Comics Illustration Synthesizer using the Stable Diffusion Model

Fine-tuning for text-to-image Dilbert Comics Generation

Mahmoud Elaref[1]

Supervisor(s): Lydia Chen[1] **, Zilong Zhao**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 25, 2023

Name of the student: Mahmoud Elaref
Final project course: CSE3000 Research Project
Thesis committee: Lydia Chen, Supervisor: Zilong Zhao , Examiner Anna Lukina

Abstract

Synthetic art is the end result of artificial intelligence models that have been trained to generate images from text prompts. "Comic synthesis" is one such use case, where comic illustrations are produced from textual descriptions. Previous attempts at comic synthesis have utilized conditional Generative Adversarial Networks (cGANs), but this approach has encountered challenges in generating consistent and visually appealing comic panels. Strict data requirements and quality limitations have left room for improvement. We propose a novel approach to comic synthesis using Stable Diffusion, a powerful generative modelling technique. The study investigates the fine-tuning of the stable diffusion model specifically for the generation of Dilbert Comics from textual prompts. We explore different techniques to fine-tune the stable diffusion model for comic synthesis including Dreambooth and LoRA. Through extensive experimentation and analysis, with an FID score of 123, results produced using the Lora technique outperformed Dreambooth, excelling in understanding the Dilbert style while Dreambooth struggled with multiple-subject training. Results are also compared with previous approaches based on conditional GANs. While the quality and detail greatly improved, the transition from conditionals to text descriptions meant the results were less accurate. The results show the potential of stable diffusion in generating appealing Dilbert Comic panels, while highlighting the need for further exploration to enhance the alignment between textual descriptions and the generated images.

1 Introduction

Comic illustrations play a significant role in conveying humour, satire, and social commentary in popular culture. The ability to automatically generate comic illustrations from text descriptions, Comic Synthesis, has gathered considerable interest in the field of artificial intelligence and computer vision. In this research we explore the effectiveness of using stable diffusion models in generating high-quality Dilbert comic illustrations from text descriptions.

1.1 Motivation

The motivation behind this research stems from the desire to automate the process of comic illustration, specifically focusing on Dilbert comics. Dilbert, created by Scott Adams, is a popular comic strip that satirizes the corporate world and offers humorous insights into workplace dynamics [1]. Dilbert comics has a simple and consistent style, it is made up of a few characters and is mostly set in an office [2]. This makes it the perfect testing and starting point for text-to-image comic synthesis. However, despite the progress made in the field of generative modeling, achieving satisfactory results in automated comic generation remains a challenging task, as shown in the results produced by Morris in Figure 2 [3].

By exploring the fine-tuning of the Stable Diffusion model for Dilbert Comics, this study aims to bridge the gap in automated comic generation. The motivation lies in the potential of stable diffusion to overcome the limitations of previous approaches and generate high-quality comics that capture the essence of the Dilbert universe.

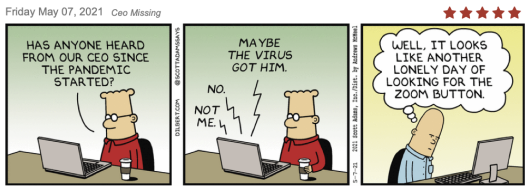

Figure 1: Example of a Dilbert Comic strip. Characters and scenes are simple to illustrate.

Figure 2: Example comic panels generated using Stability GAN by Morris.

1.2 Research Challenges

Previous approaches, particularly using conditional Generative Adversarial Networks (cGANs), have shown promising but limited success in generating high-quality and meaningful comics. Firstly, the use of conditions meant limiting the creativity of the comics and their content. In previous research by Morris et. al [3], comics could only be generated based on character presence and background colour. Additionally, the results in Figure 2 were good enough to identify the characters, yet they could be improved to show more character detail and be of better quality.

One of the key challenges encountered in the fine-tuning process of stable diffusion is the recommended dataset size, which tends to be relatively small, (3-5 images) [4]. Having to use a dataset with over 100 images to capture essential Dilbert components, might cause problems when training and produce mixed results. Moreover, the wide range of characters make it particularly problematic when attempting multi-subject training, as the fine tuning exhibits a preference for training on a single subject. which is not feasible in the context of Dilbert, which features eight main characters.

Another challenge lies in crafting suitable text descriptions. The descriptions must strike a balance between providing sufficient detail to capture key elements of the desired comic panel and avoiding excessive complexity that could hinder the training process. Finding the optimal level of de-

tail and complexity in the text prompts is crucial for training the model effectively.

1.3 Proposed Solution

The proposed solution for automated comic generation revolves around the utilization of stable diffusion, a powerful generative modeling technique. The initial step in comic synthesis is to produce an image of the scene of a comic panel, without dialogue. Focusing on capturing the visual composition and storytelling elements, so that the dialogue can be later added.

Diffusion models are machine learning systems that are trained to progressively remove random Gaussian noise in order to obtain a desired sample, such as an image. They have demonstrated impressive results in image generation tasks. However, they suffer from slow reverse de-noising and high memory consumption when operating in pixel space. To address these challenges, latent diffusion models have emerged, which perform the same diffusion process in a lower-dimensional latent space instead of pixel space [5]. In latent diffusion, the model is trained to generate compressed latent representations of the images. This approach reduces the memory and complexity associated with diffusion models, making them more feasible for training and inference.

