Project 2 Report

Thalis Goldschmidt, Blanka Visy and Paul Wollenhaupt

thalisg@uio.no, blankav@uio.no, paul.wollenhaupt@uio.no

27.09.2024

Abstract

Image generation is a critical area within computer vision and machine learning. In this work, we apply generative AI to support the mission of PUSHPOP INC., focusing on generating car images that appeal to a specific member of a focus group. We utilize a Rectified Flow model and a Variational Autoencoder (VAE) in two configurations: one with a Gaussian Mixture Model (GMM) to enhance sampling diversity, and the other with Rectified Flow in the latent space. Our approach follows three main steps: (1) training the image generation models, (2) training a regressor conditioned on Moira's preferences, and (3) selecting the top images based on regression scores. Although the generated images fell short of our expectations, likely due to the limited sample size, all three methods produced images resembling cars. However, in each case, the images were quite blurry and lacked fine detail. In terms of diversity, image quality, and estimated preference scores, the VAE with GMM outperformed the other models. Our findings suggest that while Rectified Flows can deliver high-quality results on larger datasets, there are situations—such as this one—where a VAE with GMM can yield significantly better outcomes.

1 Introduction

Image generation, a key branch of computer vision and machine learning, has evolved rapidly with advancements in deep learning. It involves generating new images from scratch or based on a given set of conditions. Image generation has various applications, from creating high-quality art [Elgammal et al., 2017], enhancing data augmentation in machine learning [Perez & Wang, 2017], to practical uses in healthcare (e.g., generating medical images) [Frid-Adar et al., 2018] and design (e.g., product or architectural visualizations) [Krahe et al., 2020].

In this project, we aim to apply generative AI to support the mission of PUSHPOP INC. Specifically, **our goal is to generate 10 images of cars that are highly appealing to one focus group member, Moira**. The dataset provided for training consists of 6,000 RGB images of various cars, each with a resolution of 96x128 pixels, alongside ratings from all members of the focus group, including Moira.

Motivated by the suggestions of our coworkers, we first reviewed existing image generation approaches. Several methods have been developed to address image generation challenges, each with unique strengths. Early techniques include autoregressive models [Graves, 2014; Oord et al., 2016], energy-based models [LeCun et al., 2006; Ngiam et al., 2011], and Variational Autoencoders (VAEs) [Kingma & Welling, 2022]. VAEs map input data to a latent space and reconstruct it, enabling the generation of variations of existing images, though often producing blurry outputs. Another approach includes flow-based models like NICE [Dinh et al., 2015] and GANs [Goodfellow et al., 2014], that dominated early image generation due to their efficiency and flexibility. Later on, diffusion models [Sohl-Dickstein et al., 2015; Ho et al., 2020; Y. Song et al., 2021] were introduced, outperforming sooner approaches. Even advancing the advantages of these models, Rectified Flows [Liu et al., 2022] balance between sample quality and computational cost, bridging the gap between one-step models like GANs/VAEs and continuous-time models like diffusion models.

After reviewing the available options, we chose to experiment with three different approaches. First, we implemented Rectified Flow in ambient space using a U-Net backbone due to its promising performance, as well as its resistance to the training instability often encountered with GANs. However, given the small size of our dataset, we also decided to employ Variational Autoencoders (VAEs), which offer higher sample efficiency compared to other methods. We used VAEs in two configurations: one with a Gaussian Mixture Model (GMM) to improve sampling diversity, and one with Rectified Flow in the latent space.

Our hypothesis was that Rectified Flow would likely outperform VAEs, primarily due to the latter's tendency to produce blurry images.

To tailor the results to Moira's specific preferences, we implemented an alternative conditioning method. Instead of relying on sampling-based approaches, we trained a regressor on the dataset to predict Moira's ratings for car images. After generating a large set of images, we used the regressor to estimate Moira's scores and selected the top 10 images with the highest predicted ratings. This approach is both sample-efficient and model-agnostic, as detailed further in Chapter 3.

Our results indicate that the Rectified Flow model in ambient space struggled to produce high-quality, diverse images. While the generated images resembled cars, they failed to capture finer details accurately. Moreover, after applying the regressor, the top 20 images appeared overly similar, suggesting a lack of diversity in the generated samples. In contrast, the VAE-based models performed better in terms of generating more varied cars. However, the Rectified Flow in the latent space also struggled to represent detailed cars, likely due to the small training dataset, which was insufficient for capturing higher-quality details in the reduced latent space.

Among the models, the VAE with GMM produced the best results visually and in terms of diversity. While the regressor seemed to overfit to colorful images, the selected car images, though somewhat blurry, were generally feasible representations.

The final submitted dataset of 10 cars had estimated Moira scores ranging from 3.91 to 4.28.

In this report, Chapter 2 provides a review of related work, Chapter 3 details our methodology, and Chapter 4 presents the experimental setup and results. Chapter 5 discusses how our findings address the research question, and Chapter 6 concludes with a summary of our approach and key takeaways.

2 Related Work

Several approaches have been developed to solve the challenges of image generation, each with unique methodologies and strengths. In this chapter, some of the major directions in image generation are introduced. As this work has a very specific research goal that is image synthesis, we focus on approaches in that direction with an emphasis on approaches that have been provided as a recommendation from the coworkers.

2.1 Variational Autoencoders

One of the first successful image generation methods was the Variational Autoencoders (VAE) [Kingma & Welling, 2022]. A VAE consists of an encoder that maps input data (such as images) to a latent space and a decoder that reconstructs the image from a sample in this space. Unlike traditional autoencoders, VAEs enforce a probabilistic constraint, where the latent space is regularized to follow a standard normal distribution. This allows for meaningful interpolations between points in the latent space, which makes VAEs particularly powerful for generating variations of existing images. VAEs are able to generate high-quality samples that are similar to the training data. However, because of the probabilistic reconstruction process that is used, the output images tend to be blurry. In subsequent research, Asperti et al. [2021] propose that state-of-the-art generative performance can be achieved—and potentially enhanced—through carefully designed VAE architectures that optimize computational and sample efficiency.

