Project II Report

Team members: Yena Kim , Krish Jhaveri , Tianna Calderon , Christopher Castrence , Adrian Balingit

Problem Statement/Motivation

Multimodal generative modeling refers to machine learning systems that jointly process multiple data modalities, such as text and images. In text-to-image generation, these models learn to synthesize images that are semantically aligned with natural language descriptions by conditioning a visual generative process on textual representations. This paradigm enables images to be generated directly from user-provided prompts and forms the foundation of modern text-to-image models.

In this project, we implement a Stable Diffusion–style text-to-image model using a rectified flow formulation. Images are generated in a latent space for computational efficiency, and text prompts are encoded using a pretrained text encoder and incorporated through a multimodal transformer backbone. Rectified flow defines a deterministic probability path between a data distribution p_0 and a noise distribution p_1 via an ordinary differential equation,

$$\frac{dy_t}{dt} = v(y_t, t)$$

where $v(y_t, t)$ is a learned velocity field. Image generation is performed by numerically integrating this equation from noise to data, simplifying sampling while providing an efficient alternative to stochastic diffusion models under constrained computational resources.

Related Work

Models like DALL-E and Stable Diffusion, which convert natural-language prompts into visuals, are examples of earlier work in multimodal text-to-image conversion. While Stable Diffusion introduces latent diffusion, which compresses images into a smaller latent space to improve training efficiency, DALL-E employs a transformer-based method to translate text directly to pixel outputs. More recent studies investigate rectified flow, a technique that uses a deterministic ODE to move samples along straight routes between noise and data in place of the stochastic denoising process of diffusion. This maintains high-class image quality while streamlining training and increasing computing efficiency.

Multimodal Diffusion Transformers (MM-DiT) and SNR-shaped timestep sampling are two recent works that improve the selection of noise levels during training and the interaction between text and image representations in rectified-flow models. Metrics like Fréchet Inception Distance (FID) are improved and prompts and generated images are better aligned because of these improvements. Our idea, however, makes use of a simpler multimodal transformer architecture and the rectified-flow (RF) formulation using the default logit-normal timestep sampling.. Despite being more constrained than large-scale systems, this configuration retains the fundamental physics of rectified flow and enables us to examine its behavior under practical computational limitations.

Methodologies

Our work follows the rectified flow (RF) formulation, which the Esser et al. (2024) paper recasts as a noise-prediction loss with an explicit Signal-to-Noise Ratio (SNR) structure, allowing a principled analysis of optimal training timesteps.

Caption: "A man with his skate board posing for a picture."

We used the standard RF setup but implemented the Logit-Normal timestep sampling distribution (rf/lognorm), which is an SNR-shaped sampling method proposed by the paper to focus on challenging intermediate timesteps. However, we used the default, un-tuned hyperparameters (rf/lognorm(0.00,1.00)). This choice proved ineffective for our simplified architecture and, due to a lack of hyperparameter optimization, loosened the connection between loss and image quality. The paper demonstrates that tuning these parameters is crucial for improving metrics like FID and CLIP.

Architecturally, we used a simplified MM-DiT backbone that did not fully separate modality-specific parameters or include stabilizing methods like QK-normalization, reflecting our compute constraints. This streamlining likely hampered text-image alignment.

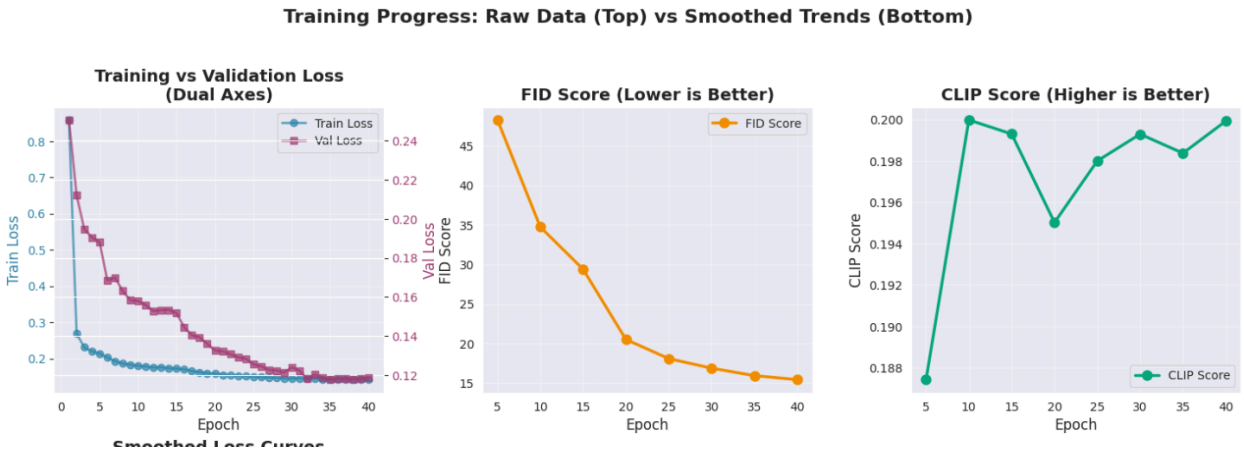
Differences in data handling, such as using on-the-fly encoding and only original MS COCO captions, further contributed to instability and made direct comparison to the paper's large-scale, preprocessed results difficult. These choices collectively impacted our interpretation of validation loss, which, despite decreasing, had a weaker correlation with final visual quality compared to the paper's highly refined methodology.

