NANYANG TECHNOLOGICAL UNIVERSITY

COLLEGE OF COMPUTING AND DATA SCIENCE

# SC4062 Generative Artificial Intelligence - Advanced Topics

Group Assignment

Visual Content Generation

April 10, 2025

| No. | Name | Matriculation Number |
|---|---|---|
| 1 | **Chen Zihang** | U2121486H |
| 2 | **Li Zihan** | U2121598G |
| 3 | **Lye En Lih** | U2121387B |
| 4 | **Zhang Jing Wen** | U2121853G |

# 1 Introduction

## 1.1 Background and Motivation

Neural style transfer enables an image to adopt the artistic appearance of another by exploiting deep feature representations. Since the seminal work of Gatys et al. (2016), a variety of methods have tackled photo-to-art translation, ranging from conditional GANs such as Pix2Pix and CartoonGAN to attention-augmented UGATIT (Kim et al., 2020). However, existing GAN-based approaches can require large datasets with complex training, and still struggle with artifacts or loss of facial fidelity.

Diffusion models now offer a compelling alternative. Stable Diffusion, in particular, shows strong style understanding and high-resolution synthesis from text prompts, suggesting it could learn direct photo-to-comic translation with fewer data and more stable training. This project asks: **Can a pre-trained diffusion model be efficiently fine-tuned to convert a portrait photo into a convincing comic illustration while preserving identity?**

## 1.2 Problem Statement

Translating real portrait photographs into comic-style illustrations is a challenging task due to the significant visual differences between the source (photographic) and target (comic) domains, where identity and facial structures must be preserved while adapting textures, shading, and colors into a stylized, hand-drawn appearance. This project formulates the task as a supervised image-to-image translation problem using a paired dataset of facial photos and corresponding comic-style renderings with the following core objective:

*"Develop a model that transforms real portrait images into colored comic-style illustrations, preserving identity and artistic consistency."*

## 1.3 Project Objectives

Following the defined problem statement, the objectives of this project are:

- **Train a diffusion-based model for comic stylization:** Fine-tune Stable Diffusion v2.1 to translate real human face images into Western comic-style portraits while preserving identity.
- **Leverage LoRA for efficient fine-tuning:** Apply Low-Rank Adaptation to update only a subset of the model parameters, reducing training cost and memory usage.
- **Use supervised learning on paired data:** Train with aligned photo-comic pairs using a combination of pixel-wise and perceptual loss functions to ensure both accuracy and visual appeal.
- **Enable single-step inference:** Design an efficient pipeline that converts input photos to comic-style images in one pass without iterative refinement.
- **Constrain scope and evaluate effectively:** Focus on frontal face portraits and a single comic style, and evaluate results through both qualitative inspection and quantitative metrics.

Through these objectives, we introduce *ComicDiff*, a lightweight diffusion-based pipeline that generates identity-preserving, high-quality comic portraits, showcasing the effectiveness of LoRA fine-tuning for style transfer.

# 2 Literature Review

## 2.1 Introduction to Image-to-Image Translation and Stylization

Early image stylization methods relied on non-photorealistic rendering and hand-crafted rules, often struggling with natural scenes or human faces. Gatys et al. (2016) introduced *Neural Style Transfer* (NST) which uses deep CNN to separate and recombine image content and style, synthesizing impressive results. Later, Generative Adversial Networks (GANs) introduced frameworks for image-to-image translation, such as Pix2Pix (Isola et al., 2017) for paired image translations. For cartoon-related transformations, Chen et al. (2018) introduced CartoonGAN, which incorporates specialized losses for bold edges and smooth color fields, the key characteristics of cartoons.

## 2.2 Recent Advances in Diffusion Models for Stylization

Denoising diffusion probabilistic models (DDPMs) have emerged as powerful alternatives to GANs, by progressively denoising random noise into structured images (Ho et al., 2020). Among these, Latent diffusion models (LDMs), which operate in a learned latent space, greatly reduced computational overhead while preserving fine details (Rombach et al., 2022).

