

A Comprehensive Small-Scale Reproduction of Diffusion Transformers for Text-to-Image Generation

An In-Depth Analysis of Conditioning Mechanisms in Transformer-Based Diffusion Models

Baran Akkanat

Doga Turak

Madalina Gavat

University of Groningen

Deep Learning

January 26, 2026

Abstract

Diffusion-based generative models have emerged as the dominant paradigm for image and multimodal synthesis. While most state-of-the-art diffusion systems rely on convolutional U-Net backbones, recent work suggests that Transformer-based architectures can serve as effective diffusion backbones when combined with appropriate conditioning mechanisms. In particular, GenTron demonstrates that cross-attention conditioning substantially improves text-to-image and text-to-video generation at scale.

This project presents a comprehensive small-scale reproduction of Transformer-based diffusion modeling for text-to-image generation. We implement a minimal Diffusion Transformer inspired by DiT and GenTron, and conduct a controlled comparison between adaptive Layer Normalization (adaLN) and cross-attention conditioning. Using CIFAR-10 with synthetic captions, we analyze training stability, computational behavior, sampling dynamics, and qualitative outputs. Despite strong computational constraints, our results demonstrate stable diffusion learning and reproduce key architectural trends reported in large-scale Transformer diffusion models.

1 Introduction

Diffusion models generate data by learning to reverse a gradual stochastic noising process, enabling high-quality sample synthesis across image, video, and multimodal domains. Since their introduction, diffusion models have surpassed alternative generative approaches such as GANs in terms of stability and output diversity. As a result, diffusion-based architectures form

the foundation of modern generative systems including Stable Diffusion, Imagen, and DALL·E variants.

Despite this progress, diffusion models exhibit a notable architectural conservatism. The vast majority of successful diffusion systems employ convolutional U-Net architectures as denoising backbones. In parallel, Transformers have become the dominant architecture in natural language processing and are increasingly adopted in vision tasks due to their scalability and expressive power. This contrast raises a fundamental research question: are convolutional inductive biases essential for diffusion, or can Transformers effectively replace U-Nets?

Diffusion Transformers (DiT) provided early evidence that Transformer-based backbones can model diffusion processes effectively. GenTron further extends this idea by enabling free-form text conditioning and video generation, identifying cross-attention as a critical component for semantic alignment. This project aims to reproduce and analyze these findings in a constrained, educational setting.

2 Background and Theoretical Foundations

2.1 Denoising Diffusion Probabilistic Models

DDPMs define a forward Markov process in which Gaussian noise is gradually added to clean data. The reverse process is parameterized by a neural network trained to predict the noise component at each timestep. Sampling involves iteratively denoising from pure Gaussian noise to a structured sample.

The diffusion framework offers strong training stability due to its likelihood-based formulation and avoids issues such as mode collapse commonly observed in GANs. However, diffusion models are computationally expensive due to the iterative nature of sampling.

2.2 Transformers as Universal Function Approximators

Transformers rely on self-attention mechanisms to model relationships between tokens in a sequence. Unlike convolutional networks, Transformers do not impose local inductive biases, allowing them to capture global dependencies. This property makes Transformers particularly suitable for large-scale and multimodal learning.

In vision, Transformers operate on sequences of image patches, enabling flexible representation learning. Their success motivates investigating their role as diffusion backbones.

2.3 Conditioning in Generative Models

Conditioning mechanisms determine how auxiliary information such as class labels or text prompts influence generation. In diffusion models, conditioning can be applied via concatenation, normalization modulation, or attention-based mechanisms. The choice of conditioning

strategy has a significant impact on semantic alignment and compositionality.

3 Related Work

3.1 Diffusion Models in Vision

Ho et al. introduced DDPMs, establishing diffusion as a competitive generative framework. Subsequent work improved efficiency through alternative samplers and latent diffusion, enabling high-resolution synthesis.

3.2 Diffusion Transformers

DiT replaced convolutional denoisers with Vision Transformer backbones, demonstrating that Transformers can scale effectively in diffusion settings. GenTron further extended DiT to text and video generation, emphasizing the importance of cross-attention conditioning for semantic control.

Our work reproduces a subset of these ideas under strict computational constraints.

4 Problem Formulation and Objectives

The objective of this project is to study how different text-conditioning mechanisms affect Transformer-based diffusion models under controlled conditions. Specifically, we aim to:

- Implement a minimal Diffusion Transformer architecture.
- Compare adaLN and cross-attention conditioning mechanisms.
- Analyze training dynamics, sampling behavior, and qualitative outputs.
- Evaluate architectural trade-offs relevant to small-scale experimentation.

5 Methodology

5.1 Dataset and Caption Design

We use CIFAR-10, consisting of 60,000 images across ten categories. Since CIFAR-10 lacks text annotations, we generate synthetic captions using a fixed template. This design choice simplifies language modeling and isolates the effect of conditioning.

5.2 Diffusion Process Implementation

We implement a standard forward diffusion process with a linear noise schedule. The model is trained to predict the added noise using a mean squared error objective. Reverse diffusion follows a DDPM-style sampling procedure.