Evaluation

We trained and evaluated our rectified flow text-to-image model on a random subset of 10,000 MS-COCO 2017 image–caption pairs, which we split into 9,000 for training and 1,000 for validation. Due to compute constraints, images were resized and center-cropped to 128×128,

converted to tensors, and normalized to [−1,1]. At each training step, we sampled one caption from the available COCO captions for that image and used its CLIP text embedding for conditioning.

To obtain quantitative, objective measurements, we tracked three metrics throughout the training: validation MSE loss on the rectified-flow velocity-prediction objective; Fréchet Inception Distance (FID) for image quality; and a CLIP score that measures prompt-image semantic similarity. Evaluation was performed every 5 epochs on up to 500 validation samples to keep the evaluation computationally reasonable. Training was stable and convergent: train loss decreased from approximately 0.86 to roughly 0.15, while validation loss decreased from approximately 0.25 to roughly 0.12 by epoch 40. Image quality improved substantially, with FID dropping from approximately 48 at Epoch 5 to ~15 or 16. Prompt alignment improved more modestly, with CLIP score increasing from roughly 0.187 at Epoch 5 to ~0.200 at Epoch 40. The smoothed curves show clear trends that additional training improves both visual quality and text alignment.



We chose our model and training settings (e.g., batch size 16) to balance reproducibility with practical resource limits, which prevents direct score comparisons with large-scale systems. Our final metrics, achieved at Epoch 40, were an FID of 15.42 and a CLIP score of 0.1999.

Compared to the paper's benchmark using the identical rf/lognorm(0.00,1.00) sampling strategy, our FID is significantly lower, which is likely due to our limited, 128×128 evaluation set being easier to fit. Conversely, our CLIP score is substantially lower, an expected quantitative result of our simplified MM-DiT architecture and on-the-fly encoding. Ultimately, our results support the paper's findings: validation loss decreases while perceptual metrics improve, providing quantitative evidence that our constrained implementation of rectified flow behaves as expected.

Conclusions

Our comprehension of rectified-flow-based text-to-image models' operation and their sensitivity to design decisions like timestep sampling, loss weighting, and multimodal architecture structure has improved as a result of this effort. We were able to run and evaluate the model with restricted hardware resources because of our reduced Stable Diffusion implementation, which matched the fundamental behavior of rectified flow and demonstrated its computational efficiency.

However, the experiment also exposed a number of flaws: on-the-fly encoding produced instability in validation loss, the default logit-normal timestep sampling parameters decreased training effectiveness, and the streamlined multimodal architecture hampered text–image alignment. In order to obtain more dependable training dynamics and higher-quality image creation in subsequent work, these findings assisted us in identifying particular areas for improvement, such as optimizing the parameters of the Logit-Normal distribution (or implementing a different SNR-shaped distribution), separating modality-specific parameters, and applying more robust preprocessing.

Description of individual effort

Yena designed and implemented the Simplified MM-DiT architecture, focusing on the transformer backbone and multimodal token processing as well as wrote the Conclusion. Krish was responsible for translating the theoretical framework into code, specifically implementing the Rectified Flow velocity-prediction objective and integrating the Logit-Normal timestep sampling mechanism as well as wrote the Related Work section. Tianna managed all data ingestion and preprocessing tasks, including preparing the 10k MS-COCO subset, image transformations (resizing, normalization), and integrating the CLIP text encoder and wrote the Problem Statement/Motivation section. Adrian managed the training environment and execution, setting up the main training and validation loops, optimizing memory usage, and conducting initial, unsuccessful attempts at hyperparameter tuning for the Logit-Normal distribution as well as wrote the Evaluation section. Christopher set up the pipeline for computing the FID and CLIP scores, generated the final metric visualizations, and ensured overall code consistency as well as wrote the Methodologies section.

References

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Lacey, K., Goodwin, A., Marek, Y., & Rombach, R. (2024). *Scaling rectified flow transformers for high-resolution image synthesis*. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*. arXiv. https://arxiv.org/abs/2403.03206

Multimodality - Wikipedia

Multimodality. (n.d.). *Wikipedia*. Retrieved December 15, 2025, from https://en.wikipedia.org/wiki/Multimodality Wikipedia