Project Proposal: Phase Transitions in the Latent Space of Hierarchical VAEs

Huan Souza,[1, *] Insung Hwang,[1, †] Kristian Munnis,[1, ‡] and Shreyas Raman[1, §]

[1] *Department of Physics, Boston University, Boston, MA, 02215, USA*

The paper "A phase transition in diffusion models reveals the hierarchical nature of data" [1] demonstrates that the reverse diffusion process exhibits a phase transition, where high-level features (like object class) decohere at a sharp "time" threshold, while low-level features (like texture) evolve smoothly. This project will first reproduce this key finding using a standard diffusion model. We will then extend this analysis by replacing the paper's synthetic "Random Hierarchy Model" with the *discrete, hierarchical latent space* using a Vector-Quantized Variational Autoencoder (VQ-VAE). We will search for an analogous phase transition in this structured, discrete space to provide a more concrete link between data hierarchy and the statistical physics of generative models.

I. INTRODUCTION

Sclocchi et al. provide compelling evidence that diffusion models learn the hierarchical structure of data [1]. Their "forward-backward" experiments (Fig. 2 in their paper) show that as the diffusion time t increases, the ability to reconstruct an image's class identity suddenly collapses, while low-level features persist and are "re-used" in the new, class-switched image. This is a hallmark of a phase transition.

The paper's theoretical explanation relies on a synthetic, tree-based "Random Hierarchy Model" (RHM). While effective, this is a toy model. A more powerful test of their thesis would be to use a model that learns a *real* hierarchical data structure.

This is precisely what hierarchical VQ-VAEs do. They explicitly learn a set of *discrete, hierarchical* latent codes (e.g., top-level codes for global structure, bottom-level codes for texture). This provides a concrete, structured laboratory to test the phase transition hypothesis, bridging concepts from statistical physics (phase transitions, decoherence) with representation learning.

II. GOALS AND OBJECTIVES

1. **Reproduce:** Reproduce the key findings of Sclocchi et al. on a manageable dataset using a pre-trained unconditional diffusion model. This involves implementing the "forward-backward" experiment and plotting feature similarity vs. diffusion time to observe the phase transition.

2. **Extend:** Extend this concept to VQ-VAEs by playing with proposals in the latent space. As a first idea, we would proposal Monte Carlo steps in the latent space and study the behavior of the decoded image

3. **Analyze:** Investigate this new discrete latent-space process for an analogous phase transition. We hypothesize that the reconstruction probability of high-level (top) VQ codes will show a sharp drop at a critical noise threshold, while "low-level" (bottom) codes will decay smoothly. Moreover, we believe that the structure will be able to be read from the different layers of the latent space in VQ-VAEs.

[*] hsouza@bu.edu

[†] his5624@bu.edu

[‡] kmunnik@bu.edu

[§] shreyasr@bu.edu

III. METHODOLOGY

Part 1: Reproduction (Baseline)

- Use a pre-trained Diffusion Model.
- Implement the "forward-backward" experiment:
  1. Take an image x_0.
  2. Noise it to x_t using the forward diffusion process.
  3. Denoise x_t back to x'_0 using the pre-trained model.
- Use a pre-trained ResNet classifier to extract hidden-layer activations for x_0 and x'_0.
- Plot the cosine similarity of high-level (logits) and low-level (early layers) features vs. t. This will be our baseline, expected to match Fig. 2 of the paper.

Part 2: Extension (VQ-VAE Latent Diffusion if time permits)

- Use a pre-trained, hierarchical VQ-VAE.
- Define a discrete noising process (MCMC steps maybe) on the latent codes, where at each step t, codes are randomly flipped to another codebook index with some probability $p(t)$.
- Decode the new latent space.
- Plot the cosine similarity between images or layes in a classifier to study the phase transition as a function of noise.

IV. EXPECTED OUTCOMES

- A successful reproduction and verification of the phase transition discovered by Sclocchi et al.
- A novel analysis demonstrating (or refuting) an analogous phase transition in the discrete, hierarchical latent space of a VQ-VAE.
- A final report with clear visualizations comparing the two systems and discussing the implications for understanding data hierarchy through the lens of statistical physics.

[1] A. Sclocchi, A. Favero, and M. Wyart, A phase transition in diffusion models reveals the hierarchical nature of data, Proc. Natl. Acad. Sci. U.S.A. 122 (1) e2408799121, https://doi.org/10.1073/pnas.2408799121 (2025).