

Mukh-Oboyob: Stable Diffusion and BanglaBERT Enhanced Bangla Text-to-Face Synthesis

Aloke Kumar Saha, Noor Mairukh Khan Arnob*, Nakiba Nuren Rahman, Maria Haque,
Shah Murtaza Rashid Al Masud, Rashik Rahman*
Department of CSE, UAP, Dhaka, Bangladesh

Abstract—Facial image generation from textual generation is one of the most complicated tasks within the broader topic of Text-to-Image (TTI) synthesis. It is relevant in several fields of scientific research, cartoon and animation development, online marketing, game development, etc. There have been extensive studies on Text-to-Face (TTF) synthesis in the English language. However, the amount of relevant existing work in Bangla is limited and not comprehensive. As the TTF field is not vastly prospected for Bangla language, the objective of this study sets forth to explore the possibilities in the field of Bangla Natural Language Processing and Computer Vision. In this paper, a novel system for generating highly detailed facial images from textual descriptions in the Bangla language is proposed. The proposed system named Mukh-Oboyob consists of two essential components: a pre-trained language model, BanglaBERT, and Stable Diffusion. BanglaBERT, a transformer-based pre-trained text encoder, is a language model used to transform Bangla sentences into vector representations. Stable Diffusion is used by Mukh-Oboyob to generate facial images utilizing the text embedding of the Bangla sentences. Moreover, the work utilizes CelebA Bangla, a modified version of the CelebA dataset consisting of face images, Bangla facial attributes, and Bangla text descriptions to develop and train the proposed system. This paper establishes a system for image synthesis with excellent performance and detailed image outcomes, as evidenced by a comprehensive analysis incorporating both qualitative and quantitative measures, leading to the system under consideration achieving an impressive FID score of 34.6828 and an LPIPS score of 0.4541.

Keywords—Bangla text-to-face synthesis; Natural Language Processing (NLP); Bangla NLP; Computer Vision (CV); Generative Model; stable diffusion; BanglaBERT

I. INTRODUCTION

Diffusion models have emerged to be a useful tool for generating realistic images in a variety of domains, such as human faces and natural landscapes. The ability to generate high-quality images from textual representations has attracted a lot of attention because of its possible applications in the development of content, augmented reality, and customized advertising.

Text-to-image generation is an approach to generating a picture from a given textual input. TTF is a subsection of TTI generation in which a human face description is provided and a facial image is generated based on the description. Compared to text-to-image generation, creating images of faces is a more difficult piece of work considering the complexity of facial features. TTF synthesis has plenty of applications that could be used, including internet marketing, animation, development of games, forensic science, and the metaverse. A significant

amount of literature has been penned about the creation of faces and images from text in recent years, as this field of study has grown in prominence. Interestingly, a significant proportion of academics have focused on image creation in the English language [1].

While substantial progress has been made in the area of text-to-image synthesis based on the English language, there is a lack of advanced and enhanced research concerning non-English languages, particularly Bangla. The Bangla language presents unique challenges to the synthesis of TTF due to its unique linguistic and cultural nuances. Facial image generation from Bangla text requires an extensive knowledge of the phonological, syntactic, and semantic structures of the language. To create authentic and culturally relevant facial portrayals, it is essential to accurately capture the visual diversity and unique facial features of Bangla-speaking people.

GAN (Generative Adversarial Network)-based models used for text-to-image synthesis face unstable training, mode collapse and non-convergence intrinsically due to adversarial training [2]. Diffusion models [3] are more capable of synthesizing realistic images compared to GANs as they seldom fall into such issues, thanks to a more stable training process. Vector quantized diffusion models produce better results compared to GAN-based models using diffusion strategy to avoid error assembling for image synthesis. Moreover, It achieves improved image generation speed while maintaining excellent image quality [4].

In the field of text to face synthesis, there is a lack of extensive research for Bangla language. Therefore, this paper proposes a novel Diffusion-based system, Mukh-Oboyob, specifically to generate face images from Bangla textual input to progress TTF generation for the Bangla language. The objective of the proposed system, Mukh-Oboyob, is to mitigate an existing void in the domain of TTF generation by addressing the difficulties associated with generating images with varying structures, that differ in appearance, and level of detail while upholding the realism of the images generated from Bangla descriptions that will significantly contribute to the advancement of the field of Bangla natural language processing.

The suggested system, Mukh-Oboyob, consists of two major parts: a pre-trained language model and a latent diffusion-based model, namely, BanglaBERT and Stable Diffusion respectively. BanglaBERT [5] is utilized to learn bi-directional contexts from Bangla sentences and extract semantic information essential to the text by encoding Bangla descriptions into vector representations and performing transformations over them in order to extract contextual information. Following

that, a Stable Diffusion [3] model which is used to synthesize images from text descriptions. The model is trained and evaluated following a modified version of the CelebA dataset [6] called CelebA Bangla [7]. This dataset incorporates 40 facial attributes derived from a semantically accurate Bangla vocabulary and includes a collection of facial images aligning with the 40 corresponding facial attributes. The CelebA Bangla dataset follows a novel algorithm [7] to create Bangla textual descriptions of facial images. The evaluation of the quality, diversity, and accuracy of the generated facial images is carried out by comprehensive testing and the application of both quantitative and qualitative evaluation metrics. The system under consideration achieved an FID (Fréchet Inception Distance) score of 34.6828 and an LPIPS (Learned Perceptual Image Patch Similarity) score of 0.4541.

The following sections of the paper are constructed as follows: Section II contains the literature review. The dataset is described in Section III. Section IV discusses the methodology followed by the system. Result Analysis is elaborated in the Section V. Section VI takes through the discussion whereas Section VII states the limitations of this work. Finally, Section VIII draws the conclusion and the references are added at the end of the paper.

