International Journal of Computer Applications Technology and Research
Volume 15–Issue 05, 01 – 06, 2026, ISSN: -2319–8656
DOI:10.7753/IJCATR1505.1001

Text-Guided Face Generation Using LoRA Fine-Tuned Stable Diffusion on the CelebA Dataset

Ceena Mathews
Department of Computer Science
Prajyoti Niketan College, Kerala, India

Abstract: This paper presents a text-conditioned face generation framework that fine-tunes a pretrained diffusion model using Low-Rank Adaptation (LoRA). The proposed method uses the CelebA dataset, in which facial attributes are converted into descriptive captions to enable controlled image synthesis. Stable Diffusion v1.5 is fine-tuned using LoRA to achieve efficient domain adaptation with reduced computational overhead. The model is evaluated using CLIPScore to measure semantic alignment and Fréchet Inception Distance (FID) to assess visual realism. Experimental results show that the model achieves a mean CLIPScore of approximately 0.32, indicating strong correspondence between textual prompts and generated images. However, the FID score of 107 suggests a gap in distributional similarity with real images. The study highlights the trade-off between semantic alignment and realism, as well as the importance of training dynamics in diffusion-based fine-tuning.

Keywords: Text-to-Image Generation, Stable Diffusion, LoRA, CelebA, CLIPScore, Face Synthesis, FID

1. INTRODUCTION

Recent advancements in deep generative models have significantly improved the synthesis of realistic human faces. Generative Adversarial Networks (GANs), particularly StyleGAN, have demonstrated remarkable capabilities in producing high-quality images. However, these models lack effective mechanisms for controlling generation using natural language descriptions.

Text-to-image generation has gained significant attention with the introduction of multimodal learning frameworks. Diffusion models, particularly Stable Diffusion, have emerged as powerful tools for generating high-quality images conditioned on textual input. These models leverage iterative denoising processes and pretrained text encoders to achieve strong semantic alignment.

Despite their effectiveness, fine-tuning diffusion models for domain-specific tasks such as face generation remains computationally expensive. Low-Rank Adaptation (LoRA) offers a parameter-efficient alternative by introducing low-rank updates to pretrained weights, enabling efficient model adaptation with minimal resource requirements.

In this work, an efficient and interpretable framework for text-guided face generation is implemented by integrating LoRA-based fine-tuning with a pretrained latent diffusion model. Unlike conventional approaches that require full model retraining, the use of low-rank adaptation enables parameter-efficient learning while preserving the generalisation capabilities of the pretrained model. Furthermore, the incorporation of text conditioning through cross-attention mechanisms allows precise control over generated facial attributes, facilitating semantically meaningful synthesis. The framework is specifically tailored to leverage the diversity of the CelebA dataset, enabling the generation of high-quality and attribute-consistent face images. This combination of efficiency, controllability, and high-fidelity generation distinguishes the proposed approach from existing methods and makes it suitable for scalable real-world applications. The objective is to achieve controllable synthesis while maintaining computational efficiency.

2. RELATED WORK

Text-to-face generation is a specialised sub-domain of text-to-image synthesis that focuses on generating

realistic human faces conditioned on natural language descriptions.

Early research in text-to-face generation primarily relied on GAN-based architectures. Conditional GANs introduced the concept of generating images conditioned on auxiliary information such as text. Text2FaceGAN [1] demonstrated the feasibility of generating faces from attribute-derived textual descriptions using the CelebA dataset. However, this approach was limited by low resolution and weak semantic alignment.

Subsequent GAN-based models, such as StyleGAN2 [2], significantly improved image quality and resolution but lacked direct text conditioning capabilities. Methods such as attribute disentanglement [3] and transformer-based GANs [4] attempted to bridge this gap by incorporating semantic control. While these approaches improved realism and attribute control, they suffered from training instability and limited scalability.

Diffusion models have recently emerged as a robust alternative for image synthesis. Denoising Diffusion Probabilistic Models (DDPM) [5] introduced a probabilistic framework for iterative image generation. Stable Diffusion [6] extended this concept to latent space, enabling high-resolution image generation with efficient computation. These models integrate CLIP-based text encoders [7], allowing strong semantic alignment between text and images.