Stable Diffusion, a Latent Diffusion model, is chosen as the method in this study. Fine-tuning the Stable Diffusion model means adjusting its parameters and optimizing its performance for a specific task or dataset, in this case training it on Dilbert comics. Various techniques have been found to fine-tune the Stable Diffusion model. Dreambooth [4], LoRA [6] and Textual Inversion [7]. The paper will explore these methods and their trade-offs in the context of Dilbert comic synthesis.

To address the challenges posed by the limited dataset size and the preference for single-subject training in fine-tuning methods, we test and evaluate two different approaches: Dreambooth and LoRA. We compare the effectiveness of these methods in capturing the Dilbert style and generating high-quality comic panels. By evaluating their performance, we aim to determine the most suitable approach for our task.

The stable diffusion model will be trained using datasets from Dilbert archives [8]. Through an iterative process of training and evaluation, the expected results will be evaluated against predetermined criteria, such as image quality, and matching of pre-conditions. This evaluation will shed light on the extent to which stable diffusion can enhance automated comic generation and provide valuable insights into the strengths and limitations of the approach.

1.4 Research Questions

Conducting this research will answer the the following questions: **"How can the Stable Diffusion model be fine-tuned such that it generates high quality Dilbert comic illustrations from text descriptions?"**. While answering this we will aim to answer the following questions:

- "How do the results of the fine-tuned Stable Diffusion model compare to those of the conditional GAN model in terms of quality and accurately matching preconditions"
- "Which fine-tuning method, Dreambooth or Lora, produces the highest quality Dilbert comic panels?"
- "How accurately do Dreambooth and Lora fine-tuning methods convey the textual descriptions of Dilbert comic panels?"

Structure of the Paper The paper is structured as follows, Section 2 discuses the Prior related work. Section 3 describes the methodology of how the Stable Diffusion model is fine-tuned for Dilbert comic synthesis, and the experimental design. The results and discussion can be found in section 4. Section 5 reflects on the ethical implications, reproduciblity and integrity of the research. Section 6 provides a conclusion of the research as well as suggestions for future work.

2 Prior Art and Preliminary

2.1 Generative models for comics

Relevant existing work in this domain includes a the paper by Morris [3] and another by Proven-Bessel [9]. Both exploring the generation of comic illustrations using conditional Generative Adversarial Networks (cGANs), demonstrating their capability to generate comic images from conditional inputs and text inputs respectively. The papers' contributions included the conditioning of the generative process on relevant input information, enabling Dilbert comic synthesis.

While their work provides valuable insights into generating comics using cGANs, latent diffusion models (LDMs) discussed in the "High-Resolution Image Synthesis with Latent Diffusion Models" [5] paper emphasises their advantages over GANs, it remains to be seen whether stable diffusion models specifically are more effective than cGANs in the context of generating Dilbert comics.

There is also a model that trained the stable diffusion model using Dreambooth on the Dilbert concept [10]. However, trained on a dataset of 7 panels, the model merely captured the Dilbert concept and style, not enough to understand any characters other than Dilbert himself. Despite this, the model's ability to capture the style indicates its strength and potential.

2.2 Background

Stable diffusion has emerged as a powerful generative modeling technique in text-to-image generation [5]. It operates by iteratively updating an image through a diffusion process, gradually revealing the desired target image. This process involves transforming a given noise vector or an initial image into a coherent and visually appealing final image.

Stable diffusion is a multi-step process. Firstly, text descriptions are translated into CLIP embeddings [11], which capture word meaning and contextual information. Secondly, as seen in Figure 3, an autoencoder encodes the image x, into a lower-resolution but detailed representation Z. In the red block, the diffusion happens in latent space by de-noising the noisy latent vector Z_T step by step all the way back to a prediction of Z. Finally, another autoencoder decodes the latent vector back into an image result $\tilde{x}$.

The network predicts the noise in an image at a specific time step, and this prediction is used to remove noise iteratively and generate a more accurate image at the initial time

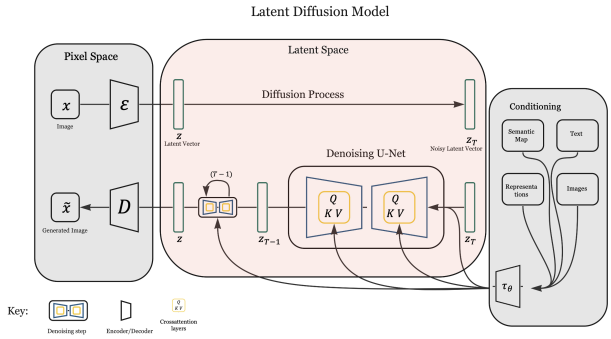

Figure 3: Latent diffusion models diagram.

step. The process begins with an image consisting of only noise and is conditioned with text embeddings. To align the generated image with the provided text, The conditioning block controls the synthesis process through inputs such as text or semantic maps. Classifier Free Guidance (CFG) is applied, each step is applied once with the text embedding guidance and once without, where the difference between the results is amplified and fed into the next step. This process effectively directs the network to align the generated image with the given text, improving the correspondence between the two.

Stable diffusion is well-suited for fine-tuning to produce images in a specific style or to understand complex concepts. By using Dilbert datasets for training, the fine-tuning process will enhance stable diffusion's capacity to generate images that align with the desired style, Dilbert comics.