Several studies have focused on the advancement of Variational Autoencoders (VAEs). Sohn et al. [2015] extends the standard VAE to incorporate conditional information, making the model more effective for generation tasks that require specific conditions, such as image restoration and style transfer. This adaptation significantly improves performance in these areas. One of the first text-to-image models, alignDraw, also builds on VAEs [Mansimov et al., 2016]. They employ a modified recurrent VAE with an attention mechanism, enabling the model to generate images based on textual descriptions, even when the input sentences were not part of the training data.

Oord et al. [2018] argue that in many domains, including image generation, discrete representations are more natural than continuous ones. In their work, they introduce a new class of generative models that combine VAEs with vector quantization (VQ). This approach aims to retain the strengths of VAEs while addressing key challenges like "posterior collapse" and variance issues. The core innovation is the discretization of the continuous latent space using a set of discrete "codebook vectors". The encoder maps the input to its nearest codebook vector, while the decoder transforms this discrete vector back into a continuous latent representation. Razavi et al. [2019] demonstrates that VQ-VAEs are capable of generating high-quality images with a strong level of diversity.

2.2 Normalising Flows

Another direction of likelihood-based generative models is flow-based models. The work by Dinh et al. [2015] introduces Non-Linear Independent Components Estimation (NICE), one of the earlier flow-based models used to represent data distributions. Unlike probabilistic systems such as diffusion models, NICE focuses on deterministic transformations and uses a discretized-time version to avoid the large sampling process. The model is based on invertible transformations, which support both forward (data generation) and backward (density estimation) processes. This framework enables exact likelihood computation and efficient sampling.

However, the necessary invertibility of the transformations poses challenges because it preserves volume, particularly with scalability in high-dimensional spaces, such as those of image data. Consequently, implementing NICE demonstrates less flexibility regarding the tasks it can handle. The idea of flow-based models was further elaborated in the work of Dinh et al. [2017], and also further work expanded the possibilities by applying a simple convolutional architechture [Kingma & Dhariwal, 2018],

2.3 Generative Adversarial Networks

An alternative approach to generative modeling is based on game theory. Goodfellow et al. [2014] introduced an innovative method named Generative Adversarial Nets (GANs) for estimating generative models through the concurrent training of two models: a Generator and a Discriminator. The Generator synthesizes images, while the Discriminator assesses them. The advantages of GANs include the absence of Markov chains, their reliance solely on backpropagation for gradient calculation, and the lack of inference during learning. Additionally, the model design allows for any differentiable function. Due to these benefits, GANs achieved significant success and dominated the field of image generation in its early stages [Li et al., 2024]. However, there are also some disadvantages. Maintaining the balance between the Generator and Discriminator is crucial, yet it is often challenging to achieve in practical scenarios. Moreover, GANs can encounter a phenomenon known as "mode collapse," resulting in generated images that are overly similar and lack diversity [Li et al., 2024].

Numerous advancements have been made in the development of GANs. Mirza & Osindero [2014] propose Conditioned GANs (CGANs), which integrate supplementary information during the training phase to guide the image generation process. This conditioning is achieved by incorporating auxiliary data as an additional input layer to both the Generator and the Discriminator, with the efficacy of this approach being demonstrated in their publication. Karras et al. [2018] present Progressive-GAN as a solution to instability issues and to speed up training. Their method incrementally increases the resolution of both the Generator and Discriminator. Radford et al. [2016] introduced DCGAN. They employ a convolutional architecture to replace fully connected layers in both the Generator and Discriminator of GANs, along with incorporating batch normalization in both components. This change elevates stability and often also demonstrates better performance than GANs.

2.4 Diffusion Models and Flow Matching

An additional alternative research direction in the domain of image generation involves diffusion models [Sohl-Dickstein et al., 2015; Ho et al., 2020; Y. Song et al., 2021; J. Song et al., 2022]. They produce data by simulating a stochastic process that gradually converts noise into meaningful data.

One of the most cited models is the Denoising Diffusion Probabilistic Model (DDPM) by Ho et al. [2020]. This model is constructed as a parametric Markov chain that produces images by progressively eliminating noise from the image through two distinct processes: a forward and a reverse process. The primary concept supporting this model is the simulation of diffusion and restoration mechanisms observed in physical systems to synthesize high-quality data samples. The forward process represents the diffusion aspect, systematically introducing noise to an image until it reaches the state of random noise. In contrast, the reverse process functions as the denoising generator mechanism, transforming the random noise input into a coherent image by sequentially removing noise. Following the advent of DDPMs, they have outperformed GANs in producing high-quality images [Dhariwal & Nichol, 2021]. However, a significant limitation of DDPMs is their slow sample generation: this is because drawing samples requires numerical integration of probability flow ordinary Ordinary Differential Equations (PF-ODEs).

Another notable recent approach is Rectified Flows [Liu et al., 2022], which focuses on a deterministic setting and uses ordinary differential equation (ODE) models to transport one distribution to another. Unlike other diffusion models that rely on more mathematically complex stochastic differential equation (SDE) models, Rectified Flows avoids this complexity both conceptually and algorithmically. The model can be applied to both generative modeling, where points from a simple distribution are transformed into an unknown empirical distribution, and transfer modeling, where both distributions are unknown and empirical. The core idea of Rectified Flows is to constrain the transport map to linear paths. This is advantageous for two reasons: theoretically, it serves as an essential component in achieving optimal transport, and computationally, it allows for exact simulation of ODEs along straight paths without the need for time discretization. Additionally, straight flows offer a principled method for learning ODEs with fast inference, enabling efficient training of one-step models with ODEs serving as intermediate steps. Rectified Flows are also more flexible than traditional diffusion models in the transformation of arbitrary distributions, which can pose challenges for other methods [Cao et al., 2023]. Moreover, rectified flows can flexibly adjust the balance between sample quality and computational cost by altering the number of function evaluations.