II. LITERATURE REVIEW

In this section, significant studies utilizing generative models in the field of TTI and TTF synthesis are presented.

A. Text to Image Generation

This section provides a concise overview of a few methods that are notable and have come across to achieve impressive results for text-to-image synthesis.

Reed *et al.* [8] suggested a method of translating single-sentenced text descriptions directly into image pixels by introducing a deep convolutional GAN based on text description embedding compressed using a fully connected layer and leaky-ReLU activation and text features used to perform feed-forward inference by the generator and discriminator network fundamentally demonstrating enhanced text to image synthesis. In Paper [9], they proposed the use of the Bangla Attentional Generative Adversarial Network (AttnGAN) to generate high-quality images from texts through multi-staged processing and incorporation of specific details in distinct parts of images, achieving an enhanced inception score on the CUB dataset.

Naveen *et al.* [10] examined the combination of various Transformer models and the Attentional GAN (AttnGAN) to create the AttnGANTRANS architecture to generate images from texts and validates the effectiveness of the Transformer models by assessing the performance of generated images using the Fréchet Inception Distance and Inception Score evaluation metrics. In another work [11], the Cross-Modal Contrastive GAN (XMC-GAN) for text-to-image synthesis was proposed which generates images that are well-aligned with the texts and accomplish significant improvements in image quality utilizing multiple contrastive losses. Tao *et al.* [12] proposed a novel Deep Fusion GAN that generates high-quality images through a one-stage architecture and uses a deep fusion block which helps to integrate text and visual characteristics entirely. In another work by Siddharth *et al.*

[13], a combination of a GAN-based model and pre-trained text encoder, Attentional GAN and ROBERTa, respectively, were used, which resulted in a significant decrease in the FID score. In paper [14], they used diffusion models for image synthesis contextual to natural language descriptions and compared CLIP and classifier-free guidance as guidance strategies. Saharia *et al.* [15], presented a text-to-image diffusion model named Imagen that achieved a high level of photorealism and text and image alignment, leading towards deep language understanding by using diffusion models along with large transformer language models for text understanding.

B. Text to Face Generation

This section presents a comprehensive summary of various methods that have garnered attention and demonstrated remarkable outcomes in the field of TTF synthesis.

Deorukhkar *et al.* [16] employed three GAN-based architectures, DCGAN, DFGAN, and SAGAN for TTF synthesis. In this paper they used the CelebA dataset [6] consisting of celebrity images, Sentence BERT for sentence embeddings to encode the textual descriptions of the images from the dataset and compared the results of the three models using IS and FID evaluation metrics. In another work [17], a GAN-based two-stream architecture was introduced to generate images with great quality and diversity. They extracted features from the images through a Contrastive Language-Image Pre-training encoder and used Cross-Modal Distillation to align the image and text features. Xia *et al.* [18] suggested a novel GAN architecture, TediGAN, which generates multi-modal images from text descriptions and proposed a Multi-Modal CelebA-HQ dataset to evaluate the result and achieve the FID score. Ayanthi *et al.* [19], used StyleGAN2 to generate visually impressive facial images with accurate rendering of the facial features and utilized BERT for text embedding to generate high-quality images that align with the text description. Paper [20] proposed a generative architecture, OpenFaceGAN that creates facial images from natural language descriptions with improved inference speed, image quality and efficiency in text and image alignment. In another paper [21], a novel network called PixelFace was proposed for TTF synthesis which utilizes a dynamic parameter-generating method to transform text features into embeddings for predicting continuous values of pixels. They validated the experimental results on the MM-CelebA dataset [18]. Nair *et al.* [22], suggested the utilization of diffusion-based models for multi-modal image synthesis that demonstrated impressive results compared to uni-modal network.




There had been research work on TTF synthesis for Bangla language but the images generated by the models were not of high quality, and the FID score was comparatively poor. Previous Bangla TTF works were unable to depict some Bangla facial attributes in synthetic images accurately. Low-quality image generation and limited consistency between Bangla text and generated images in the domain of TTF synthesis has emerged to be the existing gap for Bangla language. Therefore, the purpose of our paper is to elevate the image quality, TTI consistency and betterment of FID score for Bangla TTF synthesis.

III. DATASET

In this paper, we have used the modified version of the CelebA dataset [6] (containing celebrity images and English captions), called CelebA Bangla [7]. The novel algorithm utilized by the CelebA Bangla dataset to generate Bangla textual descriptions of facial images was originally introduced by [7]. The dataset comprises forty facial attributes that have been extracted from Bangla vocabulary that ensures semantic accuracy. Additionally, it contains around 202,599 facial images of size 128×128 of celebrities that correspond to the forty aforementioned facial attributes. The CelebA Bangla dataset¹ is available publicly on Kaggle.

Some of the samples of the CelebA Bangla dataset and text descriptions from CelebA Bangla are shown in Tables I and II.

TABLE I. SAMPLES FROM THE CELEBA BANGLA DATASET

image_name	হালকা দাড়ি (light_beard)	কুচকানো_ফ্রু (arched_eyebrows)	আকর্ষণীয় (attractive)	অল্পবয়স্ক (young)
000012.jpg 	-1	-1	1	1
000013.jpg 	-1	-1	-1	1
202378.jpg 	-1	1	1	1



IV. METHODOLOGY

A. BanglaBERT

BanglaBERT [5] is an ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) transformer language model. Mukh-Oboyob uses BanglaBERT for obtaining accurate text embeddings from Bangla textual descriptions. Of all the available pre-trained Bangla text encoders, BanglaBERT was chosen for this study due to its superior performance on a plethora of NLP tasks. BanglaBERT has its own tokenizer written for the Bangla language, with