Stable Diffusion is an LDM that uses a Variational Autoencoder (VAE) and U-Net-based diffusion process, combined with cross-attention layers that enable text-guided image synthesis. For stylization tasks, these models can be adapted by fine-tuning small subsets of parameters or by editing the diffusion sampling process.

## 2.3 Low-Rank Adaptation (LoRA) for Efficient Fine-Tuning

Training large diffusion models like Stable Diffusion from scratch or fully fine-tuning them can be computationally expensive. Hu et al. (2021) proposed *Low-Rank Adaptation (LoRA)*, which inserts a small number of learnable parameters (in low-rank form) into the weight matrices of a pretrained model. By updating only these added parameters, LoRA allows the model to capture new styles without catastrophic forgetting of its general-purpose generation capabilities.

In recent work, LoRA modules have allowed Stable Diffusion models to learn new artistic styles rapdily with a fraction of the memory cost (Chen et al., 2024; Frenkel et al., 2024).

## 2.4 Paired Dataset Training versus Prompt-Based or Unpaired Transfer

A key advantage of our setup is the availability of paired real-to-comic portrait data from the "Comic Faces Paired Synthetic v2" dataset (Sxela, 2022) obtained from Kaggle. Unlike unpaired approaches (e.g., CycleGAN) that rely on cycle-consistency, or text-prompted methods that may not match the intended style, paired data directly shows the model how each real face should look in the stylized domain, leading to more precise identity preservation and style fidelity (Isola et al., 2017). Paired datasets have also historically yielded higher-quality translations in other tasks such as edges-to-photo or semantic-map-to-image tasks (Isola et al., 2017).

# 3 Methodology

## 3.1 Model Architecture

The proposed solution, **ComicDiff**, is built upon the Stable Diffusion v2.1 architecture, integrating Low-Rank Adaptation (LoRA) modules for efficient fine-tuning. Stable Diffusion is a latent diffusion model that generates images in a lower-dimensional latent space instead of directly in pixel space. Its architecture comprises three main components:

- **Variational Autoencoder (VAE)**: The encoder $\mathcal{E}$ compresses an input image $x$ into a latent representation $z = \mathcal{E}(x)$, and the decoder $\mathcal{D}$ reconstructs the image $\tilde{x} = \mathcal{D}(z)$. This latent representation (e.g., $64 \times 64 \times 4$ for a $512 \times 512$ image) captures essential semantic features. In ComicDiff, the VAE is kept frozen during training to preserve its ability to reconstruct facial structures accurately.
- **U-Net Denoiser**: The U-Net is a time-conditioned denoising network with skip connections and cross-attention modules. It learns to denoise latent representations by iteratively removing noise during the diffusion process. In ComicDiff, we use the pre-trained U-Net from Stable Diffusion v2.1 and fine-tune it for style transfer using LoRA.

### LoRA Fine-Tuning Process

To adapt the large-scale U-Net without fine-tuning all parameters, we employ LoRA to insert two trainable matrices $A$ and $B$ into selected linear layers (typically attention projections), and freeze the original weights $W$.

Instead of directly learning a full-rank update $\Delta W \in \mathbb{R}^{d\times k}$, LoRA approximates it using:

$$\Delta W = A \cdot B, \quad A \in \mathbb{R}^{d\times r}, \quad B \in \mathbb{R}^{r\times k}$$

where $d$ is the number of *output features* from the layer, $k$ is the number of *input features* to the layer, and $r \ll \min(d, k)$ is the *rank* of the adaptation.

The adapted weight is then calculated as:

$$W_{\text{adapted}} = W + \Delta W = W + A \cdot B$$

This low-rank formulation significantly reduces the number of trainable parameters and accelerates training. The rank $r$ controls the expressiveness of the adaptation:

- **Lower** $r$ reduces memory and training time but may limit expressiveness.
- **Higher** $r$ allows more flexibility but increases compute cost.

The benefits of LoRA in our model fine-tuning process include the following:

- **Preservation of pre-trained knowledge:** The original model weights remain frozen, preventing catastrophic forgetting of previously learned image representations.
- **Reduced resource usage:** Only a small subset of additional parameters are trained, enabling efficient fine-tuning even on limited hardware.
- **Faster convergence:** With fewer parameters to optimize and a strong starting point from the pre-trained model, training is more stable and requires fewer iterations.