This approach effectively bridges the gap between one-step models, such as GANs or VAEs, and continuous-time-based models, such as diffusion models, combining the lower computational costs and fast inference of the former with the superior performance of the latter. Furthermore, the authors not only provide a theoretical foundation for the advantages of this method but also empirically demonstrate its outstanding performance in tasks like image-to-image translation and image generation.

In the following, we summarize the advantages in Table 1 and the disadvantages in Table 2 of the aforementioned methods.

Table 1: Advantages of Different Approaches

Approach	**Pros**
GANs	- One step generation - High fidelity
VAEs	- Computational and sample efficiency - Compact representation not just sampling
Normalized Flows	- Invertible transformations - Exact likelihood estimation
Diffusion Models	- High-quality image generation
Rectified Flows	- High-quality output images - Arbitrary input distribution - Invertible transformation

Table 2: Disadvantages of Different Approaches

Approach	**Cons**
GANs	- Training instabilities - Low mode coverage
VAEs	- Produces blurry output images - Unspecified latent sampling
Normalized Flows	- Memory requirement - low sample fidelity
Diffusion Models	- Large inference time
Rectified Flows	- Medium inference time - Novel/less established

3 Method

Given the previously summarized advantages and disadvantages, we experimented with three main image generation approaches, all conditioned using a separate regressor trained on Moira's scores. Our decision-making process focused on the following factors:

- GANs, while powerful, often suffer from instability due to the imbalance between the generator and discriminator. This challenge is mitigated by Rectified Flows.
- Diffusion models represent the state of the art in image generation, though they tend to have slow inference times during sampling.
- VAEs are known for producing blurry images, even in more advanced variants. However, they exhibit high sample efficiency, which is particularly important given the small size of our training set.
- Rectified Flows—although slower compared to one-step models—offer a strong balance between quality and flexibility, making them a promising choice for this project. They also achieve state-of-the-art results. In our case, since we are not generating an exceptionally large number of images, the inference time factor that could be a drawback, has not too high of an importance.

Based on these considerations, our initial model choices were:

1. **Rectified Flow with U-Net Architecture in Ambient Space**: We first explored using a Rectified Flow directly on the image space, with a U-Net backbone for better structure.
2. **VAE with GMM in Latent Space**: To address challenges related to the small training set, we also employed a Variational Autoencoder (VAE), using a Gaussian Mixture Model (GMM) in the latent space to better capture complex distributions and generate diverse samples.
3. **VAE with Rectified Flow in Latent Space**: Lastly, we tested a Rectified Flow applied to the VAE's latent space, aiming to enhance sampling fidelity and improve the quality of the generated images.

For all models, conditioning was managed by a separate regressor trained on the original car dataset with Moira's scores. This regressor evaluated and ranked the generated images, allowing us to select the top matches without altering the generative models. This approach was chosen for its model-agnostic and flexible nature, enabling the generation of a wide variety of images independent of Moira's preferences. By separating the evaluation process, the workflow remained streamlined and adaptable, allowing for quick adjustments to changing preferences without the need to retrain the generative models, thereby enhancing overall efficiency and scalability. Additionally, sample-based conditioning techniques are more sensitive to sample size. Therefore, given our limited training dataset, we opted for regression for this reason as well.

3.1 Rectified Flow with U-Net Architecture

Rectified Flow, developed by [Liu et al., 2022], is a type of a continuous normalizing flow designed to implicitly learn a transport map T between two distributions, π_0 (the initial distribution) and π_1 (the target distribution), using an Ordinary Differential Equation (ODE). This method constructs an ODE defined by a drift force ν, guiding the transformation of samples from π_0 to π_1.

The ODE describing this transformation is given by:

$$dZ_t = \nu(Z_t, t)\, dt, \quad t \in [0, 1], \quad \text{starting from } Z_0 \sim \pi_0,$$

where:

- Z_t is the state of the sample at time t.
- $\nu(Z_t, t)$ is the drift force that directs how samples evolve over time from the initial state at $t = 0$ to the final state at $t = 1$.

The objective is for the ODE, when starting from a sample Z_0 drawn from π_0, to evolve such that the solution at $t = 1$ (denoted Z_1) matches the target distribution π_1. The main challenge is learning the drift force ν from data to effectively guide the transformation between these distributions.

A significant difficulty lies in inferring the correct intermediate paths between π_0 and π_1. Without explicit constraints, the trajectories within the ODE can be overly flexible, resulting in inefficiencies during training. To mitigate this, the concept of straight trajectories is introduced: favoring direct paths between the start and end distributions simplifies the model, reduces computational overhead, and accelerates inference.

These straight paths, inspired by optimal transport theory, streamline the model by minimizing discretization errors and enhancing the inference speed. This approach enables Rectified Flows to merge the theoretical strengths of ODE-based models with the rapid inference capabilities typical of simpler generative models like GANs and VAEs [Liu et al., 2022].

The U-Net architecture plays a critical role in our implementation of Rectified Flows, acting as the backbone that parameterizes the drift force $\nu(Z_t, t)$ within the ODE framework. Originally introduced by Ronneberger et al. [2015], the U-Net is a convolutional neural network (CNN) known for its symmetric encoder-decoder structure, designed to handle image-to-image tasks, such as segmentation and generation.

The U-Net architecture employs an encoder-decoder structure, where the encoder progressively down-samples the input to capture spatial features, and the decoder up-samples these features to reconstruct the image. A key innovation of U-Net is its use of long skip connections (LSCs), which directly transfer high-resolution information from encoder layers to their corresponding decoder layers. This design helps preserve critical details that might otherwise be lost during down-sampling, making U-Net particularly effective for complex image generation tasks.

Due to these advantageous properties, especially LSCs that connect distant, symmetrical blocks within the network, U-Nets have become the preferred architecture for diffusion models [Huang et al., 2023]. These connections ensure that high-resolution details are maintained throughout the generation process, resulting in superior performance in tasks such as image denoising and restoration [Huang et al., 2023]. Recognizing these strengths, we selected U-Net as the backbone for implementing the Rectified Flows approach.