¹<https://www.kaggle.com/datasets/rashikrahmanpritom/celeba-bangla-dataset>

TABLE II. SAMPLE TEXT DESCRIPTIONS FROM CELEBA BANGLA DATASET

image_name	text_description
000012.jpg 	ছেলেটির চোখের নিচে কালি ছিল। ছেলেটির কালো চুল ছিল। ছেলেটির সোনালী চুল ছিল। ছেলেটির উঁচু গালের হাড় ছিল। ছেলেটির মুখ কিছুটা খোলা ছিল। ছেলেটির দাড়ি নেই। ছেলেটির চেহারা ডিম্বাকৃতির। ছেলেটির মুখে ছিল হাসি। ছেলেটির সোজা চুল ছিল। (The male has pretty high cheekbones and an oval face. He has black and straight hair. He has bushy eyebrows and a slightly open mouth. The male is smiling, seems young and attractive.)
000013.jpg 	ছেলেটির সোনালী চুল ছিল। ছেলেটির উঁচু গালের হাড় ছিল। ছেলেটির মুখ কিছুটা খোলা ছিল। ছেলেটির দাড়ি নেই। ছেলেটির চেহারা ডিম্বাকৃতির। ছেলেটির মুখে ছিল হাসি। ছেলেটির সোজা চুল ছিল। (The man has high cheekbones and an oval face. His hair is blond and straight. He has a slightly open mouth. He seems young and is smiling.)
202378.jpg 	মেয়েটির ফ্রু কুচকানো ছিল। মেয়েটির বড় ঠোঁট ছিল। মেয়েটির সোনালী চুল ছিল। মেয়েটির মুখে ভারী মেকাপ ছিল। মেয়েটির মুখ কিছুটা খোলা ছিল। মেয়েটির চেহারা ডিম্বাকৃতির। মেয়েটির চোখা নাক ছিল। মেয়েটির চেঁউ খেলানো চুল ছিল। মেয়েটির কানে দুল পরা ছিল। মেয়েটির লিপস্টিক পরা ছিল। মেয়েটির নেকলেস পরা ছিল। (The woman has an oval face. She has brown and wavy hair. She has arched eyebrows, big lips, a slightly open mouth and a pointy nose. She looks young, attractive and has heavy makeup. She is wearing earrings, lipstick and a necklace.)

a rich vocabulary and customized tokenization process. Due to this tokenizer, Bangla text is tokenized properly without loss of valuable information present in subtle parts of the text. As shown in Fig. 1, BanglaBERT turns the input text into tokens $Tok_1, Tok_2, \dots, Tok_N$. These tokens are given to the ELECTRA model, which comprises of two main components: Electra embedding layers and 12 Electra layers. An Electra embedding layer consists of word embedding, position embedding, token type embedding, layer norm and dropout layers. For the purpose of capturing the semantic meaning of the tokens, the Electra layers transform the tokens into continuous vector representations. These representations are fed into 12 Electra layers. Each Electra layer consists of three components: Electra Self Attention, Electra Intermediate, and Electra Output layers. Electra layers help the model to capture contextual information from the input sequence. Electra Self Attention has two constituents: Electra Attention and Electra Self output. For capturing dependencies between tokens, the Electra layer assigns weights by using a Self-Attention mechanism. The Electra Intermediate layer consists of a dense layer and a GELU (Gaussian Error Linear Unit) activation to present contextual information. The Electra Output layer assists the model in grasping thematic insight efficiently. Finally, the output of the 12 Electra layers is a text embedding of dimensions $[num_prompts \times max_length \times embedding_dim]$. Here, $num_prompts$ is the number of prompts/input Bangla textual descriptions given to the text encoder. max_length is the max-