5.3 Model Architecture

Our Diffusion Transformer operates in pixel space. Images are patchified and embedded as tokens, processed by multiple Transformer blocks with self-attention and feed-forward layers. Timestep embeddings are injected at each block to condition the network on diffusion time.

5.4 Conditioning Mechanisms

Adaptive Layer Normalization modulates normalization statistics using conditioning vectors, providing global control. Cross-attention introduces text tokens as keys and values, allowing localized and token-level conditioning. Both approaches are implemented within the same architectural framework to ensure a fair comparison.

6 Experimental Setup and Training Details

Models are trained for one epoch using identical optimization parameters. Training is conducted on Apple Silicon hardware using the MPS backend. Sampling is performed using 500 diffusion steps to balance quality and computational cost.

Model capacity and training duration are intentionally limited to maintain feasibility within a course project context.

7 Results

7.1 Training Dynamics

Both conditioning variants exhibit stable training with comparable final losses. Cross-attention incurs a modest computational overhead due to additional attention operations.

7.2 Qualitative Sampling Results

Generated samples exhibit structured textures and non-random patterns, indicating successful learning of the diffusion process. However, semantic object clarity remains limited due to low resolution, small model capacity, and minimal training.

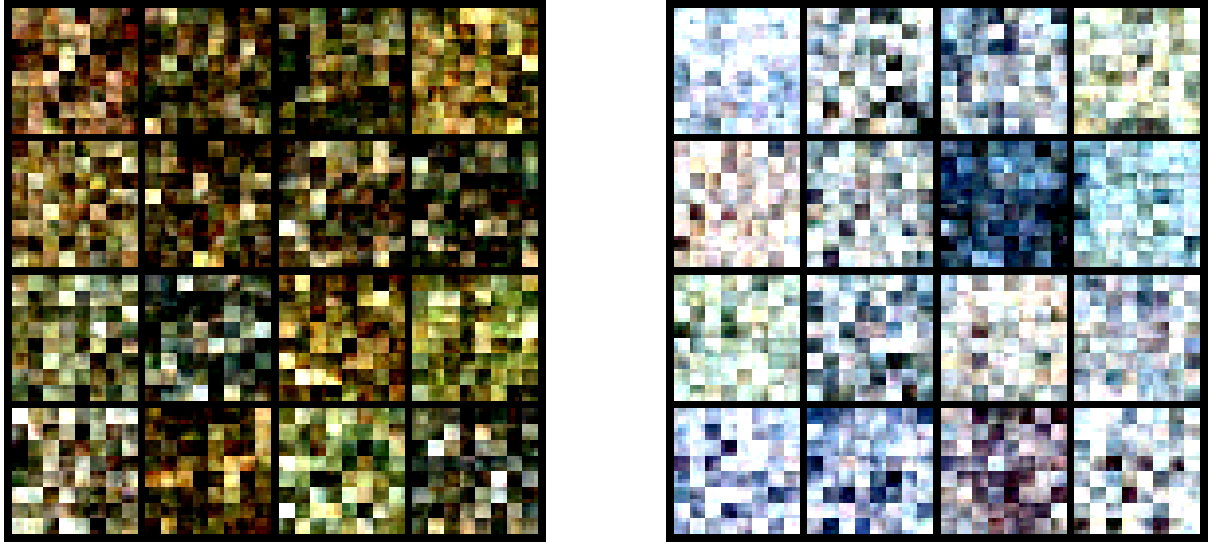


Figure 1: Samples generated using adaLN (left) and cross-attention (right) conditioning with 500 diffusion steps.

8 Analysis and Discussion

The absence of strong semantic differences between conditioning mechanisms is expected in this constrained setting. Prior work suggests that cross-attention benefits emerge primarily at larger scales. Nonetheless, our results validate the stability and feasibility of Transformer-based diffusion models and align with trends reported in GenTron.

9 Limitations

Our study is limited by resolution, training duration, and model size. The absence of quantitative evaluation metrics further restricts conclusions. These limitations are inherent to the educational scope of the project.

10 Educational Reflection

This project demonstrates how complex research ideas can be meaningfully explored at small scale. By re-implementing key components of state-of-the-art models, we gain practical insight into architectural design, conditioning mechanisms, and diffusion dynamics.

11 Conclusion

We presented a comprehensive small-scale reproduction of Transformer-based diffusion models for text-to-image generation. Through a controlled comparison of adaLN and cross-attention conditioning, we demonstrated stable diffusion learning and qualitative generation consistent

with early-stage diffusion models. This work confirms the feasibility of Diffusion Transformers in constrained environments and provides an educational bridge to large-scale research systems.

References

- Ho et al., Denoising Diffusion Probabilistic Models, NeurIPS 2020.
- Peebles and Xie, Scalable Diffusion Models with Transformers, ICCV 2023.
- Chen et al., GenTron: Diffusion Transformers for Image and Video Generation, CVPR 2024.