The time information t is Fourier encoded, fed through a linear layer, and then added to the encoder activations, allowing the network to effectively model the drift force ν across the transformation process.

3.2 Variational Autoencoders

Variational Autoencoders (VAEs), introduced by Kingma & Welling [2022], are generative models that, like U-Nets, use an encoder-decoder structure, but they fundamentally differ in how they generate images. While U-Nets are primarily designed for direct image-to-image tasks, VAEs model data distributions through a probabilistic framework, allowing them to generate new images from latent representations.

A VAE consists of an encoder that compresses input images into a latent space and a decoder that reconstructs images from these latent representations. The key distinction of VAEs lies in their probabilistic approach: the latent space is regularized to follow a standard normal distribution. This regularization enables the model to generate coherent and diverse outputs by sampling from this space and ensures smooth transitions between points in the latent space, making VAEs particularly powerful for generating variations of existing images.

3.2.1 Gaussian Mixture Models (GMMs) in VAE Latent Space

To improve sampling from the VAE's latent space, we utilized a Gaussian Mixture Model (GMM). GMMs capture complex multimodal structures by fitting a mixture of Gaussian distributions, providing a more expressive latent representation compared to direct sampling from the mean and variance of the VAE. This approach helps generate diverse and high-quality samples, making it an effective method for navigating the latent space. GMMs use the Expectation Maximization algorithm to fit a mixture of Gaussians to a distribution and are typically used in low-dimensional settings, making them ideal for use within the compact latent space of a VAE.

3.2.2 Rectified Flow in VAE Latent Space

Rectified Flows can also benefit from the compact representation of the VAE latent space, which is why combinations of Rectified Flows and VAEs are employed in state-of-the-art large-scale image generation models like StableDiffusion 3 and the FLUX model family [Esser et al., 2024]. Therefore, we decided to explore Rectified Flow applied within the VAE's latent space, combining the compact representation of the VAE with the high-fidelity sampling capabilities of Rectified Flows.

4 Experiments and Results

4.1 Overall Workflow

Our approach consists of three main steps: (1) training image generation models, (2) training a regressor conditioned on Moira's preferences, and (3) selecting top images based on regressor scores. The image generation was conducted using three different methods: Rectified Flow with U-Net in ambient space, Variational Autoencoders (VAEs) with Gaussian Mixture Models (GMMs) in the latent space, and VAEs with Rectified Flow in the latent space. In parallel, the regressor was trained on the original car dataset to learn Moira's ratings and later used to score the generated images.

4.2 Architecture Overview

For both the U-Net and VAE, we employed convolutions with a kernel size of 3. Each resolution step comprised two convolutional layers, with one convolution using a stride of two for downsampling (or a transposed convolution with stride 2 for upsampling). The number of channels doubled with each downsampling step. We interleaved GELU activations and layer normalization throughout the convolutional layers to enhance feature extraction and stability during training.

4.3 Implementation and Results

4.3.1 Rectified Flow with U-Net Architecture

The Rectified Flow model was implemented using a U-Net backbone to parameterize the drift force $\nu(Z_t, t)$ within the ODE framework. This setup was initially validated on 16x16 pixel MNIST images [Deng, 2012], allowing us to establish the feasibility of the approach with low computational cost. We then applied the model to the higher-resolution car dataset, which presented a significantly more complex challenge due to the dataset's higher resolution and smaller size.

Despite initial difficulties, the resulting images, shown in Figure 1, demonstrate that the model successfully generates recognizable car-like structures. Distinct automotive features such as grilles, tires, and hoods are clearly visible in most samples, illustrating that the model has effectively captured essential elements of the dataset. The images also show a high degree of diversity, with variations in perspectives and color schemes, showcasing the model's ability to generate a broad range of car images with different visual aspects.

However, some images appear distorted, indicating that while the model has learned the general distribution of car features, it struggles with accurately rendering finer details. This highlights the limitations of the approach in achieving high-fidelity outputs consistently, particularly when generating complex or intricate elements of the car images.

Figure 1: Generated samples using Rectified Flow in the ambient space.

4.3.2 Variational Autoencoders (VAEs) with Gaussian Mixture Models (GMMs)

The VAE was trained with a latent dimension of 64 and a KL-divergence weight $\beta = 0.1$, facilitating a compact representation of the input data. The training process was optimized through VRAM caching and Torch's just-in-time compilation, completing 1000 epochs.

Upon examining the generated images in Figure 2, we observe similar car-like structures as seen in the Rectified Flow with U-Net in ambient space approach. While many of the generated images can be easily recognized as cars, they often lack high-frequency details, resulting in slightly blurred appearances. Notably, the VAE-GMM model shows interesting color inconsistencies within the generated cars, such as patches of a gray car turning red or a red car showing blue sections (e.g., most left in the second row, and second left from the last row). This phenomenon was not as prevalent in the Rectified Flow model and highlights the VAE's tendency to blend color information unpredictably, adding an unexpected variation to the generated

samples.

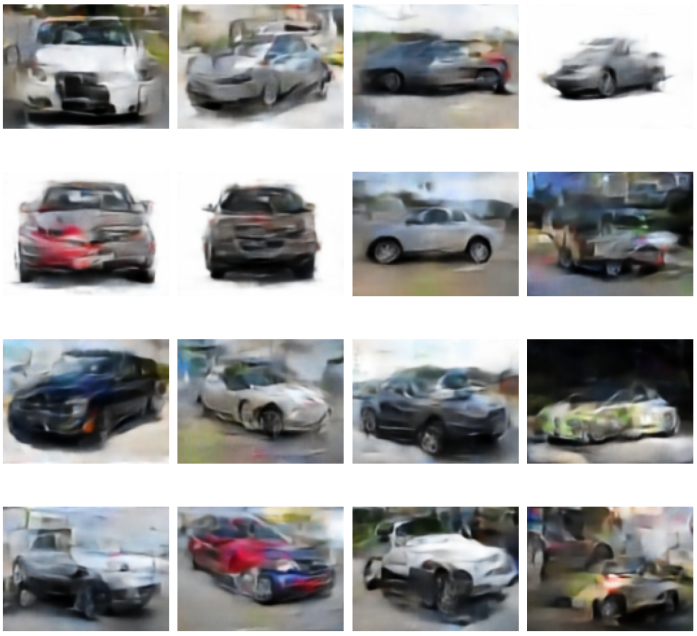

Figure 2: Generated samples using VAE with GMM in the latent space.