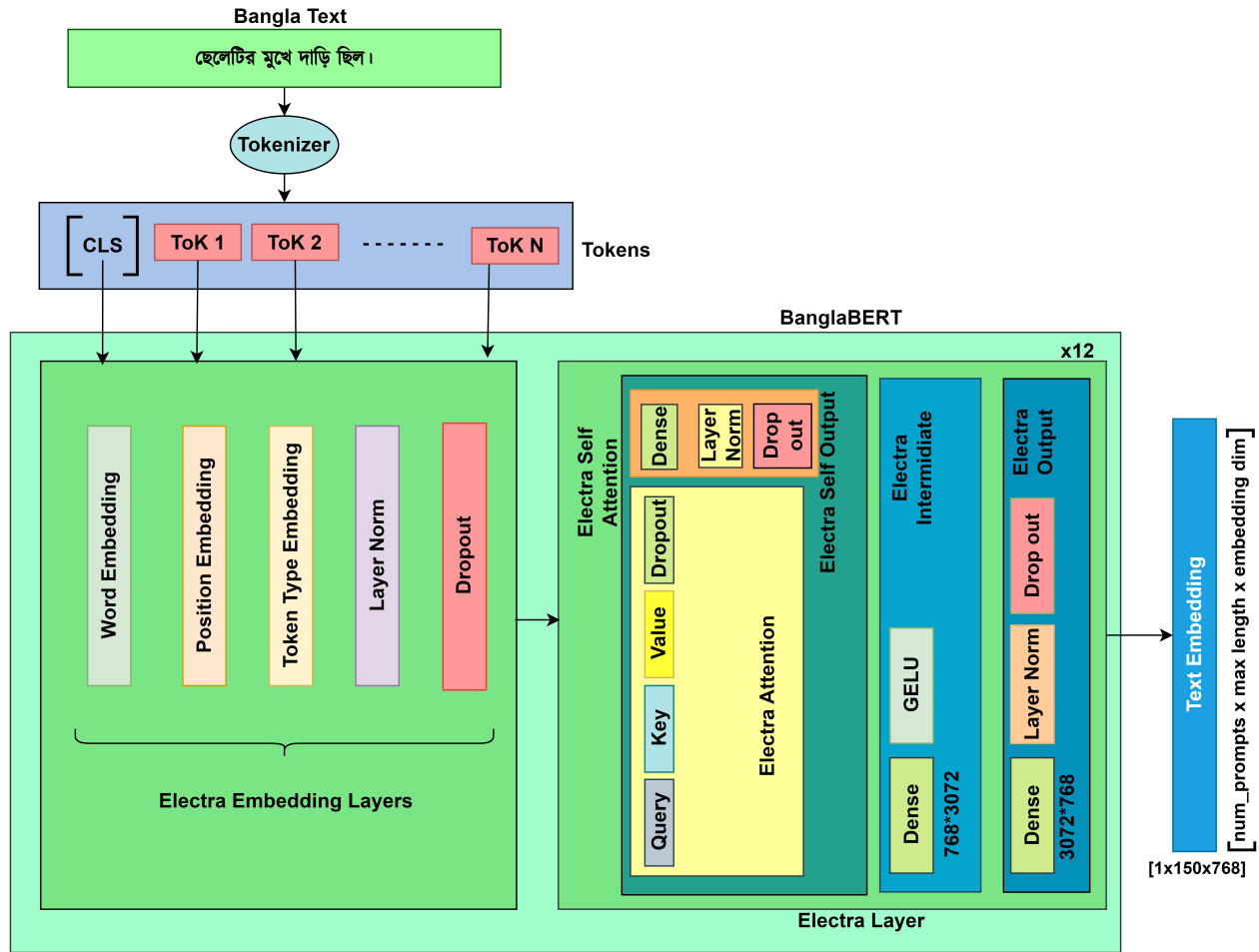


Fig. 1. Internal layered architecture of BanglaBERT.

imum number of tokens allowed to be given by the tokenizer. $embedding_dim$ represents the embedding dimension of the embedding layers.

B. Stable Diffusion

Stable Diffusion [3] is based on the latent diffusion model that produces synthetic images from text input. Stable diffusion mainly comprises a Variational Autoencoder(VAE), some Schedulers, a Text Encoder, a U-Net Model, and Classifier Free Guidance. The training and image generation process is explained briefly below.

Training: The training process of Stable Diffusion is outlined in Fig. 2. During the training of Stable diffusion, an input facial image is passed through the VAE encoder to obtain a latent vector. As shown in Eq. 1, the latent vector is scaled by a scaling factor defined in the configuration of the VAE. Scaling the latent vector allows Mukh-Oboyob to control the amount of randomness of the probability distribution of synthetic facial images.

$$latent_vector_{scaled} = scaling_factor \times vae_encoder(image_face) \quad (1)$$

It is shown in Eq. 2 that the scaled latent vector, a random noise vector, and timestep are passed to a noise scheduler for timesteps $t = 1 \dots T$. This is the Forward Diffusion process. In this process, the noise scheduler gradually adds noise to the latent image, thus obtaining a noisy latent vector.

$$latent_vector_{noisy} = noise_scheduler(latent_vector_{scaled}, noise_{random}, timesteps) \quad (2)$$

A Bangla textual description is passed through BanglaBERT to obtain a text embedding. The Text embedding, Random noise, and noisy latent vector are fed into the U-Net for the purpose of predicting a noise vector; as demonstrated in Eq. 3. The U-Net utilizes its contracting path and expansive path to better predict a noise vector close to the random noise previously used in the forward diffusion process.

$$Noise_{predicted} = U-Net(Text_Embedding, timesteps, latent_vector_{noisy}) \quad (3)$$

The predicted noise and random noise are compared using Mean Squared Error Loss as defined in Eq. 4. Here, N signifies