The overall architecture remains consistent with Stable Diffusion v2.1. However, during fine-tuning, only the LoRA-inserted weights within the U-Net are updated. The VAE component remain untouched, ensuring that the model retains its core generative capabilities while learning to apply the desired stylistic transformation.

### 3.2 Dataset Preparation and Preprocessing

We trained the model on Kaggle's "Comic Faces Paired Synthetic v2" dataset, consisting of 10,000 pairs of real portraits and corresponding comic-style versions. All images are high resolution (up to 1024×1024 pixels) and algorithmically generated to simulate a consistent comic style that can be described as dark-outlined and reddish-toned.

The faces in the dataset are already aligned and cropped so that facial features between the real/comic pair will correspond pixel-to-pixel, and the model will not need to account for differences in poses. The dataset also covers a variety of identities (e.g. genders and ethnicities), which will help the model generalize for new faces.

Before feeding the data to the training pipeline, we perform the following pre-processing steps:

- Resolution standardization: The original 1024×1024 images are down-sampled to a fixed resolution of 768×768 pixels while preserving aspect ratio. This is the native resolution for SD2.1's latent model.
- Normalization: We normalize pixel values to the range $[0, 1]$ (or $[-1, 1]$ after a transformation) as expected by the VAE encoder. We apply the same normalization so that the encoded latents are compatible with the pre-trained latent space.
- Dataset split: We split the dataset into a training set and a validation set, with the training set containing 95% of the data and the validation set containing 5% of the data.

Through this preparation, we ensure the training data is well-aligned and normalized for the model to learn the photo-to-comic mapping. The high-quality paired data provides a strong supervision signal, which will be easier for the model to learn from compared to unpaired translation tasks.

## 3.3 Training Setup and Fine-Tuning Procedure

With the architecture defined, ComicDiff is trained in a supervised manner to stylize face images. The training fine-tunes only the LoRA-inserted parameters in the U-Net while keeping the rest of the Stable Diffusion model frozen.

### LoRA Fine-Tuning Configuration

LoRA modules are inserted into attention and selected feed-forward layers of the U-Net. We then proceed to set the rank $r = 8$ to control the dimensionality of the adaptation matrices $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times k}$, allowing the update to the original weight $W \in \mathbb{R}^{d\times k}$.

### Loss Functions

We train the model using a weighted combination of pixel-level and perceptual similarity losses:

- **Mean Squared Error (MSE):**

$$\mathcal{L}_{\text{MSE}} = \|\hat{y} - y\|^2$$

  Encourages the output to match the ground-truth image at the pixel level.

- **LPIPS (Learned Perceptual Image Patch Similarity):**

$$\mathcal{L}_{\text{LPIPS}} = \|f(\hat{y}) - f(y)\|^2$$

  Where $f(\cdot)$ denotes features from a pre-trained VGG network, this loss aligns higher-level visual features to improve stylistic realism.

The total loss is given by:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MSE}} + \lambda \cdot \mathcal{L}_{\text{LPIPS}}$$

where $\lambda$ is a weighting factor that balances the contribution of perceptual loss relative to the MSE loss. In our case, we set $\lambda = 0.8$ to place slightly more emphasis on perceptual similarity while still maintaining structural accuracy. The combination encourages outputs that are not only visually similar in appearance and style but also maintain fine facial details.

### Training Procedure

Given a training pair of real and comic images $(x, y)$:

1. **Encode to Latents:** Use the frozen VAE encoder $\mathcal{E}$ to get latents $z_x = \mathcal{E}(x)$ and $z_y = \mathcal{E}(y)$.
2. **Latent Translation:** Feed $z_x$ to the U-Net (with LoRA weights), conditioned on a text embedding. The U-Net outputs the predicted comic-style latent $\hat{z}_y$.
3. **Decode Output:** Use the VAE decoder to reconstruct the stylized image $\hat{y} = \mathcal{D}(\hat{z}_y)$.