4.3.3 VAE with Rectified Flow in Latent Space

This approach applied Rectified Flow directly within the VAE's latent space, utilizing the compact structure of the VAE to refine the sampling process and generate high-quality images with improved fidelity. The Rectified Flow was trained on the VAE's latent vectors using a Multi-Layer Perceptron (MLP) rather than a U-Net, given the vector nature of the input. The results, shown in Figure 3, demonstrated similar image quality to the VAE-GMM approach.

4.3.4 Latent Space Analysis

To further understand the behavior of the VAE, we conducted a detailed analysis of its latent space using dimensionality reduction techniques such as t-SNE. This analysis aimed to assess the structure of the latent space and identify potential correlations between Moira's scores and specific latent dimensions. We found that the latent space was continuous and lacked distinct clusters, suggesting that Moira's high-rated images were not confined to specific regions but were distributed across the entire space, as shown in Figure 4.

Additionally, a significant fraction of the latent activations before reparametrization were consistently near zero, prompting us to refine sampling by focusing on the 18 most activated neurons, as seen in Figure 5. Both the Rectified Flow and GMM trained on this reduced latent space slightly improved the final sampling fidelity, although the performance of Rectified Flow did not exceed that of the GMM, as visually assessed.

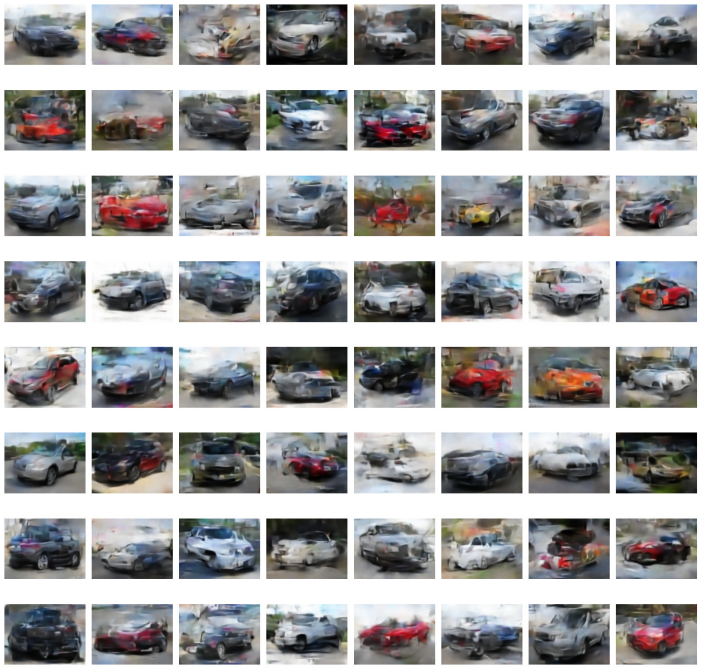

Figure 3: Generated samples using VAE with Rectified Flow in the latent space.

4.4 Regressor for Conditioning on Moira's Preferences

The regressor, a modified pre-trained ResNet18 model, was tuned through grid search for optimal hyperparameters and then trained on the full dataset with Moira's scores as targets. This model's role was crucial for scoring the generated images and determining which samples best matched Moira's preferences. By using the regressor, we indirectly conditioned the image selection process without altering the generative models themselves.

4.5 Image Selection

A total of 16,000 images were generated per each of the three methods: Rectified Flow with U-Net, VAE with GMM, and VAE with Rectified Flow. Each generated image was then scored by the regressor based on Moira's preferences. From each method, the top 20 images with the highest predicted scores were selected as the final output. Figure 6 shows the top 20 images selected from the VAE-GMM model, while Figure 7 displays the corresponding images for the Rectified Flow with U-Net. The highest-scored image is located in the bottom right of each figure.

5 Discussion

After initially believing that our Rectified Flow with U-Net approach generated a broad variety of new images, the evaluation with the regressor revealed a different reality. The model consistently produced highly similar outputs, particularly among the top 20 images, as shown in Figure 7. This lack of diversity was evident, with 18 out of the top 20 images featuring nearly identical red cars and the remaining 2 being nearly identical silver cars. This suggests that the model's output homogeneity

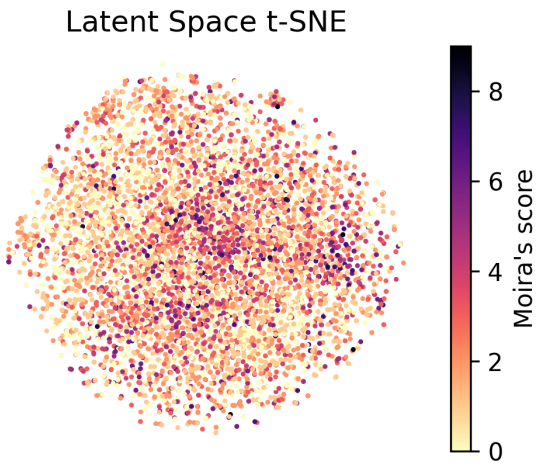

Figure 4: t-SNE plot of the 18 latent space dimensions. The dots are colored by Moira's score of the car.