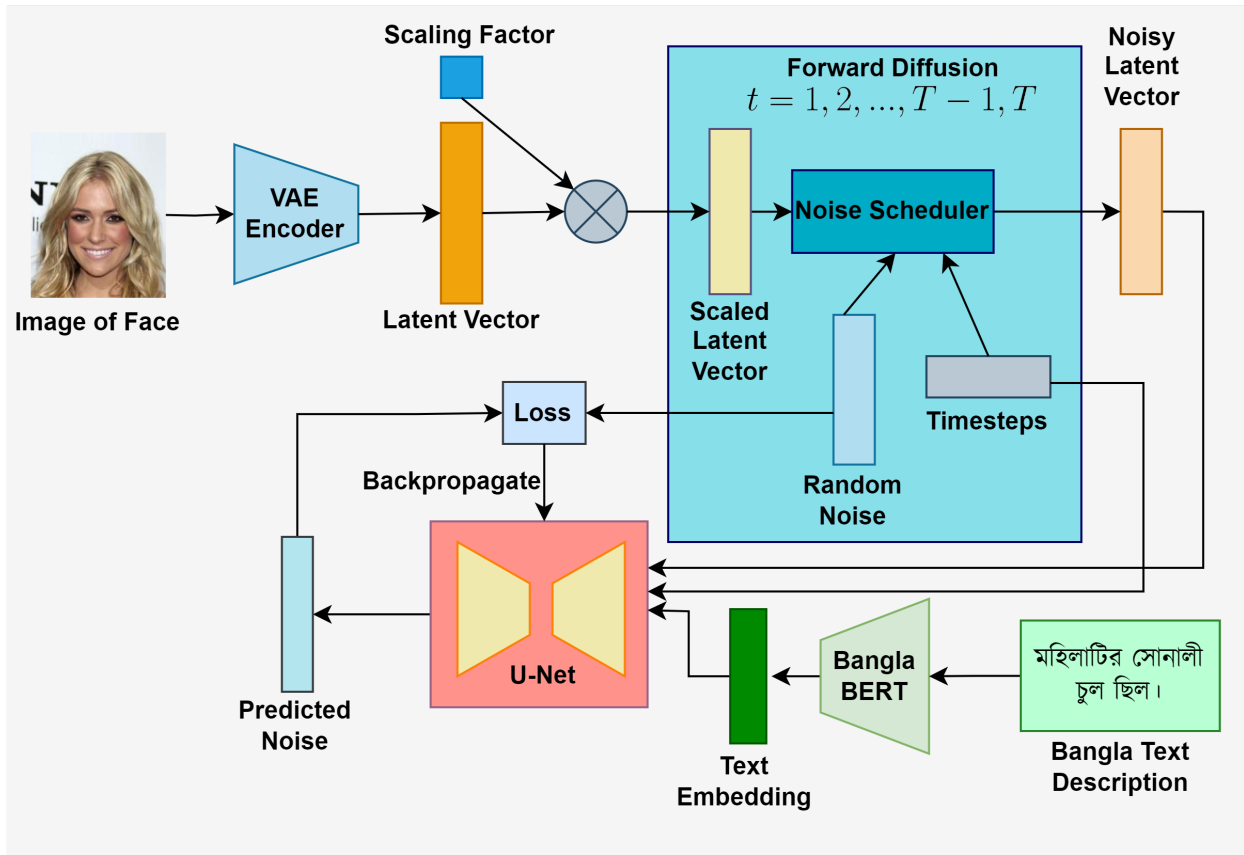


Fig. 2. Training procedure of the stable diffusion model used by Mukh-Oboyob.

the number of rows in the noise vectors. This Loss is back-propagated through the U-Net to help the U-Net predict better noise vectors in future iterations.

$$Loss_{MSE} = \frac{1}{N} \sum_{i=1}^N (Noise_{random} - Noise_{predicted})^2 \quad (4)$$

Image Generation: The Image Generation Phase as depicted in Fig. 3 generates synthetic images from Bangla Text. During the Image Generation or Prompting phase, a Bangla Text prompt is sent to BanglaBERT to obtain a prompt embedding. A random latent vector is scaled by following Eq. 5. Here σ is used to control how much noise is added to the latent text representation.

$$latent_model_input = \frac{random_latent}{\sqrt{\sigma^2 + 1}} \quad (5)$$

The latent model input, prompt embedding and timestep are given to the U-Net for predicting Noise in Eq. 6. This Noise vector is the U-Net's attempt to produce a latent representation of the text; which can later be decoded into an image.

$$Noise_{predicted} = U-Net(latent_model_input, timestep, Text_Embedding) \quad (6)$$

However, This predicted noise is not satisfactory at timestep $t = T - 1$. Therefore, as shown in Eq. 7, the predicted noise and timesteps are iteratively passed on to the scheduler which produces another latent vector for timesteps $t = T-1, T-2, T-3, \dots, 3, 2, 1$. This is the Reverse Diffusion process.

$$latent_vector = Scheduler(Noise_{predicted}, timesteps) \quad (7)$$

Finally, in Eq. 8, the latent vector achieved at timestep $t = 1$ is scaled by a scaling factor defined in the variational autoencoder's configuration. This scaling is done to ensure that the latent vector is normalized and has values in a specific range, thus helping to improve consistency across a multitude of samples.

$$latent_vector_{scaled} = \frac{latent_vector}{scaling_factor} \quad (8)$$

The scaled latent vector is now passed to the Decoder of the Variational Autoencoder used in Mukh-Oboyob. As depicted in Eq. 9, The Decoder produces an image of a face in accordance to the textual prompt given to the Text Encoder earlier.