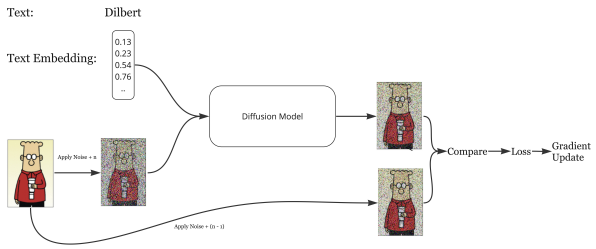

Figure 4: Fine tuning stable diffusion.

There are essentially three different techniques that have been found to train stable diffusion, Dreambooth, LoRA and Textual Inversion. Each using a different method to train the model to associate a 'subject' with its 'identifier', for example, the image of the Dilbert character with the word 'Dilbert'.

All three techniques aim to try and train the model to associate the word Dilbert with the picture of the Dilbert character, as shown in Figure 4. The diffusion model takes two inputs, first text embedding that represents the text input. Second, an image with noise of n steps added. The diagram, then tries to predict what that image looks like after $n-1$ steps of noise. The result is then compared, to the actual image with $n-1$ noise added. A loss function is then calculated so that a gradient update can be performed.

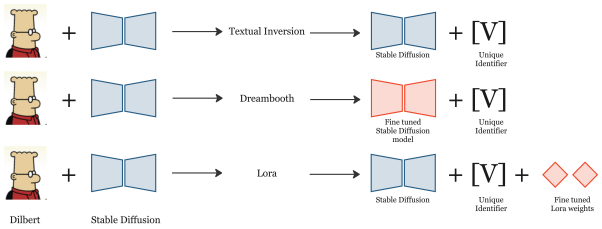

Figure 5: The difference between each fine tuning technique.

Dreambooth

Dreambooth is the first and most common method for fine tuning text-to-image diffusion models, it aims to improve the control and quality of synthesized images by incorporating 'subject-driven generation' [4]. The paper introduces this fine-tuning approach to enhance the subject fidelity and consistency of the generated images.

When training with Dreambooth, the model takes both inputs, the image(s) of the subject, and the text embedding, its goal is to associate a unique identifier in the text prompt, with the subject. Throughout several steps of training, the loss function is used to perform a gradient update on the model, in order to improve its understanding of the desired subject, and link it to its identifier in the prompt.

Essentially, as shown in Figure 5, a whole new diffusion model is produced, after constantly altering the internal structure of the diffusion model during training. This makes Dreambooth the most effective method when training stable diffusion. However, it is also quite storage inefficient, as each model is large in size, this becomes increasingly problematic when trying to train multiple concepts and end up with multiple models.

Textual Inversion

Textual Inversion is a different technique that learns to represent the target subject through new 'words' in the embedding space of the frozen text-to-image model [7]. Textual Inversion is the same in the sense that it tries to associate a unique identifier in the prompt, with its subject. However, rather than updating the diffusion model, it updates the text embedding, specifically, it updates the vector that that corresponds to the new word. By changing the meaning of this word, bringing it closer to what it should look like. The result is that a vector representation of the word (identifier) that is understood by the model as the target subject. As shown in Figure 5, the stable diffusion model remains unchanged.

LoRA

Low-Rank Adaption, LoRA, is a novel technique introduced to deal with the problem of fine-tuning large-language models [6]. Powerful models with billions of parameters, such as GPT-3, are prohibitively expensive to fine-tune. LoRA consumes less memory while accelerating the training of large models. It works by adding rank-decomposition weight matrices (update matrices) to the existing weights, and only trains the newly added weights, while the original weights are kept frozen.

LoRA was adapted to fine tune Stable Diffusion [12]. It works in a similar way to Dreambooth, but instead of updating the whole model, it adds new weights inside it, without changing existing weights in the model, as shown in Figure 5. This makes LoRA training easily portable because only the new weights need to be shared. This also allows more flexibility, the extent to which the model is used can be controlled via a scale parameter.

LoRA focuses on making small changes to the critical cross-attention layers, the yellow boxes in Figure 3, these are the layers where the image and prompt interact, this has been found to be sufficient for effective training. By modifying the weights of the cross-attention layers, LoRA fine-tunes the model, utilizing smaller low-rank matrices to reduce the size of the LoRA model files. An article [13] suggests that while LoRA excels in modifying styles, it may have limitations in handling facial features. This implies that it might be good in illustrating the Dilbert comic style.

3 Methodology

3.1 Solution

Out of the three fine-tuning techniques, textual inversion is the weakest in terms of results. According to Ruiz et al. [4], there is an "overwhelming preference" for Dreambooth over Textual Inversion. Dreambooth and Lora appear to be better options. Thus, we will employ Dreambooth and Lora and fine-tune them while carefully controlling the training factors. This includes using the same dataset of images and captions, training them for an equal amount of time, utilizing the same baseline stable diffusion model (v1.5), and employing the same autoencoder (VAE). The performance of both fine-tuned models will be evaluated by generating sets of images using the same text prompt. The quality and accuracy of the generated images will be compared to determine which model produces superior results in terms of capturing the Dilbert style and conveying the intended meaning.

3.2 Dataset

The dataset used is obtained from the Dilbert archives [8]. Firstly, the dataset is processed by splitting the panels, training and generating is a lot simpler on one panel rather than a whole comic strip. Next, we remove the dialogue, to focus solely on the visual content. This is due to image generation models' struggle to produce text on an image. It makes sense to train and produce comics without dialogue for better results, and the dialogue can always be added on top later. To ensure data quality, a filtering process is implemented to remove any erroneous or low-quality entries. Resulting in a dataset that provides a reliable foundation for training the model. The dialogue is removed using an object detection model, fine tuned on Dilbert panels, where a bounding box is placed on top of the text. The color used to cover is decided by taking the most frequent color in that box, usually being the background colour.