Fine-tuning approaches such as DreamBooth [8] and LoRA [9] have been proposed to adapt large diffusion models to specific domains. DreamBooth enables personalised generation but requires extensive computational resources. In contrast, LoRA provides a lightweight alternative by updating only low-rank matrices, making it suitable for domain-specific tasks.

Despite these advancements, limited work has explored LoRA-based fine-tuning for text-guided face generation using structured datasets such as CelebA. Additionally, most prior works emphasize FID as the primary evaluation metric, with limited focus on semantic alignment metrics such as CLIPScore.

3. MATERIALS AND METHODS

3.1 Dataset

The CelebA dataset [10] is used in this study. It contains over 200,000 face images annotated with 40 binary attributes, including gender, age, facial expression, accessories, and hair characteristics.

To enable text-based conditioning, attribute labels are converted into natural language captions using a rule-based approach. For example:*"A portrait photo of a smiling young woman with black hair and eyeglasses."*

Images are preprocessed by cropping, resizing to 512×512 resolution, and normalization. A subset of 100000 images is used for training to balance computational efficiency and model performance.

3.2 Methods

The proposed framework is based on Stable Diffusion v1.5, a latent diffusion model designed for high-quality text-to-image generation. The architecture consists of three main components: a Variational Autoencoder (VAE), a U-Net based denoising network, and a CLIP-based text encoder. The VAE is responsible for encoding input images into a latent space and decoding them back into pixel space, thereby reducing computational complexity. The U-Net network performs the core denoising process by iteratively refining noisy latent representations. The text encoder, derived from CLIP, converts input textual prompts into semantic embeddings that guide the image generation process. Figure 1 illustrates the overall architecture of the proposed framework for text-guided face generation using LoRA fine-tuned stable diffusion on the CelebA dataset.

Low-Rank Adaptation (LoRA)

To enable efficient domain adaptation, LoRA is employed for fine-tuning the pretrained diffusion model. Instead of updating the full set of model parameters, LoRA introduces trainable low-rank matrices to approximate weight updates. The modified weight matrix can be expressed as:

$$W' = W + \Delta W \quad (1)$$

where the update term is defined as:

$$\Delta W = A \; x \; B$$

where A and B are low-rank matrices with significantly fewer parameters compared to the original weight matrix W. This approach reduces memory consumption and computational cost while maintaining the model's ability to adapt to new data. In this work, LoRA is applied to the U-Net layers of the diffusion model, allowing efficient fine-tuning for face generation tasks using limited training data. The model is trained on the CelebA dataset using attribute-derived textual captions. Each image is paired with a corresponding textual description to enable supervised text-conditioned learning. The training process is conducted for 5000 steps with a learning rate of 5×10^{-5}. A batch size of 1 is used due to memory constraints, with gradient accumulation employed to stabilize training. All images are resized to a resolution of 512×512 pixels to match the input requirements of the stable diffusion model.

3.3 Evaluation Metrics

The performance of the proposed model is evaluated using CLIPScore [11] and FID [12]. CLIPScore measures the semantic similarity between general images and their corresponding textual prompts. CLIPScore is computed using cosine similarity between image and text embeddings extracted from a pretrained CLIP model. A higher CLIPScore indicates stronger alignment between the generated image and the input description. This metric is particularly suitable for text-conditioned generation tasks where semantic correctness is more critical than distributional similarity.

FID evaluates the similarity between distributions of real and generated images, measuring visual realism.

4. RESULTS AND DISCUSSION

The proposed framework consists of four major stages: input, text prompt encoding, diffusion-based image generation with LoRA adaptation, and output face synthesis.

In the first stage, images from the CelebA dataset are used as the training data. This dataset contains a large collection of face images with diverse identities and attributes, enabling the model to learn rich facial features. In the second stage, a textual description provided by the user is processed using a text encoder based on the CLIP model. The text encoder converts the input prompt into a high-dimensional embedding that captures the semantic information required for conditional image generation.

