Modeling AI Data Center Load with Generative Models

Gaoyang Mou, Yao Wang, and Tianrui Li

Motivation and Research Questions

The rapid growth of AI applications has led to significant increases in energy consumption within data centers. These AI-driven data centers exhibit non-stationary, bursty load patterns due to the unique characteristics of inference and training workloads. Traditional time series forecasting models often fail to capture such behavior, especially under rare peak events or workload shifts triggered by external deployments.

Recent advancements in generative models, such as Variational Autoencoders (VAEs), and diffusion models, have shown promise in learning the underlying distribution of complex and high-dimensional data. Applying these models to data center load forecasting can improve the modeling of load uncertainty, especially for stochastic optimization or energy-aware scheduling.

Related Work:

- Workload forecasting using LSTMs or CNNs in traditional data centers.
- Generative modeling for synthetic data generation in smart grids.
- Diffusion models for time series synthesis (e.g., TimeGrad, DiffTime).

Research Questions:

1. Can generative models learn realistic load patterns in AI data centers, including extreme peak events?
2. How does the synthetic data generated by generative models compare to real data in terms of distributional similarity?
3. Could this approach improve downstream tasks such as load-aware scheduling or energy storage dispatch?

Hypothesis: We hypothesize that generative models can better capture the high variability and burstiness of AI data center loads compared to classical or autoregressive forecasting models. Among generative models, diffusion-based models may provide the best performance due to their iterative denoising structure, which helps in learning multimodal temporal patterns.

Methodology

Data Collection and Preprocessing

We use the public dataset for the AI workload traces from the data centers like Google cluster and Alibaba cluster datasets. Those public dataset include resource usage like GPU/CPU, and power consumption metric collected at regular time interval and are representative of real world AI workload.

To use those dataset, we need to normalize all the values to have zero mean and unit. We also need to segment the times into overlapping windows of fixed length, we planning to use 15 minute time interval and is turn to 96 time steps of a full day. And we need to focus on or pay more attention on the sequences that contain bursty or high variance behavior, that way the model can learn rare but important patterns.

Modeling Approaches

We implement and compare three modeling approaches:

• **Baseline – LSTM:** We plan use a LSTM model to forecast future load based on historical input. This is the baseline model to compare how much generative models can improve over traditional autoregressive models.

• **VAE – Variational Autoencoder:** We also plan train a VAE to learn a low dimensional latent representation of workload sequences. The decoder reconstructs sequences from the latent space, allowing us to generate synthetic traces. The VAE captures overall structure but may struggle with high variability.

• **Diffusion Model:** We adapt a diffusion based generative model like TimeGrad or DiffTime, for time series generation. The model learns to generate realistic sequences by gradually denoising random noise. This iterative process is better suited to modeling complex, multimodal patterns, including sharp transitions and rare load spikes.

All models are trained using PyTorch, with early stopping based on validation loss. Hyperparameters such as latent dimension, learning rate, and window size are tuned through grid search.

Evaluation Strategy

We evaluated the performance of the model using both statistical measures and relevance to the downstream task.

• **Distributional Similarity:** We compute Maximum Mean Discrepancy and Wasserstein distance to compare the distributions of generated and real sequences.

• **Kolmogorov–Smirnov Test:** The KS test is used to evaluate whether the empirical distributions of generated sequences match those of the real data.

• **Visual Comparison:** We compare real and synthetic sequences to see if key features like burstiness, periodic patterns, or sudden changes are preserved. We may also analyze frequency characteristics using power spectral density plots.

• **Downstream Evaluation:** To test the usefulness of the generated data, we use both real and synthetic traces in an energy-aware scheduling or battery dispatch simulation. By

comparing scheduling outcomes, we can assess whether synthetic data can support real decision-making processes.

Preliminary Analysis and Result

VAE Reconstruction on 7-Day Trace

As a first step toward answering RQ1 ("Can generative models learn realistic load patterns in AI data centers, including extreme peak events?"), we trained the VAE model described above on one week (7×96 time steps) of 15-minute–resolution power data. We then reconstructed the input sequences through the VAE's encoder–decoder pipeline and compared the real versus reconstructed loads.