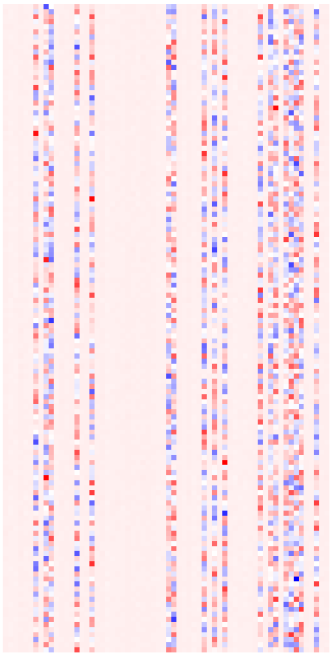

Figure 5: Plot of the data matrix of the latent space, rows are individual datapoints. The white space indicates features that are not activated.

likely extends to the rest of the generated dataset as well. In sharp contrast, the VAE with GMM approach demonstrated a broader variety of generated samples and overall better results, highlighting a significant limitation of the Rectified Flow model in terms of variety and novelty.

The VAE-GMM approach achieved moderate success in generating images that align with Moira's preferences, as seen in Figure 6, though some notable limitations were also observed. Analysis of the highest-scored images in the training data showed a strong preference for colorful sports cars, particularly in shades of red, orange, and yellow (see Figure 8). Silver cars, although frequently rated highly by Moira and generated often (Figure 2), were conspicuously absent from the top 20 images selected by the regressor in the VAE-GMM approach. This suggests that the regressor may have overfitted the color features, favoring brighter and more vibrant cars over others, regardless of their overall visual quality. Additionally, while Moira's highest scores reached 9, the regressor's maximum predictions peaked at around 4.4, indicating that while the generated images did not fully match the quality of the highest-rated real images, the relative ranking of preferences was still effectively maintained.

The relatively poor performance of the Rectified Flow model can likely be attributed to the limited size of the training dataset. Models from the diffusion family, including Rectified Flow, are typically trained on large-scale image datasets with millions of samples, unlike our smaller dataset of only a few thousand images. This lack of data hinders the model's ability to learn complex visual distributions effectively.

For the VAE, the lower fidelity of the generated images may result from the latent distribution not perfectly matching the imposed standard normal distribution. The KL-divergence penalty in the VAE loss function regularizes the latent space but may not fully capture the intricacies of the car images, leading to blurriness and artifacts in the generated outputs.

Contrary to our initial expectations, the combination of VAE and Rectified Flow in the latent space did not outperform other methods. Despite using a reduced 18-dimensional latent space, the 6000 training samples might still be insufficient for the Rectified Flow to accurately model the distribution. The computational complexity and data requirements of Rectified Flow suggest that more data would be necessary to fully leverage this model's capabilities.

The use of a GMM within the VAE latent space provided a slight improvement in image fidelity, which aligns with the GMM's capacity to capture more complex, multimodal structures compared to the simpler standard normal distribution. This marginal boost suggests that even slight increases in expressiveness can enhance the visual quality of generated samples, though the improvement is not dramatic.

Our approach of outsourcing the conditioning to a separate regressor model had its own set of advantages and disadvantages. The primary advantage of this strategy is its flexibility; the generative models were not directly conditioned on Moira's preferences, allowing us to generate a wide variety of images first and filter them post hoc. This method simplifies the generative process and makes it adaptable to changes in target preferences without retraining the generation models themselves.

However, this approach also presented several drawbacks. The regressor, trained on a limited dataset, likely overfits specific features like color rather than evaluating the overall quality of the cars, as indicated by the disproportionately high scores for colorful vehicles. This reliance on the regressor for conditioning can result in suboptimal selection, especially if the regressor fails to generalize well beyond the training data. Enhancing

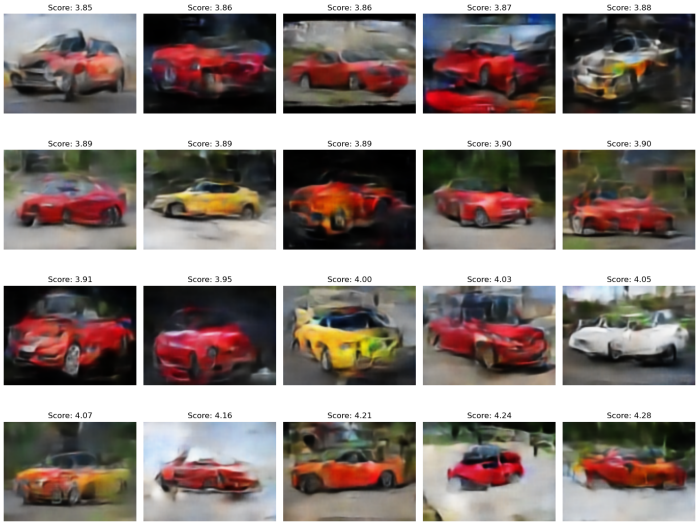

Figure 6: Top 20 images selected by the regressor for the VAE-GMM model. One can see that more features from Moira's favourite cars are incorporated compared to Figure 2.

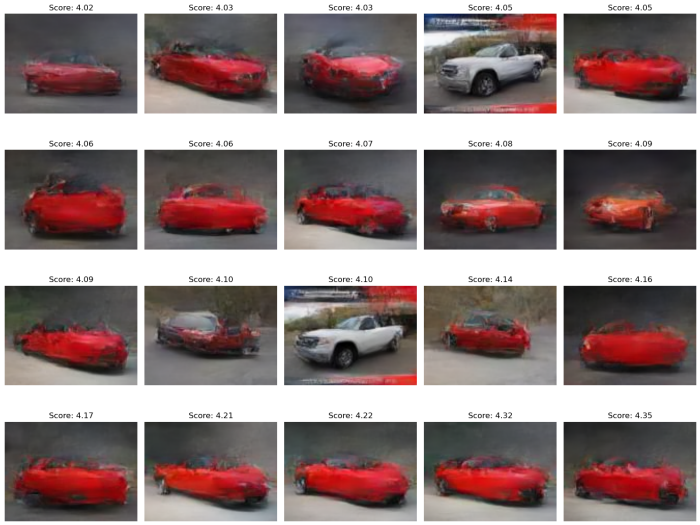

Figure 7: Top 20 images selected by the regressor for the Rectified Flow with U-Net model.

the regressor's performance through training on a larger, more diverse dataset would be crucial for improving the quality of selected images. Furthermore, our method requires generating a large number of images to identify those that align with Moira's preferences, rather than directly generating conditioned images, which is less efficient and resource intensive.