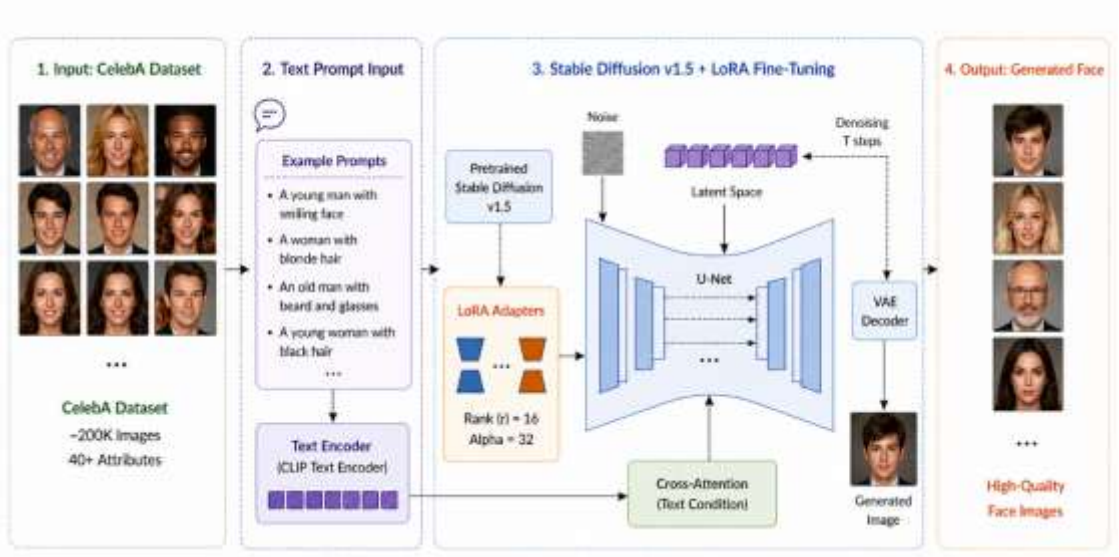

Figure 1. Proposed framework for text-guided face generation using LoRA fine-tuned Stable Diffusion.

In the third stage, the pretrained Stable Diffusion v1.5 model is fine-tuned using LoRA. LoRA introduces low-rank trainable matrices into selected layers of the model, allowing efficient adaptation without modifying the full set of parameters. During generation, random noise is sampled in the latent space and iteratively refined through a denoising process using a U-Net architecture. The encoded text features are incorporated through cross-attention mechanisms, enabling the model to generate images consistent with the given textual description. The final latent representation is then passed through a VAE decoder to reconstruct the generated face image. In the final stage, the model produces high-quality face images that align with the input text prompt. The generated outputs demonstrate the ability of the proposed framework to synthesize realistic and semantically consistent faces while maintaining computational efficiency through LoRA-based fine-tuning.

Figure 2 shows the performance of the model in generating human face images. The model achieves a maximum mean CLIPScore of **0.3214**, indicating strong semantic alignment between input prompts and generated images. The model successfully captures prominent attributes specified in the text, including hair color (e.g., blond, black, and gray), the presence of eyeglasses, and facial expressions such as smiling. These results validate the ability of the proposed framework to learn meaningful associations between textual descriptions and visual representations.

Despite these strengths, the model exhibits variability in capturing subtle or complex attributes. Features such as age-related characteristics, fine skin texture, and intricate facial details are less consistently represented across generated samples. This limitation can be attributed to the relatively small training dataset and the simplicity of the attribute-to-caption conversion process, which may not fully capture nuanced semantic variations.

A detailed analysis of the training behavior reveals that the model's performance is highly sensitive to the number of training iterations. The CLIPScore improves during the early stages of training, reflecting progressive learning of text–image relationships. However, beyond a certain point, performance begins to decline, indicating the onset of overfitting. This phenomenon suggests that the model gradually shifts from learning generalised representations to memorising training-specific patterns, thereby reducing its ability to generalise to unseen prompts. Consequently, selecting an appropriate intermediate training stage is essential for achieving optimal performance.

However, the FID score of 107 indicates a noticeable gap between the distribution of generated images and real images. This suggests that while the model performs well in aligning with textual descriptions, it does not fully capture the underlying distribution of real-world face images. This limitation can be attributed to factors such as the reduced training dataset size, simplified caption generation process, and the inherent trade-off in diffusion models between semantic conditioning and distributional accuracy

An important observation from this study is the balance between computational efficiency and performance. The use of LoRA enables parameter-efficient fine-tuning by updating only a small subset of model parameters, thereby reducing memory requirements and training time. Despite this lightweight adaptation, the model achieves strong semantic alignment, demonstrating the effectiveness of LoRA for domain-specific customisation of large diffusion models.