3.3 Captioning

Originally, the idea was to use the panel dialogue to describe the scene, this quickly changed as textual descriptions that describe characters and objects proved to be much more effective. To caption each image with a description, initially BLIP captioning was used. BLIP is a model that is known for image captioning [14]. The results were limited in quality. BLIP could not identify the characters and gave poor descriptions because it is unfamiliar with the Dilbert universe. For example, all characters were addressed as men and women, making it difficult to reproduce them later, moreover, the Boss' long hair characterised him as a woman according to BLIP. To address this, a manual approach is adopted, involving character and scene labeling. This process relied on utilizing tags and character names to accurately describe the content of each panel. Short phrases separated by commas to describe everything visible in a panel. By manually annotating the dataset with precise captions, a collection of 180 text-image pairs was curated, ensuring a more reliable and informative dataset for training the stable diffusion model in the context of Dilbert comic generation. Figure 6 shows an example of captions used.

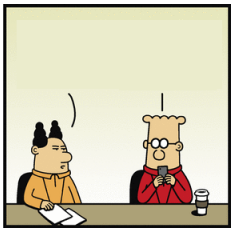

Figure 6: Image captioned: "Dilbert, Boss, Boss sitting at table, Dilbert sitting at table, Boss holding paper, Dilbert holding phone, coffee cup, table".

3.4 Training

The training of the fine-tuned stable diffusion model using both, LoRA and Dreambooth, was completed efficiently. To control the training time and keep it constant, each model was trained for a different amount of steps, but both amounting to two hours on a dataset of 180 text-image pairs. The training was conducted on the Google Compute Engine's Colab platform, utilizing a single A100 GPU in the "High-RAM" GPU run-time setting. The base model used for training was the standard stable diffusion model v1.5. Additionally, the stable diffusion (vae) was employed to further enhance the model's ability to capture and encode complex visual representations. Results presented are inferred at 20 steps, greater number of steps would yield better results, but would take longer.

3.5 Experiments

To assess the models' capability, experiments are conducted. The experiments consist of a set of prompts designed to generate panels with two Dilbert characters. The model's capability to produce panels with the desired characters will be tested using accuracy scores, showing the percentage of correct panels. More experiments will be conducted to test both the simpler or more complex scenarios, with different number of characters, to really test the limits of the model.

4 Experimental Setup and Results

This section provides the results, analysis and evaluation of the obtained models. It aims to provide a comprehensive assessment of the both Dreambooth and Lora model performance, focusing on their ability to accurately match preconditions, convey semantic meaning, and produce high quality Dilbert comic panels. Additionally, a comparison is made between the results achieved by the fine-tuned stable diffusion model and those of the conditional GAN model.

4.1 Evaluation Metrics

In evaluating the performance of the fine-tuned models for comic generation. Firstly, accuracy in following descriptions was measured by calculating the percentage of accurately predicted characters appearing in the generated comic panels. Evaluating the models' capability of understanding the difference between characters. Also, examining how effectively the generated comic panels conveyed the intended meaning from the provided text descriptions to determine the extent to which the panels successfully captured the semantic elements and context specified in the text prompts.

Moreover, the Fréchet Inception Distance (FID) score was employed as a measure of quality to assess how closely the generated comic panels resemble the original Dilbert comic style. The FID score takes into account both visual similarity and distribution matching between the generated panels and the ground truth comic panels. A lower FID score indicates a higher quality and similarity to the original style.

4.2 Stable Diffusion

Before fine-tuning, the stable diffusion model showed no resemblance to the Dilbert style and instead showcased a generic comic style. In comparison, the Lora and Dreambooth techniques show promise for enhancing the model's ability to capture specific styles. An article suggests that Lora excels in capturing styles but may face challenges in accurately depicting faces [13]. It is worth noting that there is limited research available that directly compares these fine-tuning techniques. This study aims to address this gap by evaluating and comparing the performance of Lora and Dreambooth in the specific context of Dilbert comics, thereby providing valuable insights on their effectiveness.

Pre fine-tuning

In the evaluation of the fine-tuned stable diffusion model, a comparison was made between the results generated before and after the fine-tuning process. A significant improvement in capturing the Dilbert style was observed in the fine-tuned model. This is demonstrated in Figure 7, which depicted the extent to which the fine tuned LoRA model influenced the generation process for the same prompt and seed, on a scale of 0 to 1.

At the lower end of the scale (0), where the base stable diffusion v1.5 model was employed, the model lacked specific knowledge about Dilbert and produced generic comic drawings without the distinct Dilbert style. However, as the scale approached 1, indicating the increased usage of the fine tuned LoRA model, the generated results exhibited a significant improvement and demonstrated the proper Dilbert style. The fine-tuned model exhibited characteristics such as the 2-D visual representation and more identifiable characters.

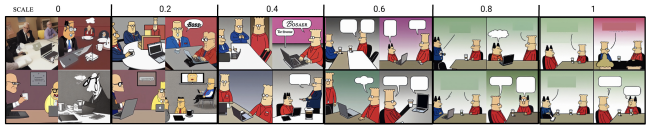

Figure 7: A graph showing the same prompt generating the same image while varying the the extent to which the fine tuned LoRA model is used from 0 to 1.