5.1 Comparison of our Results with relevant literature

One of our selected models was Rectified Flows in an ambient space. To evaluate this, we first generated an image using the state-of-the-art Rectified Flows model, FLUX.1, as presented by Labs [2024]. The FLUX.1 model employs a hybrid architecture consisting of multimodal and parallel diffusion transformer blocks, scaled to 12 billion parameters. The generated results are shown in Figure 9.

We observe that the image produced by FLUX.1 (see Figure 9) is of significantly higher quality compared to the output of our Rectified Flow model in an ambient space, as shown in Figure 1. This disparity is largely due to our model's much smaller training set and its considerably lower parameter count. Additionally, FLUX.1 benefits from a more advanced architecture, whereas our model relies on simpler convolutional U-Net layers for data efficiency. Even the original images from the FLUX.1 paper exhibit superior quality, which can likely be attributed to differences in dataset size. While we trained on a dataset of 6,000 images, FLUX.1 was trained on the much larger LSUN dataset [Yu et al., 2016], where each category contains between 120,000 and 3 million images. Another key difference is that we did not implement the Reflow technique. While this choice does not affect the quality of the generated images, it remains a notable distinction. Reflow primarily improves inference speed, but given that inference time was not a priority in our case, we opted for simplicity and chose not to incorporate it.

Our results with the VAE generally aligned with expectations

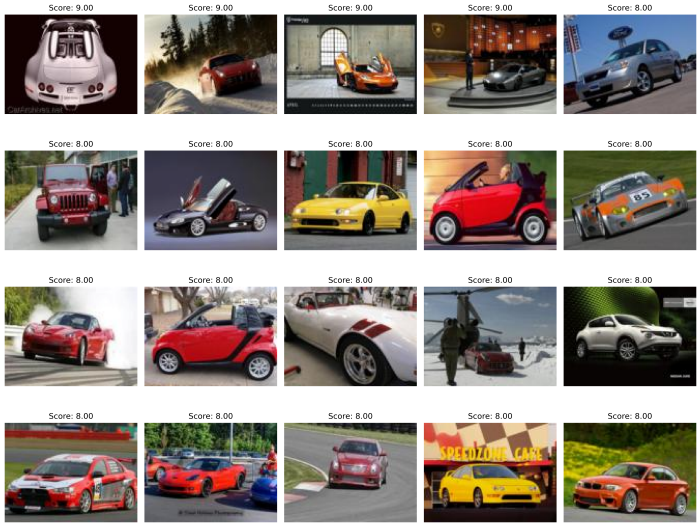

Figure 8: Top 20 images rated by Moira.

based on the literature. As mentioned, the VAE not only produced blurry images that missed fine details but also exhibited inconsistent color patches. However, it demonstrated a clear advantage in terms of sample efficiency, as it produced more diverse and better quality images than the newer Rectified Flows approach. In addition, the latent space offered interpretability benefits. In our case, this was not extraordinarily helpful, as no clusters were observed in the Miora scores, but it remains a useful feature for visualization and could be beneficial for other tasks.

6 Conclusion

The experiments revealed significant differences in the performance of our generative models. The Rectified Flow with U-Net model showed limited diversity, often generating nearly identical images, particularly red cars, underscoring its struggle with variety and finer details. In contrast, the VAE with GMM approach performed better, producing more varied and visually

Figure 9: One generated sample of the 12B parameter FLUX.1 [schnell] [Labs, 2024]

distinct samples that aligned more closely with Moira's preferences, though the images often lacked high-frequency details and exhibited some blurriness.

Future improvements could focus on expanding the training dataset with more diverse and extensive samples to enhance the models' ability to capture complex visual distributions. Additionally, integrating advanced architectures such as VQ-VAEs could improve image fidelity and detail, addressing the limitations seen in the current models. Exploring techniques with direct conditioning, such as Diffusion Models or Conditional GANs, may also offer a more efficient approach, allowing the generation of images directly aligned with target preferences, reducing the reliance on post hoc selection processes.

Regarding our task to fulfill the requirements for PUSHPOP INC., we would have selected the top 10 generated images from the VAE-GMM model, which best matched Moira's preferences. These images can be found in the Appendix 7.

7 Group contributions

Paul contributed by implementing all three image generation models, performing Flux image generation, debugging and formatting the code, and contributing to parts of the methodology and discussion. Thalis was responsible for implementing the regressor, applying it to the images, generating plots of the results, exploring the dataset, and contributing to the methodology, discussion, experiments, results and the conclusion. Blanka focused on the related work, introduction, parts of the methodology and discussion, abstract, dataset exploration, formatting the appendix, and putting the final touches on the report.

References

Asperti, A., Evangelista, D., & Piccolomini, E. L. (2021). *A survey on variational autoencoders from a greenai perspective.* Retrieved from https://arxiv.org/abs/2103.01071

Cao, H., Tan, C., Gao, Z., Xu, Y., Chen, G., Heng, P.-A., & Li, S. Z. (2023). *A survey on generative diffusion model.* Retrieved from https://arxiv.org/abs/2209.02646

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, *29*(6), 141–142.