4. **Compute Loss:** Calculate $\mathcal{L}_{\text{MSE}}$ by comparing pixel values between $\hat{y}$ and $y$ and $\mathcal{L}_{\text{LPIPS}}$ by passing both through the pre-trained VGG neural network to compute feature-wise distances between $\hat{y}$ and $y$.

5. **Backpropagation:** We backpropagate the loss and update the LoRA parameters in the U-Net by using the AdamW optimizer, while keeping all other weights frozen. This allows the LoRA matrices to reduce the loss for this pair.

This training process is repeated for every batch over multiple iterations during each epoch, as part of the gradient descent optimization.

### Hyperparameters and Optimization

Training was conducted over 5 epochs with a batch size of 4 and a learning rate of $1 \times 10^{-5}$ for the LoRA parameters. The training dataset consisted of 9500 aligned image pairs of real and comic-style faces. Optimization was performed using the AdamW optimizer, and automatic mixed-precision training was enabled via PyTorch's `GradScaler` to accelerate training and reduce memory usage.

Each training step involves encoding input pairs using the VAE, predicting stylized latents with the U-Net, decoding the outputs, and calculating the combined loss. Loss values were logged per step and plotted across epochs to monitor convergence.

Unlike many fine-tuning pipelines that implement early stopping based on validation loss trends, we deliberately chose to train for a fixed number of epochs. This decision allows the LoRA parameters to fully utilize the capacity of the dataset and learn the stylistic transformation across a wide variety of examples, especially in early-stage exploratory training.

Fixed-length training is particularly useful when:

- Validation performance is noisy or insufficiently representative of overall generalization.
- Fast convergence is expected due to pre-trained weights, and overfitting risk is minimal due to the small number of trainable parameters.
- Visual inspection and downstream perceptual quality are prioritized over pure loss metrics.

This design choice allowed for a controlled and consistent comparison across different configurations (e.g., LoRA ranks or loss weights), as the training duration remained constant.

### 3.4 Inference Pipeline

After training, the ComicDiff model can convert a real portrait into a comic-style image using an inference process. This section outlines the deployed inference pipeline implemented in PyTorch.

1. **Input preprocessing:** A real face image is first loaded and resized to 768×768 to match the training resolution. The image is normalized to the $[-1, 1]$ range and converted to a PyTorch tensor.

2. **Encoding to latent:** The input image tensor is passed through the frozen Variational Autoencoder (VAE) encoder $\mathcal{E}$ from Stable Diffusion to obtain a latent representation $z_x = \mathcal{E}(x)$. The resulting latents are scaled by a constant factor (0.18215), consistent with the model's pre-training setup.

3. **Latent transformation with U-Net:** The encoded latent $z_x$ is processed through the fine-tuned U-Net with LoRA-inserted layers. A single diffusion step is executed with $t = 1$, and no iterative denoising is performed. The text embedding is set to match the conditioning prompt used during training. The U-Net produces a residual output, which is scaled and added to the input latent:

   $$\hat{z}_y = z_x + \alpha \cdot \Delta z$$

   where $\alpha$ is the LoRA residual blending factor. This approximates a learned transformation from real to comic-style latents.

4. **Decoding the result:** The stylized latent $\hat{z}_y$ is then passed through the frozen VAE decoder $\mathcal{D}$ to produce the final output image $\hat{y} = \mathcal{D}(\hat{z}_y)$. The output is de-normalized back to the $[0, 1]$ pixel range and saved to disk as a comic-style portrait.

Unlike conventional Stable Diffusion usage, which involves dozens of iterative sampling steps, this method completes the stylization in a single forward pass, allowing for near real-time image transformation. The model consistently produces the same stylized output when given the same input image and prompt. This determinism is especially beneficial in applications where reliable and repeatable results are important.

In summary, the ComicDiff inference pipeline performs an efficient, latent translation from real faces to comic stylization. It leverages the pre-trained VAE for encoding/decoding and a LoRA-enhanced U-Net for style transformation, enabling fast, consistent, and high-quality image stylization suitable for deployment in interactive applications.