Overall, the results confirm that the proposed framework provides a practical and efficient solution for text-guided face generation. While the model performs well in capturing dominant attributes and maintaining visual consistency, further improvements can be achieved by enhancing caption quality, increasing dataset diversity, and incorporating more sophisticated conditioning mechanisms. These directions offer promising opportunities for improving both the robustness and generalisation capability of text-to-face generation systems.

Figure 2. Text-guided face generation results using LoRA fine-tuned Stable Diffusion.

5. CONCLUSION

This study presented an efficient framework for text-guided face generation using LoRA fine-tuned Stable Diffusion on the CelebA dataset. By integrating Low-Rank Adaptation with a pretrained diffusion model, the proposed approach enables parameter-efficient fine-tuning while maintaining the generative capabilities of the base model. The incorporation of text conditioning through CLIP-based embeddings allows the model to generate facial images that are semantically aligned with user-provided descriptions.

The experimental results demonstrate that the proposed method achieves a mean CLIPScore of approximately 0.32, indicating strong correspondence between textual prompts and generated images. At the same time, the relatively high FID score highlights limitations in capturing the full distribution of real-world facial images, reflecting a trade-off between semantic alignment and visual realism. The analysis of training behavior further reveals the impact of overfitting at later stages, emphasizing the importance of appropriate training duration and checkpoint selection.

Overall, the findings confirm that LoRA-based fine-tuning provides a practical and computationally efficient solution for domain-specific adaptation of diffusion models. The proposed framework successfully balances controllability, efficiency, and interpretability, making it suitable for applications such as virtual avatar generation, data augmentation, and human–computer interaction.

Future work can focus on improving realism by incorporating larger and more diverse datasets, enhancing caption generation using advanced language models, and exploring hybrid conditioning mechanisms. Additionally, integrating automatic selection strategies such as CLIP-guided candidate ranking could further improve output quality and consistency, providing a promising direction for advancing text-to-face generation systems.

6. REFERENCES

[1] Nasir, O. R., et al. (2019). *Text2FaceGAN: Face generation from fine-grained textual descriptions*. In Proceedings of the IEEE International Conference on Multimedia & Expo Workshops (IEEE BigMM).

[2] Karras, T., Laine, S., & Aila, T. (2020). *Analyzing and improving the image quality of StyleGAN*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (arXiv:1912.04958)

[3] Wang, T., Zhu, L., Yang, Y., & Liu, J. (2021). *Faces à la carte: Text-to-face generation via attribute disentanglement*. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV).

[4] Gholamrezaie, F., et al. (2022). *Cycle Text2Face: Cycle generative adversarial network via transformers*. arXiv preprint arXiv:2206.04503.

[5] Ho, J., Jain, A., & Abbeel, P. (2020). *Denoising diffusion probabilistic models*. In Advances in Neural Information Processing Systems (NeurIPS). (arXiv:2006.11239)

[6] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). *High-resolution image synthesis with latent diffusion models*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[7] Radford, A., Kim, J. W., Hallacy, C., et al. (2021). *Learning transferable visual models from natural language supervision*. In Proceedings of the International Conference on Machine Learning (ICML). (arXiv:2103.00020)

[8] Gal, R., Alaluf, Y., Atzmon, Y., et al. (2022). *DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation*. arXiv preprint arXiv:2208.12242.

[9] Hu, E. J., Shen, Y., Wallis, P., et al. (2022). *LoRA: Low-rank adaptation of large language models*. In the International Conference on Learning Representations (ICLR). (arXiv:2106.09685).

[10] Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). *Deep learning face attributes in the wild*. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).

[11] Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., & Choi, Y. (2021). *CLIPScore: A reference-free evaluation metric for image captioning*. arXiv preprint arXiv:2104.08718.

[12] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). *GANs trained by a two time-scale update rule converge to a local Nash equilibrium*. Advances in Neural Information Processing Systems.

.