Lora Results

The results obtained from the application of Lora fine tuning were highly impressive. The model demonstrated great capabilities in distinguishing between different characters in Dilbert comics, accurately capturing their unique traits and characteristics, as shown in Figure 8. Moreover, the generated comic panels showcased a remarkable adaptation of the Dilbert style across various environments, showcasing the model's ability to replicate the distinct aesthetics and visual elements associated with Dilbert comics. However, it should be noted that the model occasionally produced duplicated characters in certain instances.

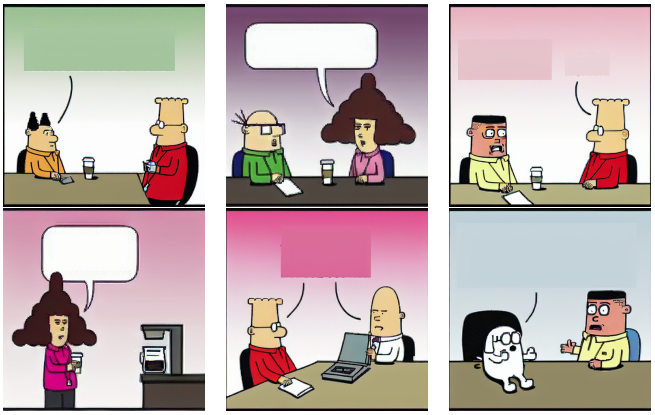

Figure 8: Lora fine tuning results for (a): Dilbert and Boss sitting, (b) Wally and Alice sitting, (c): Dilbert and Asok sitting, (d): Alice by the coffee machine, (e): Dilbert and CEO, (f): Asok and Dogbert

Dreambooth Results

The results obtained with Dreambooth were of a slightly lower quality compared to Lora; however, the model still managed to capture the essence of the Dilbert style fairly well. It successfully depicted the first character with accuracy, showcasing a good understanding of their visual attributes. However, there were instances where Dreambooth mixed up elements from different characters, leading to some confusion. Notably, the model encountered challenges when representing a second or third character, often mixing up characters. It still demonstrated an ability to capture the dominant features, such as the Boss's pointy hair and Dilbert's glasses. As shown in Figure 9, most of the images

include Dilbert or the boss, this is because the model struggled to produce the rest of the characters. Figures 9.d, 9.e were attempted with Dogbert, Alice respectively, and they failed to display their prominent features. Figure 9.f was a panel attempted without the Dilbert character, it can be seen that the loss of the word 'Dilbert' from the prompt, unlike Lora, resulted in loss of style from the model. Appears that the Dreambooth model, although promising, has some limitations.

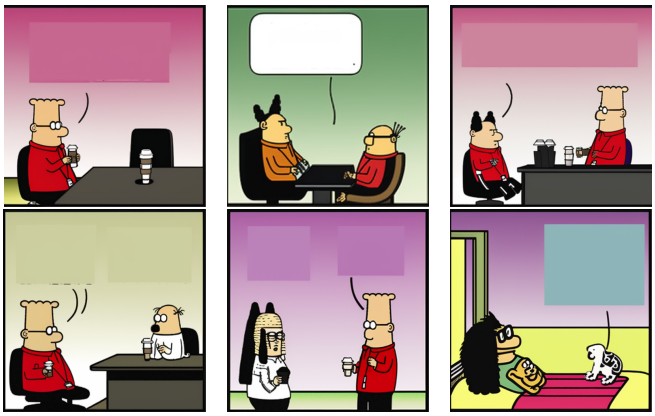

Figure 9: Dreambooth fine tuning results for (a): Dilbert, (b): Boss and Wally, (c): Dilbert and Boss sitting, (d): Dilbert and Dogbert, (e): Dilbert and Alice, (f): Wally and Dogbert in a bedroom.

Accuracy scores

In the comparison of accuracy experiments, both Lora and Dreambooth were evaluated by generating 50 images for each model using the same prompts. One set of prompts expected Dilbert and the Boss, while the other set expected the Boss and Wally. The results of the first experiment in Table 1 revealed interesting insights. In terms of accurately representing the first character, both models performed admirably, with high accuracy observed in almost all generated images. However, when it came to depicting the second character, the models struggled. Nonetheless, Lora clearly outperformed Dreambooth. While Lora produced scenes where the second character was depicted accurately, one out of every two images showcasing a perfect representation, Dreambooth showed difficulties.

To further investigate the issue observed with the second character in the previous accuracy experiments, the Boss, a second experiment was conducted. In this experiment, the order of the characters was changed, with the Boss appearing first followed by Wally this time. Surprisingly, the results, shown in Table 2 showed a significant improvement in the accuracy of the Boss character, which was now consistently generated with great precision. Which suggests that the problem for Dreambooth is not specific to certain characters but rather related to the order in which they are presented. Lora images excelled with Wally showing 94% accuracy and the Boss with 72%. Highlighting Lora's superiority in understanding the distinction of characters compared to Dreambooth, producing perfect scenes 46% of the time, almost once every two panels. For a visual reference of all the generated images, please refer to the appendix.

Model	Character 1	Character 2	Perfect Scene
Lora	100%	60%	60%
Dreambooth	96%	24%	18%

Table 1: Results for each model on the first prompt: Character 1: Dilbert, Character 2: Boss, Perfect Scene: Both characters present alone without any extra characters.

Model	Character 1	Character 2	Perfect Scene
Lora	72%	94%	46%
Dreambooth	100%	12%	12%

Table 2: Results for each model on the second prompt: Character 1: Boss, Character 2: Wally, Perfect Scene: Both characters present alone without any extra characters.