$$image_{face} = Decoder_{VAE}(latent_vector_{scaled}) \quad (9)$$

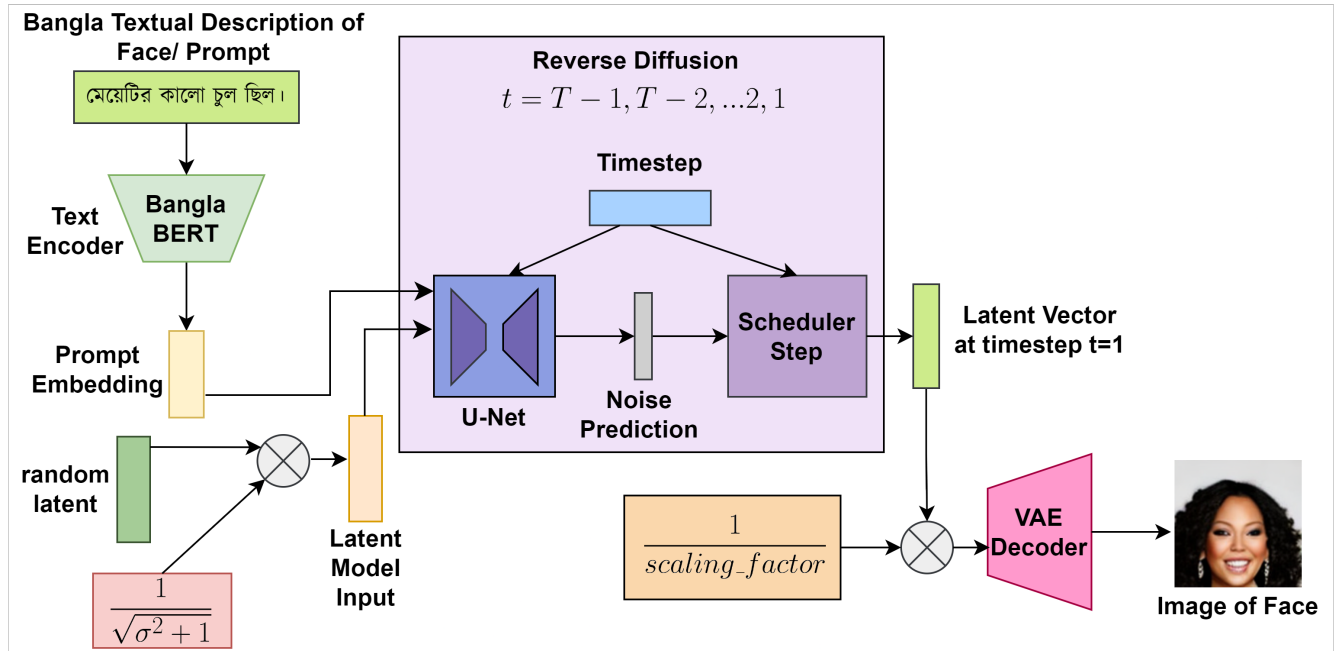


Fig. 3. Image generation procedure used in stable diffusion of Mukh-Oboyob.

C. LoRA

Fine-tuning the entire Stable Diffusion model can be a hardware and time-consuming task; often unfeasible in limited hardware and electric power support. Therefore Mukh-Oboyob uses LoRA (Low-Rank Adaptation) [23] to fine-tune the cross-attention layers in the U-Net model part of Stable diffusion. Let the weight matrix of a cross-attention layer be C_0 . LoRA will selectively update C_0 by using the low-rank decomposition in Eq. 10. During the training or fine-tuning process, the C_0 matrix is not updated in the backward pass. Only T_A and T_B are updated while training.

$$C_0 + \Delta C = C_0 + T_B T_A \quad (10)$$

By significantly reducing the number of trainable parameters in the process outlined above, LoRA reduces training time and VRAM consumption drastically; without causing noticeable degradation in synthetic image quality.

V. RESULT ANALYSIS

In this section, a comprehensive discussion of the experimental details during training and validation of the proposed model is provided.

A. Experimental Setup

Stable Diffusion v1-4² was fine-tuned using LoRA on a single RTX 3060 GPU for developing the proposed system, Mukh-Oboyob. Stable Diffusion uses a CLIP Text Encoder [24], which only works for English text inputs. For Mukh-Oboyob, BanglaBERT's text encoder and tokenizer was used. Input image resolution was changed to 128×128 to make

it compatible with the CelebA Bangla dataset. Furthermore, CLIP tokenizer has a maximum sequence length of 77, whereas the textual descriptions of CelebA Bangla produce upto 150 tokens when tokenized with the BanglaBERT tokenizer. When tokenized with $max_length = 77$, BanglaBERT's tokenizer discards significant parts of the Bangla text. Therefore, for compatibility issues, the maximum sequence length was set to 150 in the proposed system, Mukh-Oboyob. The batch size was set to 16. an initial learning rate of 10^{-4} was used. The constant scheduler was chosen as the learning rate scheduler. Number of warmup steps was set to 0 for the learning rate scheduler. The hyperparameters for the Adam optimizer used are: $\beta_1 = 0.9, \beta_2 = 0.9, weight_decay = 10^{-2}, \epsilon = 10^{-8}$. The dimension of the LoRA update matrices was set to 4 for training the proposed method, Mukh-Oboyob. A Variational Autoencoder³ trained with Exponential Moving Average weights was used during prompting Mukh-Oboyob.

B. Qualitative Analysis

As shown in Fig. 4, Mukh-Oboyob produces images with far better quality and diversity compared to previous GAN methods. The synthetic images produced by Mukh-Oboyob are more semantically aligned with the Bangla textual descriptions of faces. Almost all facial attributes written in the input textual descriptions are accurately depicted in the corresponding images produced by Mukh-Oboyob. This shows that BanglaBERT successfully provided meaningful text embeddings which were properly comprehended by Stable Diffusion.

The effect of the hyperparameter called number of inference steps was explored in Fig. 5. With only 1 inference step, a noisy image is produced. While inference steps increase, inference time and image quality also increase. Regardless of

