FaceSynth: text-to-face generation using CLIP and its variants with generative adversarial networks

Priyadharsini Ravisankar[1], Shruthi Dhanvanth[2], Vaishnave Jenane Padmanabhan[2]
[1]Department of Artificial Intelligence and Data Science, Rajalakshmi Engineering College, Chennai, India
[2]Department of Computer Science, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India

Article Info

Article history:

Received Jul 30, 2024
Revised Jun 27, 2025
Accepted Jul 13, 2025

Keywords:

CLIP
Generative adversarial network
MetaCLIP
Multilingual-CLIP
StyleGAN
Text-to-face generation
Text-to-image generation

ABSTRACT

In recent years, there have been massive developments in the field of generative AI, especially in generative adversarial networks (GANs). GANs generate original images that haven't been seen during training and have had several advancements like StyleGAN, StyleGAN2, and StyleGAN2-adaptive discriminator augmentation (ADA). Contrastive language-image pre-training (CLIP), by OpenAI, is a visual linguistic model that has been trained to associate texts with images. Recently, new CLIP variants were developed, such as metadata-curated language-image pre-training (MetaCLIP), released by Facebook and trained on a larger dataset, and Multilinigual-CLIP, which adapts CLIP to multiple languages. We compare CLIP and its variants in text-to-face synthesis with a custom StyleGAN2-ADA model and a pre-trained StyleGAN2 model. Our training-free algorithm starts with an initial image latent code that is iteratively manipulated to match a given text description. It achieves this by minimizing the distance between the text and image embedding in the multi-modal embedding space of the CLIP models. An examination of CLIP and its variants showed that MetaCLIP outperformed its competitors in LPIPS similarity and closeness of the synthesized image to the actual prompt. CLIP produced the most realistic images with the best FID score and multilingual-CLIP presented a choice of input text language and generated decent images.

Corresponding Author:

Priyadharsini Ravisankar
Department of Artificial Intelligence and Data Science, Rajalakshmi Engineering
Chennai, India
Email: priyadharsini.r@rajalakshmi.edu.in

1. INTRODUCTION

Artists draw pictures from their imagination and are proficient at depicting various entities like birds, animals, scenery, and human faces. To give their drawings a lifelike appearance, they incorporate color, texture, composition, and expressions. Furthermore, when an artist is provided a text description, they can draw a sketch that captures all of the features specified while simultaneously maintaining the realistic aspects. Using machine learning models, we can mimic this unique ability of artists and automate the creation of an image from text, thus dramatically reducing manual labor. Generative artificial intelligence has evolved to be one of the most noteworthy advances in recent years in computer vision. To create new data, generative AI learns patterns and sequences found in dataset samples. The primary goal of text-to-image synthesis, a branch of Gen-AI, is to create an image from an input caption. The attributes specified in the text guide image generation, and this process has several uses in art, storytelling, education, and more. Text-to-face synthesis, a subfield in text-to-image synthesis, produces a facial image from a description and requires greater attention to detail. Human faces contain many subtleties and mistakes in generated faces are easily detectable. There have been many advancements in the field of text-to-face synthesis, through generative adversarial networks (GANs) [1] and diffusion models [2].

GANs [1] are machine learning models that create original data that consist of two sub-networks, a generator and a discriminator. These networks compete adversarially throughout the training period. Here, the generator produces artificial images and the discriminator categorizes them as real or fake. The generator takes feedback from the discriminator and makes iterative improvements, in which manner the two sub-networks challenge each other to create unique data. There have been several refinements to the original GAN to improve the quality and realism of synthesized images, like conditional generative adversarial network (CGAN) [3] and deep convolutional generative adversarial network (DCGAN) [4]. Some models like attentional generative adversarial networks (AttnGAN) [5] and stacked generative adversarial networks (StackGAN) [6] are popular in text-to-image generation for birds and other objects but do not generalize well and have the nuance needed for face synthesis. Style-based generative adversarial networks (StyleGAN) [7], based on progressive growing of generative adversarial network (ProGAN) [8], is a state-of-the-art generative model that creates high-quality, realistic images. In StyleGAN, the network contains many layers, with the initial ones producing a lower-dimensional image that concentrates on the basic features and the other layers focused on adding more complex details to the image. Advancements in StyleGAN have seen the development of StyleGAN2 [9] and StyleGAN3 [10]. All these StyleGAN models require an enormous dataset for training, which is extremely expensive and computationally intensive. StyleGAN2-adaptive discriminator augmentation (ADA) [11] was created to fix this problem and can be trained on a smaller, limited dataset. It builds on the original StyleGAN architecture by adding data augmentation techniques. StyleGAN2 was used with text encoders like BERT in ‘text to face generation with StyleGAN2’ [12], and DistilBert in ‘StyleT2F’ [13] to generate faces from text descriptions.

