun Du***, Yen-Chen Lin, Alberto Rodriguez, Leslie Kaelbling, Tomas Lozano-Perez, Antonio Torralba, Pulkit Agrawal. "SE(3)-Equivariant Relational Rearrangement with Neural Descriptor Fields." *Conference on Robot Learning*, 2022.

[20] Ethan Chun, **Yilun Du**, Anthony Simeonov, Tomas Lozano-Perez, Leslie Kaelbling. "Local Neural Descriptor Fields: Locally Conditioned Object Representations for Manipulation" *International Conference in Robotics and Automation*, 2023.

[21] Linqi Zhou, **Yilun Du**, Jiajun Wu. "3D Shape Generation and Completion through Point-Voxel Diffusion. " *International Conference in Computer Vision*, 2021.

[22] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong, Zheng, **Yilun Du**, Zhenfang Chen, Chuang Gan. "3D-LLM: Injecting the 3D World into Large Language Models. " *Neural Information Processing Systems*, 2023.

[23] Cameron Smith, **Yilun Du**, Ayush Tewari, Vincent Sitzmann. "FlowCam: Training Generalizable 3D Radiance Fields without Camera Poses via Pixel-Aligned Scene Flow " *Neural Information Processing Systems*, 2023.

[24] **Yilun Du**, Cameron Smith, Ayush Tewari, Vincent Sitzmann. "Learning to Render Novel Views from Wide-Baseline Stereo Pairs" *Computer Vision and Pattern Recognition*, 2023.

[25] Jiahua Fu, **Yilun Du**, Kurran Singh, Joshua B. Tenenbaum, John Leonard. "NeuSE: Neural SE(3)-Equivariant Embedding for Consistent Spatial Understanding with Objects" *Robotics, Science, and Systems*, 2023.

[26] **Yilun Du**, Joshua Meier, Jerry Ma, Rob Fergus, Alexander Rives. "Energy-Based Models for Atomic-Resolution Protein Conformations." ICLR 2020.

[27] Robert Verkuli*, Ori Kabeli*, **Yilun Du**, Basile Wicky, Lukas Milles, Justas Dauparas, David Baker, Sergey Ovchinnikov, Tom Sercu, Alexander Rives. "Language Models Generalize Beyond Natural Proteins." arXiv 2022.

[28] Johnson Wang, Juntian Zheng, Pingchuan Ma, **Yilun Du**, Byungchul Kim, Andrew Spielberg, Josh Tenenbaum, Chuang Gan, Daniela Rus. "DiffuseBot: Breeding Soft Robots With Physics-Augmented Generative Diffusion Models " *Neural Information Processing Systems*, 2023.

[29] Anurag Ajay*, Seungwook Han*, **Yilun Du***, Shuang Li, Abhi Gupta, Tommi Jakkola, Joshua B. Tenenbaum, Leslie Kaelbling, Akash Srivastava, Pulkit Agrawal. "Compositional Foundation Models for Hierarchical Planning." *Neural Information Processing Systems*, 2023.

[30] Zhutian Yang, Jiayuan Mao, **Yilun Du**, Jiajun Wu, Joshua B. Tenenbaum, Tomas Lozano-Perez, Leslie Kaelbling. "Compositional Diffusion-Based Continuous Constraint Solvers." *Conference on Robot Learning*, 2023.

[31] Tailin Wu*, Takashi Maruyama*, Long Wei*, Tao Zheng*, **Yilun Du***, Gianluca Iaccarino, Jure Leskovec. "Compositional Generative Inverse Design." *Neural Information Processing Systems Workshop on AI4Science* 2023.

[32] **Yilun Du**, Sherry Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tainhe Yu, Pieter Abbeel, Joshua B. Tenenbaum, Leslie Kaelbling, Andy Zeng, Jonathan Tompson. "Video Language Planning." arXiv 2023.

[33] **Yilun Du**, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, Igor Mordatch. "Improving Factuality and Reasoning in Language Models through Multiagent Debate." arXiv 2023.

[34] Kevin Black*, Michael Janner*, **Yilun Du**, Ilya Kostrikov, Sergey Levine. "Training Diffusion Models with Reinforcement Learning." arXiv 2023.