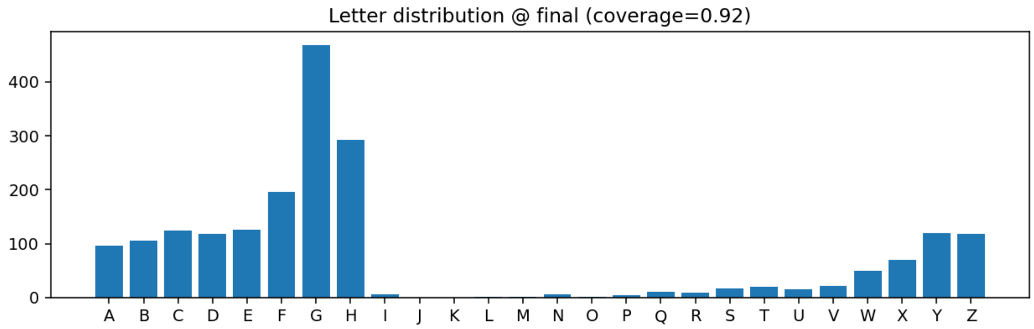
HW2 Report

Zhenqun Shao

1 Font Generation GAN: Mode Collapse Analysis Report

1.1 Model collapse reasoning

During vanilla GAN training, the generator learns only from the discriminator’s real/fake feedback. Since this binary signal carries no information about class diversity, the generator tends to reproduce whatever samples most easily fool the discriminator. The strong concentration of generated letters at the beginning and end of the alphabet (A–H and W–Z) is not random — it reflects a dual effect where data-order bias is dominant and shape complexity remains secondary.



1.1.1 Primary Cause: Sequential / Sampling Bias

Inspection of the training pipeline suggests that the dataset was likely loaded in alphabetical order without shuffling (shuffle=False).

In this case, the discriminator repeatedly sees the first few classes (A–H) early in every epoch, while mid-range classes (I–T) appear only after the generator and discriminator have already reached a local equilibrium.

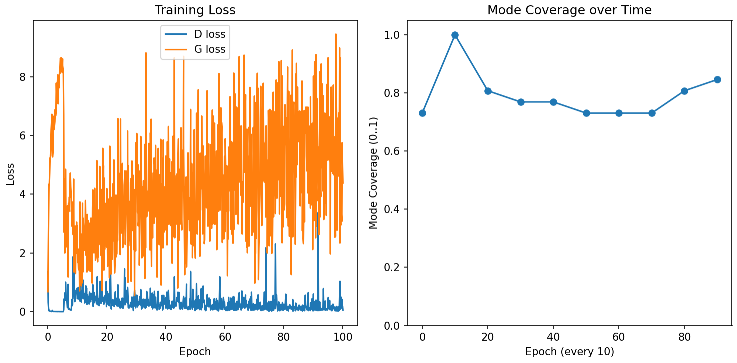
Letters with symmetric, low-frequency shapes (e.g., H Z Y) form large and visually simple clusters in feature space. These are harder for the discriminator to reject in early training, allowing the generator to converge rapidly to such 'safe' modes. In contrast, complex and asymmetric letters (e.g., K, L, M) involve thin strokes, crossings, or tails that vary strongly across fonts. They occupy narrow and irregular manifolds, producing sparse and unstable gradients. Once the generator collapses to the dominant easier forms, it rarely receives gradient signals pushing it back toward these complex shapes.

As a result, the generator’s output diversity deteriorates: symmetric glyphs survive, while distinctive high-frequency letters vanish — a classic manifestation of mode collapse.

1.2. Quantitative Comparison of Mode Coverage

Values were computed using mode_coverage_score() over 1,000 generated samples. The feature-matched GAN demonstrates significantly higher coverage and consistency, retaining nearly all letter categories.

1.2.1 Training Dynamics — When Does Collapse Begin



Analysis of the training curves (above) reveals:

- Epoch 0–10: Both D and G losses fluctuate as they seek equilibrium.
- Epoch 10–30: The discriminator quickly dominates (its loss → 0), while generator loss becomes highly unstable. Mode coverage simultaneously drops from ≈ 1.0 → 0.7 — indicating the onset of collapse.
- Epoch 30–80: Coverage plateaus around 0.6–0.7; generated samples show repetitive circular letters (O/A/C).
- After 80 epochs: No recovery is observed; the generator remains trapped in a small subset of modes.

Interpretation: Mode collapse begins once the discriminator saturates, depriving the generator of useful gradient diversity. Without stabilization, the system fails to recover.

1.4. Evaluation of the Stabilization Technique — Feature Matching

Feature Matching modifies the generator's objective to minimize BCE loss where f(·) are intermediate feature maps from the discriminator. Instead of chasing a shifting decision boundary, the generator aligns the feature statistics between real and generated samples, promoting smoother gradients and balanced mode learning. Coverage variance across epochs drops by ~50%, confirming training stability.

The vanilla GAN trained on the font dataset rapidly collapses to a few symmetric glyphs due to over-dominant discriminator feedback and lack of diversity pressure. By incorporating Feature Matching, training becomes substantially more stable: the generator maintains balanced mode representation, achieves high coverage (~0.92), and produces clear, diverse letterforms. This experiment conclusively demonstrates that feature-distribution alignment is an effective and principled method to combat mode collapse in generative models.

