Doctoral Research Plan

Research Proposal: Structured and Adaptive Latent Representations for Systematic Generalization

1. Research Motivation and Core Question

My research is driven by a foundational question in artificial intelligence:

What structural properties must a representation possess in order to support reasoning and systematic generalization?

Despite remarkable progress driven by scale, modern neural systems often struggle with *compositional extrapolation*: the ability to recombine known operations into novel, longer-horizon behaviors. This suggests that reasoning is not merely a function of capacity, but of **representation form**. These limitations have become increasingly visible as modern neural models, including large language models, exhibit striking but brittle reasoning behavior—often failing under simple compositional variations.

Understanding why such failures occur, and how they might be addressed through representation design, is both theoretically important and practically pressing. Rather than viewing structure as a set of rigid, hand-designed rules, I am interested in **structural inductive biases as adaptive inductive priors**: constraints that shape representation geometry while remaining differentiable and context-sensitive.

In particular, I aim to investigate how **algebraic, geometric, and compositional inductive biases** can be incorporated into latent representations in a manner that allows **structure-preserving mappings** to emerge through learning. This perspective aligns closely with Prof. Yue Song's research program on **structured representation learning**, especially the goal of discovering relational and compositional regularities from data rather than prescribing them a priori.

---

2. Background and Prior Research

My prior research integrates **mathematical formalism with hands-on machine learning experimentation**, focusing on models whose latent representations support reasoning beyond interpolation.

In my Master's thesis, *On Why Form Shapes Reason*, I studied **Latent Program Networks (LPNs)**—neural architectures that represent transformations as continuous latent representations inferred from input–output examples. I introduced a framework for **structuring latent representations using algebraic constraints**, drawing inspiration from category theory. Specifically, I enforced properties such as:

- **Associativity**, to promote stable multi-step composition
- **Identity**, to anchor latent representations and reduce drift
- **Closure-like semantic consistency**, aligning composed latent representations with decoder behavior

Empirically, these structural constraints enabled **systematic multi-step generalization from single-step supervision**, without symbolic rules or test-time search. These results suggest that reasoning behavior can emerge when latent representation geometry is shaped appropriately, rather than when symbolic structure is explicitly imposed.

In parallel, I have explored **geometric and Lie-theoretic perspectives** on iterative neural reasoning, viewing refinement-based models as learned dynamical systems whose trajectories in latent space encode structured transformations. This further motivates treating structure as **soft, continuous, and adaptable**, rather than rigid or discrete.

3. Proposed Doctoral Research Direction

3.1 High-Level Objective

The overarching objective of my doctoral research is to **understand how structured latent representations support compositional reasoning**, and how the *form and strength* of structural inductive biases influence generalization.

Rather than assuming a fixed notion of structure, I aim to study **what kinds of representational constraints make reasoning natural, stable, and extensible**, and whether these constraints can function as **adaptive inductive priors learned from data**.

3.2 Key Research Themes

(1) Structured Latent Representations with Adaptive Constraints

I propose to investigate latent representations endowed with algebraic and relational structure, where **structural inductive biases are not fixed a priori**, but treated as **differentiable, adaptive priors**. This includes studying:

- Compositional operators that approximately satisfy algebraic laws
- Latent identities and invariances that stabilize long-horizon reasoning
- Parameterizations in which the *strength or applicability* of structural constraints varies during learning

A central hypothesis is that reasoning benefits not from maximal structure, but from **appropriately tuned structure**, and that this tuning may depend on context, data, or task complexity.

(2) Learning Composition Beyond Supervision

A second focus is **generalization beyond observed compositions**: training models on elementary transformations while evaluating their ability to extrapolate to unseen multi-step or hierarchical combinations.

Methodologically, this includes:

- Loss-based enforcement of compositional consistency
- Architectural separation between *representation* and *composition*
- Analysis of how adaptive inductive priors affect extrapolation behavior

This treats compositional generalization as a diagnostic signal of whether learned structure genuinely supports reasoning.

(3) Geometric Views of Iterative Reasoning

I am interested in viewing reasoning as **trajectories in structured latent spaces**, where repeated application of near-identity transformations accumulates into meaningful global behavior.

This perspective enables:

- Analysis of stability and drift in long-horizon reasoning
- Connections between latent geometry and generalization

- Controlled study of how adaptive structural constraints influence iterative refinement

4. Methodological Approach

Methodologically, this research will combine:

- **Theoretical grounding**, using algebraic and geometric principles to motivate hypotheses about representation structure
- **Empirical investigation**, through controlled synthetic and semi-structured tasks designed to isolate the effects of structural inductive biases
- **Adaptive mechanisms**, exploring whether the influence of certain inductive priors can be learned or adjusted during training rather than imposed uniformly

A key open question is whether structural constraints that prove effective in controlled settings remain beneficial in richer, noisier domains. This research treats that uncertainty not as a limitation, but as an **empirical question to be investigated systematically**.

5. Alignment with Prof. Yue Song’s Research

Prof. Yue Song’s work on **structured representation learning and relational inductive biases** provides an ideal intellectual environment for this research. His emphasis on learning structure from data aligns directly with my interest in adaptive inductive priors and principled representation design.

I am particularly interested in connections between **structured latent representations** and the notions of **learned homomorphism and equivariance**. From this perspective, a well-structured representation should act as an approximate homomorphic mapping between transformations in the data domain and operations in latent space, such that composition and symmetry are preserved in a learned, task-dependent manner. Rather than enforcing fixed equivariances a priori, I am interested in whether such structure can emerge as a soft, adaptive property of the representation through learning. This viewpoint offers a principled lens for studying how representation form supports reasoning and aligns closely with Prof. Yue Song’s work on discovering relational and equivariant structure from data.

Under Prof. Song’s guidance, I am particularly excited to explore:

- How relational and compositional structure can emerge through learning
- Whether disentangled representations can be extended to learn **operators**, not just factors
- How adaptive inductive priors affect generalization across tasks and domains

I see my background in algebraically and geometrically structured latent models as a natural complement to this research direction.

6. Long-Term Vision

In the long term, I aim to contribute to a theory of **how representation form shapes reasoning capability—**clarifying when structured inductive biases are necessary, how rigid they should be, and how they can be learned rather than prescribed.

Through this work, I aim to contribute **conceptual clarity and empirically grounded insight** to the study of reasoning in neural systems, helping move the field beyond scale-driven progress toward more principled and generalizable forms of intelligence.