FID scores

The Fréchet Inception Distance (FID) is a commonly used evaluation metric in image generation tasks. FID measures the quality and similarity of the generated images compared to real images. FID is calculated by computing the Fréchet distance between two Gaussians fitted to feature representations of the Inception network [15]. Thus, quantifying the perceptual difference between the two image sets. A lower FID score indicates a higher level of similarity and quality between the generated and real images. As shown in Table 3, the results generated from the Lora model scored and FID score of 122.89 while Dreambooth scored 154.76. Showing once again that the Lora model outperformed the Dreambooth model, this time in quality and similarity to the original Dilbert comics.

4.3 cGAN

When comparing the results of the fine-tuned stable diffusion model to those of the conditional GAN model, notable improvements are observed in terms of detail clarity and overall image quality. Our model demonstrates enhanced capability in capturing finer visual details, resulting in more visually appealing and high-quality generated Dilbert comic panels.

When compared to conditional GANs, our accuracy decreases, this is expected due to the nature of textual descriptions in text-to-image models. The inclusion of character names introduces an extra dimension and complexity, leading to potential confusion. However, despite this trade off, the significant improvement in overall image quality and the extra possibilities introduced with textual descriptions justifies it. According to Morris [3] Stability GAN gives 92% accuracy on two-character panels.

The text-to-image model offers greater freedom compared to conditional models enabling us to explore diverse content.

Model	FID score
Lora	122.89
Dreambooth	154.76

Table 3: FID scores for both models.

Freedom in setting the comic in an office, bedroom, or garden and adding objects such as coffee cups, laptops and microphones result in more complex and higher quality Dilbert comics.

Furthermore, the FID score of 122.89 obtained in our evaluation demonstrates a noticeable improvement over the FID score reported in the previous work of ComicGAN, which achieved a score of 150.95 [9]. This indicates that our stable diffusion model has successfully enhanced the quality and realism of the generated comic panels compared to the previous approach.

4.4 Discussion

Capturing the Dilbert Style

The Lora fine-tuned stable diffusion model demonstrated great performance in capturing the distinctive style of Dilbert within the established scope of the Dilbert environment. The results produced show all the visual elements commonly found in Dilbert comics, including objects within the training set such as coffee cups and laptops.

To further assess the model's ability to capture the Dilbert style and explore its potential for generating novel content, additional experimentation was conducted. Thus, generating comic panels with scenarios beyond the conventional office environment, including an outdoor setting, a couch and a bedroom, as shown in Figure 10. These experiments aimed to evaluate whether the model could effectively adapt the Dilbert style to various contexts and invent new objects consistent with the overall visual style.

The results demonstrated that the model successfully extended its capabilities beyond the initial training set, generating compelling and visually consistent comic panels that maintain the essence of the Dilbert style while introducing new elements and environments. This highlights the model's potential for creative comic synthesis and the generation of engaging visual narratives that remain faithful to the established aesthetic.

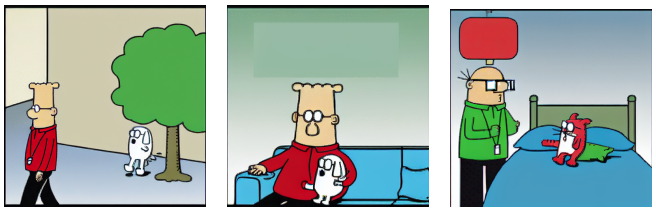

Figure 10: Lora results (a) Dilbert and Dogbert outdoors (b) Dilbert and Dogbert on a couch (c) Wally and Catbert on a bed

Dreambooth limitations

Dreambooth, although showing potential, faced certain challenges compared to Lora in the experiments. Dreambooth failed to illustrate certain characters effectively. One notable reason could be the longer training time required for Dreambooth. Due to us limiting it to the same training time as Lora, Dreambooth did not achieve as much in terms of accuracy and style adaptation. Given a longer training duration, the results may potentially improve.

Another struggle observed with Dreambooth was its difficulty in handling multiple subject training. The model occasionally mixed up the characters, suggesting a need for separate training for each character to enhance performance. The Dreambooth paper [4] recommends 3-5 images to train per subject, while Lora was able to adapt to our bigger dataset with multiple subjects, Dreambooth struggled. Lora also coped better with distinguishing between characters. Perhaps, with additional time and resources, Dreambooth could have potential to improve if single subject training was adapted, each character trained separately.

Lastly, Dreambooth lost touch of the style sometimes without Dilbert in the prompt, in Figure 9.f, while Lora proved to adapt more when it comes to capturing and mimicking the Dilbert style, even outside the office setting, agreeing with the hypothesis that Lora would excel in capturing the style of Dilbert comics.

5 Responsible Research

The basis of effective scientific research rests upon not only the methods employed and the results obtained, but also the adherence to responsible and ethical scientific practices. In line with these principles, we have strictly followed the Netherlands Code of Conduct, which serves as a guiding framework for our work. Honesty is of utmost importance to us, as we keep the integrity of our research by accurately representing our procedures, results, and interpretations. Our commitment to scrupulousness means that we approach our research with attention to detail and accuracy. Transparency is another fundamental principle that governs our work, as we strive to provide clear and comprehensive documentation of our methods, data, and findings. Independence is highly valued, as we maintain autonomy in our research process, allowing for unbiased exploration and analysis. Responsibility guides our actions, as we are mindful of the potential impact of our research on society and strive to ensure ethical considerations are at the forefront of our work. In addition, we emphasize the importance of proper citation, acknowledging the contributions of previous studies and researchers. By upholding these principles and standards, we aim to make a meaningful contribution to the scientific community through a paper that demonstrates technical accuracy as well as ethical integrity.