# 4 Experimental Results

## 4.1 Quantitative Evaluation Results

To assess the quality of the generated comic-style portraits, we evaluated the outputs using four standard image similarity metrics on a held-out test set of 500 paired samples. These metrics were chosen to capture both pixel-level accuracy and perceptual similarity:

- **Structural Similarity Index (SSIM)**: Measures the structural similarity between two images by comparing luminance, contrast, and texture. A higher SSIM indicates better preservation of structural details.
- **Peak Signal-to-Noise Ratio (PSNR)**: Evaluates the pixel-wise fidelity between the generated image and the ground truth. A higher PSNR generally suggests fewer differences at the pixel level.
- **Mean Squared Error (MSE)**: Computes the average squared difference between the predicted and ground truth pixel values. Lower MSE values indicate higher pixel-level accuracy.
- **Learned Perceptual Image Patch Similarity (LPIPS)**: Uses deep neural network features (e.g., VGG) to measure perceptual similarity. Lower LPIPS values mean that the predicted image is perceptually closer to the target, aligning more closely with human visual judgment.

| **Metric** | **Average Score** |
|---|---|
| Structural Similarity Index (SSIM) | 0.539 |
| Peak Signal-to-Noise Ratio (PSNR) | 8.934 dB |
| Mean Squared Error (MSE) | 0.129 |
| LPIPS (VGG) | 0.300 |

Table 1: Quantitative evaluation metrics computed over 500 test image pairs between generated comic-style images and their corresponding ground-truth targets.

The average SSIM of 0.539 indicates moderate structural similarity between the generated images and ground-truth comics, suggesting that the model preserves key spatial features while adapting to the comic style. A PSNR of 8.934 dB and MSE of 0.129 reflect low pixel-level fidelity, which is expected for stylization tasks that alter color, texture, and shading.

The LPIPS score of 0.300, computed using a VGG-based perceptual network, indicates that the outputs remain perceptually close to the targets. As LPIPS aligns well with human perception, this supports the observation that the model produces visually appealing, identity-preserving comic portraits.

### 4.2 Qualitative Evaluation Results

To assess the visual quality of the stylized outputs produced by our model, we conduct a qualitative comparison between the original real-life images, their corresponding ground truth comic illustrations, and the outputs generated by our fine-tuned model.

The final comic-style outputs are generated using the proposed inference pipeline (Section 3.4).