²<https://huggingface.co/CompVis/stable-diffusion-v1-4>

³<https://huggingface.co/stabilityai/sd-vae-ft-ema>

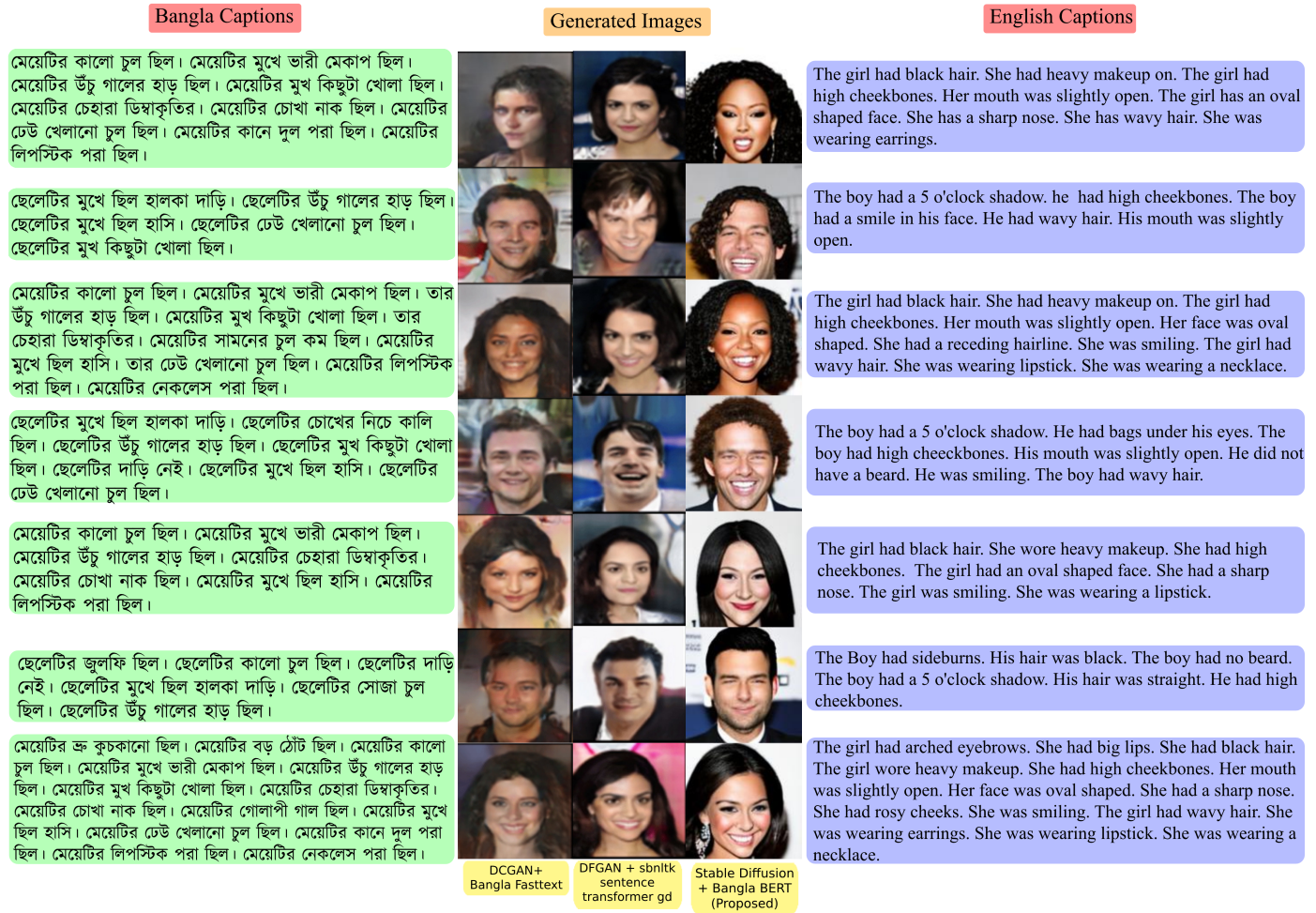


Fig. 4. Comparison of generated images between previous GAN methods and Mukh-Oboyob for different bangla captions.

the number of inference steps, VRAM consumption stays the same.

C. Quantitative Analysis

Mukh-Oboyob achieves state-of-the-art quantitative results on the performance metrics outlined in [7], overcoming previous GAN-based results, as numerically proven in Table III. All performance metrics were computed on 10,000 synthetic images, as practiced in various studies. FID is a widely used performance metric for evaluating the quality and diversity of generated images. Mukh-Oboyob achieves a better FID score of 34.6828 compared to the other models by a large margin. Although Inception Score(IS) is a metric used for assessing the quality and diversity of generated images, it is criticized in existing literature for having sensitivity to dataset bias, lack of semantic coherence, and limited applicability to different domains. Mukh-Oboyob achieves a competitive Inception Score of 11.3721, as shown in Table III. Learned Perceptual Image Patch Similarity(LPIPS) is an excellent domain agnostic performance metric for image synthesis which correlates very well with human perception. The proposed model, Mukh-Oboyob also achieves a much better LPIPS score compared to DCGAN and DFGAN. Face Semantic Similarity (FSS)

and Face Semantic Distance(FSD) are used for comparing the similarity and dissimilarity of generated and real faces. Although FSS and FSD are relevant to the TTF domain, they are not widely recognized or established metrics. Nevertheless, Mukh-Oboyob achieves a competitive FSS and FSD score as shown in Table III.

TABLE III. COMPARISON OF PERFORMANCE METRICS BETWEEN MUKH-OBOYOB AND PREVIOUS METHODS

Model	FID ↓	IS ↑	LPIPS ↓	FSD ↓	FSS ↑
Bangla fasttext + DCGAN [7]	126.71	12.3607	21.8291	20.2385	0.3427
sbnlk sentence transformer gd + DFGAN [7]	155.1593	4.78246	3.2216	20.3697	0.4203
Mukh-Oboyob (BanglaBERT + Stable Diffusion)	34.6828	11.3721	0.4541	24.8942	0.0528

Fig. 6 depicts the decreasing MSE loss at each epoch during the training of the proposed method, Mukh-Oboyob. At each epoch, there were 12630 updates; so each epoch was very time-consuming.






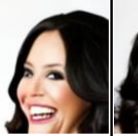

Text Description/Prompt	inference steps = 1	inference steps = 5	inference steps = 10	inference steps = 15	inference steps = 20	inference steps = 25	inference steps = 30
মেয়েটির কালো চুল ছিল। মেয়েটির মুখে ভারী মেকাপ ছিল। মেয়েটির উঁচু পালের হাড় ছিল। মেয়েটির মুখ কিছুটা খোলা ছিল। মেয়েটির চোখা নাক ছিল। মেয়েটির মুখে ছিল হাসি। মেয়েটির চেউ খেলানো চুল ছিল। মেয়েটির লিপস্টিক পরা ছিল। (The girl had black hair. She was wearing heavy makeup. The girl had high cheekbones. Her face was slightly open. The girl had a sharp nose. She was smiling. She had wavy hair. The girl was wearing lipstick.)							
Inference Time	366 ms	449 ms	757 ms	1027 ms	1324 ms	1601 ms	1880 ms
VRAM Consumption	5.09 GB						

Fig. 5. Effect of inference steps on the quality of generated images.



Fig. 6. MSE loss at different epochs of training the proposed Mukh-Oboyob model.