6 Conclusions

Our paper aimed to produce a comics illustration synthesizer using the Stable Diffusion model. To fine-tune stable diffusion two techniques were chosen, Lora and Dreambooth, with the goal of enhancing the generation of Dilbert comic panels from text prompts. Through our evaluation, we found that Lora exhibited superior performance in terms of accuracy and overall quality compared to Dreambooth. The fine-tuned model demonstrated a remarkable ability to understand and reproduce the unique Dilbert comic style based on textual descriptions. One in two generated images by Lora produced a scene that perfectly matched the text description. In conclusion, the utilization of LoRA in the training process of stable

diffusion for Dilbert comic generation proved to be a promising approach. The training speed of LoRA enabled multiple training rounds, allowing for exploration of various parameters and datasets to identify the optimal configurations, ultimately enhancing the performance and quality of the model. By leveraging the capabilities of LoRA, this project achieved substantial advancements in automated comic synthesis, generating Dilbert comic panels from text prompts.

6.1 Future Work

There are several potential avenues for future work based on the findings and implications of this research. Firstly, exploring the impact of improved prompts on the results. Investigating how to design more effective prompts could further optimize the generation of Dilbert comic panels. In advanced text to image generation, prompts can go up to 50 words. The simplicity of Dilbert causes it to not need that much, however, more detailed prompts could still further improve results. Additionally, focusing on improving the quality and accuracy of the captioning process within the datasets could contribute to better alignment between textual descriptions and visual content. For example, if the speech bubbles were mentioned in the captioning process, they can be removed from the generated image resulting in the ability to control them and cleaner results.

Furthermore, this research represents the initial step towards automating comic generation. To progress further, future work can involve integrating text generation capabilities into the pipeline. By incorporating language generation models (LLMs), it becomes possible to generate the accompanying dialogue and captions for the generated Dilbert comic illustrations. This comprehensive approach would enable end-to-end comic synthesis, facilitating the creation of fully automated and contextually coherent comic panels.

A Results

A.1 Lora Results

Results produced using the stable diffusion model v.1.5 fine tuned using Lora.

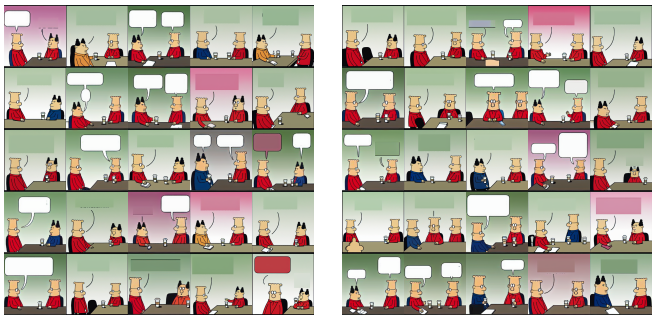

Figure 11: 50 images produced by Lora in experiment on prompt 1: ”Dilbert, Boss, sitting, table, coffee cup, office”

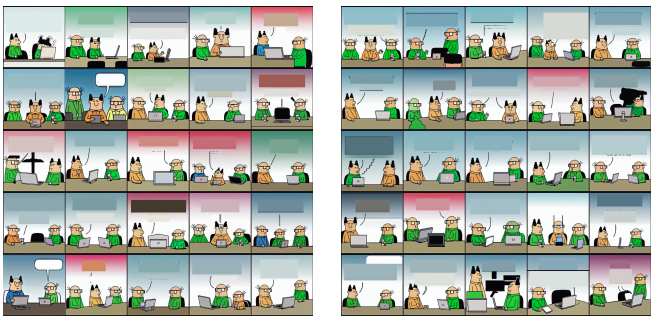

Figure 12: 50 images produced by Lora in experiment on prompt 2: ”Boss, Wally, Wally sitting, Boss sitting at table, table, laptop, office”

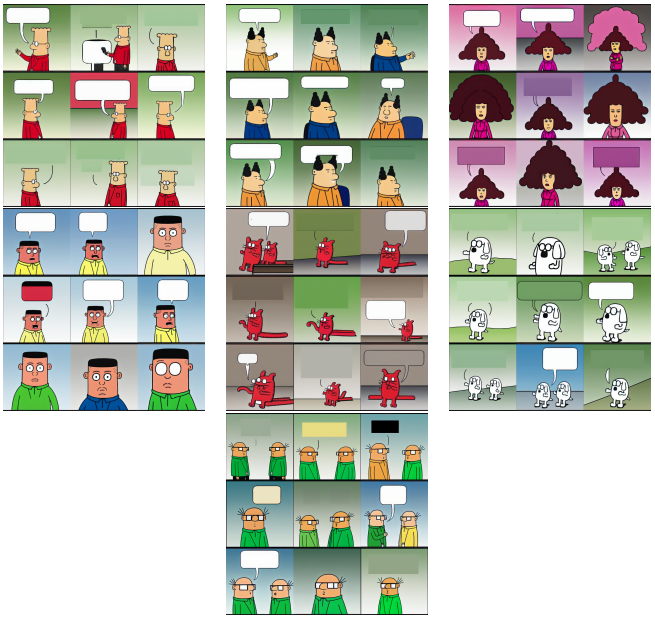

Figure 13: Lora results on one character, results score 100% accuracy.