Dhariwal, P., & Nichol, A. (2021). *Diffusion models beat gans on image synthesis.* Retrieved from https://arxiv.org/abs/2105.05233

Dinh, L., Krueger, D., & Bengio, Y. (2015). *Nice: Nonlinear independent components estimation.* Retrieved from https://arxiv.org/abs/1410.8516

Dinh, L., Sohl-Dickstein, J., & Bengio, S. (2017). *Density estimation using real nvp.* Retrieved from https://arxiv.org/abs/1605.08803

Elgammal, A., Liu, B., Elhoseiny, M., & Mazzone, M. (2017). *Can: Creative adversarial networks, generating "art" by learning about styles and deviating from style norms.* Retrieved from https://arxiv.org/abs/1706.07068

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., ... Rombach, R. (2024). *Scaling rectified flow transformers for high-resolution image synthesis.* Retrieved from https://arxiv.org/abs/2403.03206

Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). *Synthetic data augmentation using gan for improved liver lesion classification.* Retrieved from https://arxiv.org/abs/1801.02385

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). *Generative adversarial networks.* Retrieved from https://arxiv.org/abs/1406.2661

Graves, A. (2014). *Generating sequences with recurrent neural networks.* Retrieved from https://arxiv.org/abs/1308.0850

Ho, J., Jain, A., & Abbeel, P. (2020). *Denoising diffusion probabilistic models.* Retrieved from https://arxiv.org/abs/2006.11239

Huang, Z., Zhou, P., Yan, S., & Lin, L. (2023). *Scalelong: Towards more stable training of diffusion model via scaling network long skip connection.* Retrieved from https://arxiv.org/abs/2310.13545

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). *Progressive growing of gans for improved quality, stability, and variation.* Retrieved from https://arxiv.org/abs/1710.10196

Kingma, D. P., & Dhariwal, P. (2018). *Glow: Generative flow with invertible 1x1 convolutions.* Retrieved from https://arxiv.org/abs/1807.03039

Kingma, D. P., & Welling, M. (2022). *Auto-encoding variational bayes.* Retrieved from https://arxiv.org/abs/1312.6114

Krahe, C., Bräunche, A., Jacob, A., Stricker, N., & Lanza, G. (2020). Deep learning for automated product design. *Procedia CIRP*, *91*, 3-8. (Enhancing design through the 4th Industrial Revolution Thinking) doi: https://doi.org/10.1016/j.procir.2020.01.135

Labs, B. F. (2024). *flux.* Retrieved from https://github.com/black-forest-labs/flux.git (Accessed: 2024-09-16)

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., Huang, F., et al. (2006). A tutorial on energy-based learning. *Predicting structured data*, *1*(0).

Li, J., Zhang, C., Zhu, W., & Ren, Y. (2024, 06). A comprehensive survey of image generation models based on deep learning. *Annals of Data Science*, 1-30. doi: 10.1007/s40745-024-00544-1

Liu, X., Gong, C., & Liu, Q. (2022). *Flow straight and fast: Learning to generate and transfer data with rectified flow.* Retrieved from https://arxiv.org/abs/2209.03003

Mansimov, E., Parisotto, E., Ba, J. L., & Salakhutdinov, R. (2016). *Generating images from captions with attention.* Retrieved from https://arxiv.org/abs/1511.02793

Mirza, M., & Osindero, S. (2014). *Conditional generative adversarial nets.* Retrieved from https://arxiv.org/abs/1411.1784

Ngiam, J., Chen, Z., Koh, P. W., & Ng, A. Y. (2011). Learning deep energy models. In *Proceedings of the 28th international conference on machine learning (icml-11)* (pp. 1105–1112).

Oord, A. V. D., Kalchbrenner, N., & Kavukcuoglu, K. (2016). *Pixel recurrent neural networks.* Retrieved from https://arxiv.org/abs/1601.06759

Oord, A. V. D., Vinyals, O., & Kavukcuoglu, K. (2018). *Neural discrete representation learning.* Retrieved from https://arxiv.org/abs/1711.00937

Perez, L., & Wang, J. (2017). *The effectiveness of data augmentation in image classification using deep learning.* Retrieved from https://arxiv.org/abs/1712.04621

Radford, A., Metz, L., & Chintala, S. (2016). *Unsupervised representation learning with deep convolutional generative adversarial networks.* Retrieved from https://arxiv.org/abs/1511.06434

Razavi, A., Oord, A. V. D., & Vinyals, O. (2019). *Generating diverse high-fidelity images with vq-vae-2.* Retrieved from https://arxiv.org/abs/1906.00446

Ronneberger, O., Fischer, P., & Brox, T. (2015). *U-net: Convolutional networks for biomedical image segmentation.* Retrieved from https://arxiv.org/abs/1505.04597

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., & Ganguli, S. (2015). *Deep unsupervised learning using nonequilibrium thermodynamics.* Retrieved from https://arxiv.org/abs/1503.03585

Sohn, K., Lee, H., & Yan, X. (2015). Learning structured output representation using deep conditional generative models. In *Neural information processing systems.* Retrieved from https://api.semanticscholar.org/CorpusID:13936837

Song, J., Meng, C., & Ermon, S. (2022). *Denoising diffusion implicit models.* Retrieved from https://arxiv.org/abs/2010.02502

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2021). *Score-based generative modeling through stochastic differential equations.* Retrieved from https://arxiv.org/abs/2011.13456

Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., & Xiao, J. (2016). *Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop.* Retrieved from https://arxiv.org/abs/1506.03365

Appendix: Our Final Submission for the Project Moira's Dream Car

To: PushPop INC

This appendix contains 10 images generated using a Variational Autoencoder (VAE) with Gaussian Mixture Models (GMM), showcasing the images with the highest estimated Moira scores.

Image 1: Highest estimated Moira score

Image 2: Second highest estimated Moira score

Image 3: Third highest estimated Moira score

Image 4: Fourth highest estimated Moira score

Image 5: Fifth highest estimated Moira score

Image 6: Sixth highest estimated Moira score

Image 7: Seventh highest estimated Moira score

Image 8: Eighth highest estimated Moira score

Image 9: Ninth highest estimated Moira score

Image 10: Tenth highest estimated Moira score