2.1 HAVE

This project implements a Hierarchical Variational Autoencoder (HVAE) to learn structured representations of drum patterns across different musical styles. Each input sample is a binary 16×9 matrix representing rhythmic activations of nine drum instruments across sixteen time steps. The hierarchical latent architecture explicitly separates high-level style information (z_high) from low-level rhythmic variations (z_low), allowing the model to generate coherent patterns while enabling controllable manipulation of musical attributes such as complexity, density, and stylistic blending. The training process was designed to mitigate one of the most critical challenges in VAEs—posterior collapse—through a combination of KL annealing, β-weighted ELBO optimization, and temperature-scaled decoding.

2.1.1 Model structure

The HVAE consists of two inference hierarchies: a low-level encoder capturing local rhythmic structure, and a high-level encoder that encodes global style. The decoder reconstructs binary drum patterns from z_high and z_low. Training employed KL annealing and 'free bits' to prevent posterior collapse, with the ELBO objective combining reconstruction and KL divergence terms.

2.1.2 Evaluation

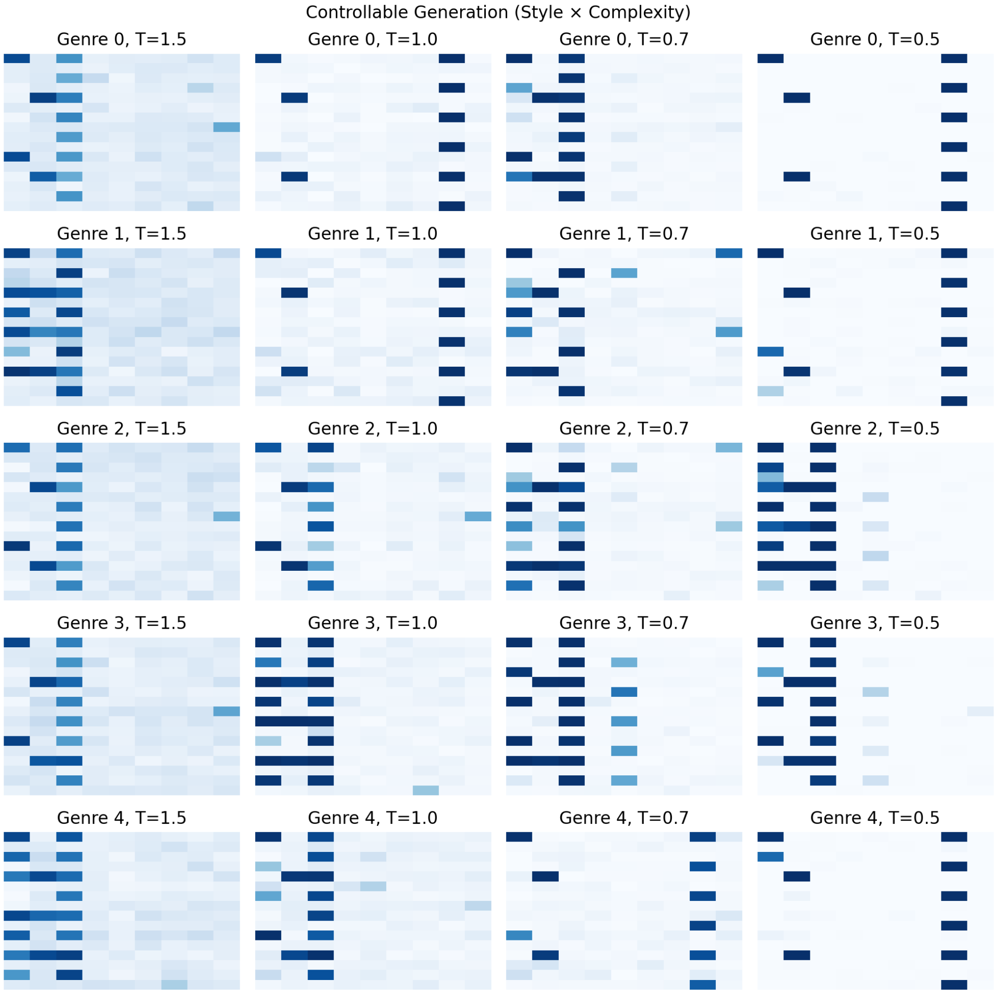
Evaluation was performed using t-SNE (for z_high), PCA (for z_low), latent interpolation, and controllable generation. All visualizations were saved to results/visualize/. These analyses reveal how the model organizes musical information in hierarchical latent spaces.

2.2 Result and analyse

The training logs demonstrate stable convergence with total loss ~33 and persistent non-zero KL values, indicating that latent variables remain active. KL annealing successfully prevented posterior collapse by gradually reintroducing the regularization term. The t-SNE visualization of z_high shows clustering by musical style, confirming that high-level latents encode genre identity. The PCA plot of z_low reveals continuous variation, implying that z_low captures fine-grained rhythmic details such as groove complexity.

2.3 Conclusion

The implemented HVAE effectively models drum patterns with controllable style and rhythmic variation. KL annealing and hierarchical encoding successfully mitigate posterior collapse and enable interpretable latent representations. The generated drum patterns demonstrate musical plausibility and stylistic control, confirming the HVAE’s potential as a controllable generative model for symbolic music.


Controllable Generation (Style × Complexity)
Genre 0, T=1.5
Genre 0, T=1.0
Genre 0, T=0.7
Genre 0, T=0.5
Genre 1, T=1.5
Genre 1, T=1.0
Genre 1, T=0.7
Genre 1, T=0.5
Genre 2, T=1.5
Genre 2, T=1.0
Genre 2, T=0.7
Genre 2, T=0.5
Genre 3, T=1.5
Genre 3, T=1.0
Genre 3, T=0.7
Genre 3, T=0.5
Genre 4, T=1.5
Genre 4, T=1.0
Genre 4, T=0.7
Genre 4, T=0.5