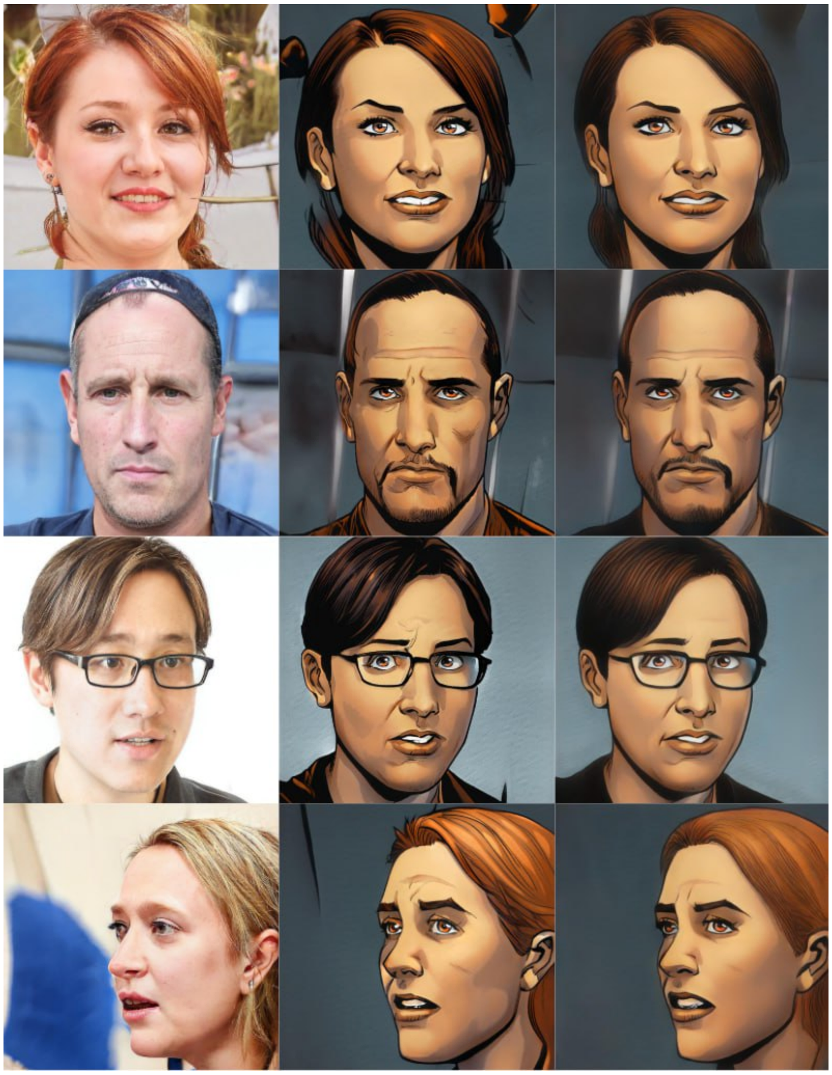

Figure 1: Visual comparison between real images (left), ground truth comic-style targets (middle), and outputs generated by the fine-tuned ComicDiff model (right).

As illustrated in Figure 1, the ComicDiff outputs effectively replicate the visual characteristics of the ground truth comic images, including color tones, contour stylization, and expression consistency, while maintaining the core identity features of the original photographs.

The results confirm that our model generalizes well across various facial structures and lighting conditions. Furthermore, the deterministic nature of the inference pipeline ensures consistent outputs for the same input image, making the system suitable for real-world applications such as avatar creation or stylized profile generation.

## 5 Future Work

Although ComicDiff has shown promising results in stylizing portraits into comic-style illustrations, there remains room for further improvement. A key direction is diversifying the training dataset to include more varied poses and camera orientations, which would improve generalization across real-world inputs.

Additionally, expanding the pipeline to support multiple comic styles by altering contour thickness, color palettes, and texture stylizations could offer greater flexibility and user control. Future work may also explore lightweight model variants or quantization for real-time deployment on edge devices.

Finally, methods that reduce dependence on paired training data, such as semi-supervised or unpaired image translation, can broaden the training scope. Enhanced user interaction through sketch guidance or reference-based control could further enrich the model's creative potential to be able to generate higher quality and accurate comics images.

## 6 Conclusion

In this project, we proposed **ComicDiff**, a lightweight and efficient pipeline for transforming real-life portraits into stylized comic illustrations. By fine-tuning the Stable Diffusion v2.1 U-Net with Low-Rank Adaptation (LoRA), we enabled the model to perform high-fidelity image-to-image translation while preserving identity features. The training leveraged paired image datasets and a combined loss of MSE and LPIPS to balance structural accuracy and perceptual quality.

Our inference pipeline produces stylized outputs in a single forward pass, requiring no iterative sampling or noise injection. This makes the system highly efficient and suitable for deployment in real-world applications where rapid and consistent stylization is desired.

Qualitative evaluations demonstrated that ComicDiff successfully captures the stylistic characteristics of comic images while maintaining realism and consistency.

# References

Chen, B., Li, K., Feng, Y., and Zhao, X. (2024). Consislora: Consistency-aware low-rank adaptation for stable diffusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–10.

Chen, Y., Lai, Y.-K., and Liu, Y. (2018). Cartoongan: Generative adversarial networks for photo cartoonization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9465–9474.

Frenkel, R., Ortiz, C., and Tyszkiewicz, M. (2024). B-lora: Bi-component low-rank adaptation for content and style disentanglement in diffusion models. In *European Conference on Computer Vision*, pages 210–225.

Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, pages 1–14.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134.

Kim, J., Kim, M., Kang, H., and Lee, K. H. (2020). U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In *International Conference on Learning Representations*.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Sxela (2022). Comic faces paired synthetic v2 [data set]. https://www.kaggle.com/datasets/defileroff/comic-faces-paired-synthetic-v2. Kaggle.