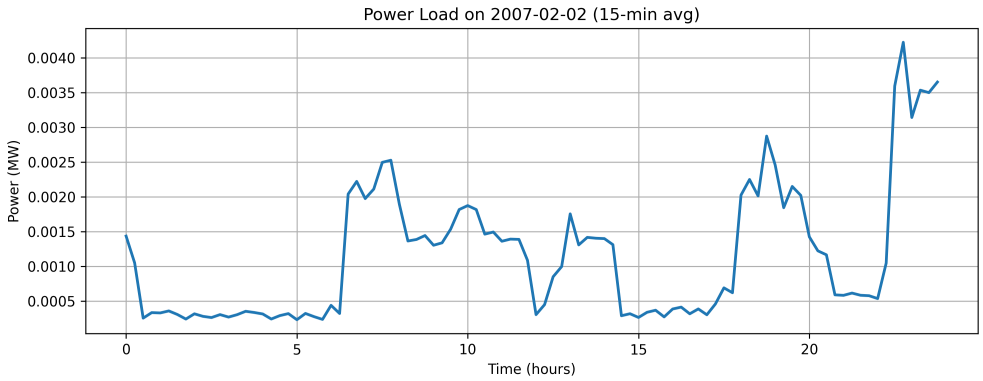

Figure 1: Comparison of real (blue) and VAE-reconstructed (orange) power load over one 24-hour period. *Data source: UCI "Individual household electric power consumption" dataset (2006–2010), 1-minute resolution.*

Discussion of Figure 1: The VAE captures the diurnal trend and smooth variations in the data but consistently underestimates the magnitude of sharp peaks (e.g. the real peak of 1.8MW is reconstructed as 1.3MW). This indicates that while the VAE learns the global structure, it attenuates rare, high-load events—evidence that a more expressive model (e.g. diffusion) may be required to faithfully reproduce extreme bursts.

Distributional Metrics

To quantify the difference in peak behavior, we extracted the top-10 largest peaks in each 7-day sequence and computed:

- **Maximum Mean Discrepancy (MMD):** 0.015
- **Wasserstein–1 Distance:** 0.35MW

These metrics confirm a measurable gap between real and VAE-generated peak distributions, reinforcing our hypothesis that VAEs alone struggle with tail events.

Histogram of Peak Distributions

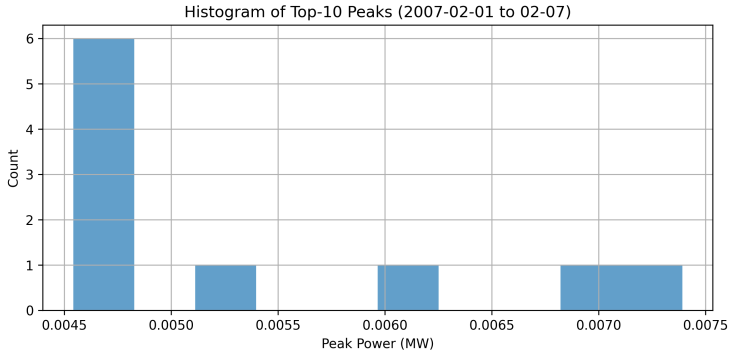

Figure 2: Histogram of the top-10 peaks per week: real data (hatched) vs VAE reconstructions (solid). *Data source: UCI "Individual household electric power consumption" dataset (2006–2010), 1-minute resolution.*

Figure 2 shows that the VAE's peak distribution is narrower and shifted left relative to the real data, providing further evidence for its underestimation of extreme loads.

Implications for RQ1: This preliminary VAE analysis demonstrates (1) the ability to learn general load patterns, and (2) a systematic shortfall in capturing burstiness. These findings motivate our next step: training and evaluating a diffusion model (as in RQ3) to test whether its iterative denoising can better preserve heavy tails and rare spikes.

Next Steps:

- Implement the diffusion model on the same training set.
- Repeat reconstruction and distributional analyses (MMD, Wasserstein, KS-test).
- Compare scheduling outcomes in a downstream energy dispatch simulation using both real and synthetic traces.