VI. DISCUSSION

Developing a system that performs well and combines state-of-the-art models on limited hardware to a new domain requires a significant amount of background knowledge and experience. The evaluation of generative models is prone to subjectivity and lack of a clear ground truth [2]. Despite these adversities, Mukh-Oboyob achieves stellar performance and establishes a new state of the art in Bangla TTF Synthesis. The most subtle bangla facial attributes are learned surprisingly by the proposed model, Mukh-Oboyob.

VII. LIMITATIONS

Even though Mukh-Oboyob achieves never-before-seen results on Bangla TTF synthesis, it is a bit behind compared to English TTF models which have achieved single-digit FID scores. The relatively less advanced performance of Mukh-Oboyob can be attributed to the lack of a pre-trained model for Bangla that adequately captures the sophisticated details of the Bangla language, as BERT or GPT captures for English. Another issue faced by Mukh-Oboyob is that some of the generated facial images contain dark or blurry eyes. Even after using a pre-trained VAE aimed at solving this issue, a few

images are still synthesized with dark eyes. This is an open research problem.

VIII. CONCLUSION

This paper proposes a novel system, Mukh-Oboyob for producing images of faces from Bangla Textual input. Mukh-Oboyob uses BanglaBERT as a Text Encoder and Stable Diffusion for image generation. The proposed Mukh-Oboyob model was trained and evaluated on the CelebA Bangla dataset. Mukh-Oboyob achieves a state-of-the-art FID score of 34.6828 and an LPIPS score of 0.4541. A limitation of this work is that the performance of Mukh-Oboyob is relatively lower compared to state-of-the-art English TTF models. Another limitation of Mukh-Oboyob work is that this work suffers from lack of established performance metrics for evaluation of the facial features of synthetic facial images. An interesting avenue of future work can be generating more diverse and realistic facial images from captions of other languages(Arabic, Hindi, Spanish, etc.).

ACKNOWLEDGMENT

We are grateful to the Institute of Energy, Environment, Research, and Development (IEERD, UAP) and the University of Asia Pacific for their financial support. We extend our sincerest gratitude to Md Shopon for his inspiration and insights. We thank A. Faiyaz for technical assistance.

REFERENCES

- [1] N. G. Nair, W. G. C. Bandara, and V. M. Patel, "Unite and conquer: Plug & play multi-modal synthesis using diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6070–6079.
- [2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.
- [3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.
- [4] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, and B. Guo, "Vector quantized diffusion model for text-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 696–10 706.

- [5] A. Bhattacharjee, T. Hasan, W. Ahmad, K. S. Mubasshir, M. S. Islam, A. Iqbal, M. S. Rahman, and R. Shahriyar, "BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla," in *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 1318–1327. [Online]. Available: <https://aclanthology.org/2022.findings-naacl.98>
- [6] Z. Liu, P. Luo, X. Wang, and X. Tang, "Large-scale celebfaces attributes (celeba) dataset," *Retrieved August*, vol. 15, no. 2018, p. 11, 2018.
- [7] N. M. K. Arnob, N. N. Rahman, S. Mahmud, M. N. Uddin, R. Rahman, and A. K. Saha, "Facial image generation from bangla textual description using dcgan and bangla fasttext," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, 2023.
- [8] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *International conference on machine learning*. PMLR, 2016, pp. 1060–1069.
- [9] M. A. H. Palash, M. A. Al Nasim, A. Dhali, and F. Afrin, "Fine-grained image generation from bangla text description using attentional generative adversarial network," in *2021 IEEE International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things (RAAICON)*. IEEE, 2021, pp. 79–84.
- [10] S. Naveen, M. S. R. Kiran, M. Indupriya, T. Manikanta, and P. Sudeep, "Transformer models for enhancing atngan based text to image generation," *Image and Vision Computing*, vol. 115, p. 104284, 2021.
- [11] H. Zhang, J. Y. Koh, J. Baldrige, H. Lee, and Y. Yang, "Cross-modal contrastive learning for text-to-image generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 833–842.
- [12] M. Tao, H. Tang, F. Wu, X.-Y. Jing, B.-K. Bao, and C. Xu, "Dfgan: A simple and effective baseline for text-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 515–16 525.
- [13] M. Siddharth and R. Aarthi, "Text to image gans with roberta and fine-grained attention networks," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 12, 2021.
- [14] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.
- [15] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 479–36 494, 2022.
- [16] K. Deorukhkar, K. Kadamala, and E. Menezes, "Fgtd: Face generation from textual description," in *Inventive Communication and Computational Technologies: Proceedings of IICCT 2021*. Springer, 2022, pp. 547–562.
- [17] J. Sun, Q. Deng, Q. Li, M. Sun, M. Ren, and Z. Sun, "Anyface: Free-style text-to-face synthesis and manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 687–18 696.
- [18] W. Xia, Y. Yang, J.-H. Xue, and B. Wu, "Tedigan: Text-guided diverse face image generation and manipulation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2256–2265.
- [19] D. Ayanthi and S. Munasinghe, "Text-to-face generation with stylegan2," *arXiv preprint arXiv:2205.12512*, 2022.
- [20] J. Peng, H. Pan, Y. Zhou, J. He, X. Sun, Y. Wang, Y. Wu, and R. Ji, "Towards open-ended text-to-face generation, combination and manipulation," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 5045–5054.
- [21] J. Peng, X. Du, Y. Zhou, J. He, Y. Shen, X. Sun, and R. Ji, "Learning dynamic prior knowledge for text-to-face pixel synthesis," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 5132–5141.
- [22] N. G. Nair, W. G. C. Bandara, and V. M. Patel, "Unite and conquer: Plug & play multi-modal synthesis using diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6070–6079.
- [23] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022. [Online]. Available: <https://openreview.net/forum?id=nZeVKeeFYf9>
- [24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.