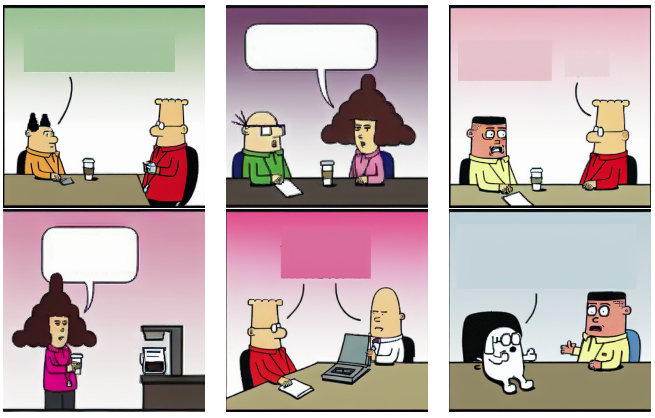

Figure 14: Lora fine tuning results for (a): Dilbert and Boss sitting, (b) Wally and Alice sitting, (c): Dilbert and Asok sitting, (d): Alice by the coffee machine, (e): Dilbert and CEO, (f): Asok and Dogbert

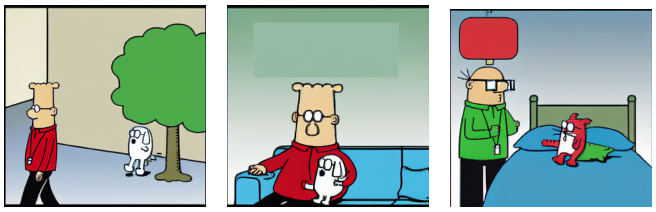

Figure 15: Lora results (a) Dilbert and Dogbert outdoors (b) Dilbert and Dogbert on a couch (c) Wally and Catbert on a bed

A.2 Dreambooth Results

Results produced using the stable diffusion model v.1.5 fine tuned using Dreambooth.

Figure 16: 50 images produced by Dreambooth in experiment on prompt 1: "Dilbert, Boss, sitting, table, coffee cup, office"

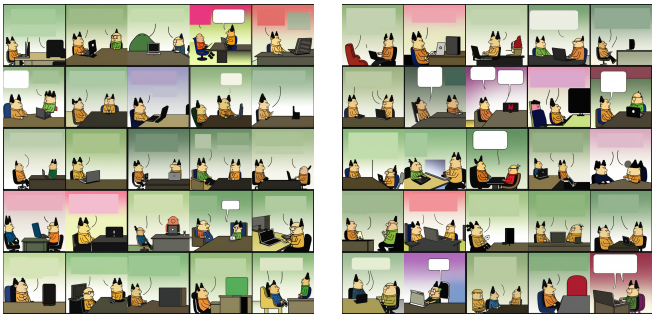

Figure 17: 50 images produced by Dreambooth in experiment on prompt 2: "Boss, Wally, Wally sitting, Boss sitting at table, table, laptop, office"

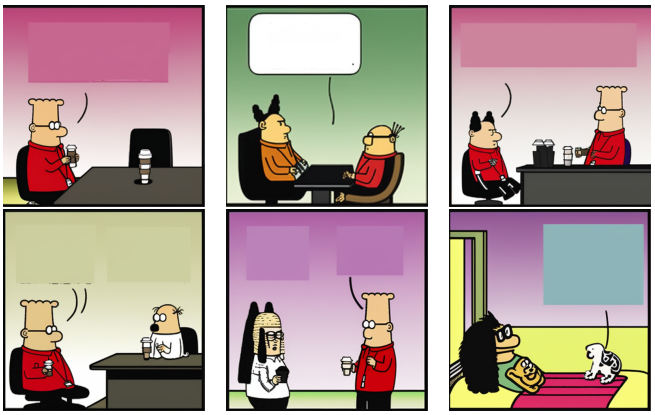

Figure 18: Dreambooth fine tuning results for (a): Dilbert, (b): Boss and Wally, (c): Dilbert and Boss sitting, (d): Dilbert and Dogbert, (e): Dilbert and Alice, (f): Wally and Dogbert in a bedroom.

References

[1] Wikipedia contributors. Dilbert — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Dilbert&oldid=1159915438, 2023. [Online; accessed 13-June-2023].

[2] Wikipedia contributors. List of dilbert characters — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=List_of_Dilbert_characters&oldid=1158387727, 2023. [Online; accessed 13-June-2023].

[3] Darwin Morris. Synthesizing comics via conditional generative adversarial networks, July 2021.

[4] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2023.

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[6] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[7] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022.

[8] Archive of dilbert comics, June 2023.

[9] Ben Proven-Bessel, Zilong Zhao, and Lydia Y. Chen. Comicgan: Text-to-comic generative adversarial network. *CoRR*, abs/2109.09120, 2021.

[10] Chris CSAle. Dreambooth model for the dilbert concept, 2023.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[12] Simo Ryu. Low-rank adaptation for fast text-to-image diffusion fine-tuning. https://github.com/cloneofsimo/lora, 2023.

[13] Simo Ryu, Andreas Jansson, Jesse Andrews, and Zeke Sikelianos. Introducing lora: A faster way to fine-tune stable diffusion. February 2023.

[14] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training

for unified vision-language understanding and generation, 2022.

[15] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August 2020. Version 0.3.0.