A visual-linguistic model called contrastive language-image pre-training (CLIP) [14] was created by OpenAI to link texts with images. It is a deep learning model trained on 400 million image-text pairs obtained from ImageNet and connects them by encoding both into a joint embedding space. It is notable for its success in zero-shot learning, which involves classifying images with labels not encountered in training. CLIP contains an image and text encoder and was used with StyleGAN in several text-to-face generation and manipulation models like StyleCLIP [15] and TediGAN [16]. Previous works have explored text-to-face generation and manipulation using a pre-trained StyleGAN [7] and CLIP model [14]. CLIPDraw [17] is a text-to-drawing algorithm that synthesizes drawings by maximizing the cosine similarity between a generated sketch and an input description. This method is biased towards drawings rather than realistic images. Reddy *et al.* [18] proposed a model that generated a random image and optimized its latent code with CLIP’s loss function.

Recently, Facebook Research released MetaCLIP [19], based on CLIP, trained on a massive dataset of 1 billion images fetched from CommonCrawl. CLIP’s accomplishments were said to lie in the quality of data it was trained on and not in its architecture. Since there is inadequate information on how CLIP obtained its training data, MetaCLIP intended to unveil and refine CLIP’s method of acquiring data. MetaCLIP is relatively new and has not had any applications in text-to-image synthesis. Vision-language models and CLIP, in particular, are crucial models in artificial intelligence that have made a considerable impact in the field. However, most focus only on English texts, which is a consequence of the scarce number of image-text datasets in other languages. Multilingual-CLIP [20] was introduced to address this limitation and leveraged the strength of CLIP’s pre-trained text encoder to train a student model to process multiple languages. It has several pre-trained models designed for diverse languages and can be used in multilingual text-to-face generation.

Each of these models has its advantages and disadvantages. Our main contribution involves experimenting, analyzing, and comparing CLIP, MetaCLIP, and multilingual-CLIP in English and Tamil in text-to-face generation. Multilingual-CLIP can be used to enable high-quality text-to-face synthesis in several other languages to make it accessible worldwide. We also assess the performance of a custom StyleGAN2-ADA model [11], trained with a subset of images from the FairFace dataset [21], and a pre-trained StyleGAN2 model trained with the FFHQ dataset. Our goal is to synthesize realistic images that contain the fine-grained attributes mentioned in the description. Our proposed system integrates CLIP and its variants with StyleGAN2 or StyleGAN2-ADA. CLIP [14], along with its variants, MetaCLIP [19] and multilingual-CLIP [20], connect text descriptions with their visual representations, while StyleGAN2 or StyleGAN2-ADA [11] generates the high-quality images. StyleGAN2-ADA with the FairFace dataset is used to ensure diversity and fairness in the synthesized images and can be contrasted with the pre-trained StyleGAN2 model. The overall algorithm operates in two phases. In the first phase, a given text description is encoded using CLIP or its variants text encoder [8], [14], [19]. A starting latent vector, which is a numerical representation of the features of a text or image in a higher dimensional latent space, is generated. This is done by randomly synthesizing a few images, storing the loss between the encoded generated image and text, selecting the image with the lowest loss, and obtaining its latent vector as the starting latent code. In the second phase, this starting latent code is optimized with a loss repeatedly. The loss is found in every iteration after synthesizing and encoding the image, following which it uses an arcsin function to find the distance between the embeddings of the image and text, which corresponds to how similar the input text is to the generated image. The calculated loss is passed backward to update the latent code. This procedure is repeated 100 times to generate the final image. The results obtained from our work highlight the trade-offs in using a custom-trained StyleGAN2-ADA model versus a pre-trained StyleGAN model, as well as the trade-offs between CLIP and its variants. The conclusion is that while images synthesized with MetaCLIP match the text description better, CLIP produces more realistic-looking images. MultilingualCLIP makes it possible for text-to-face synthesis to function in multiple languages. However, its performance is subpar relative to the other two models in realism and closeness to the text. Finally, while the pre-trained StyleGAN2 model performed better, the custom-trained StyleGAN2-ADA model ensured more diversity and fairness in its synthesized images. This research paper intends to provide useful insights that help in making informed decisions and choosing models that best-fit requirements.

In section 2, a literature survey is conducted that goes over related work, reviewing various text-to-image and text-to-face generation models. Section 3 discusses the modules used in our algorithm, StyleGAN [7], [9]–[11], CLIP, MetaCLIP, and multilingual-CLIP, our proposed architecture, which consists of the dataset description, overall architecture, algorithm, explanation of our loss function and the evaluation metrics, and finally the experiments we conducted. Section 4 contains the results and a comprehensive discussion section where the results are presented and studied. Finally, we present our conclusions in section 5 and discuss future work.

2. LITERATURE SURVEY

Text-to-face and text-to-image synthesis using GANs are well-researched topics of generative AI that have large-scale applications. Some of the most significant papers are outlined along with their contributions. Zhang *et al.* [6] introduced one of the first text-to-image models called StackGAN, trained on bird images from the CUB and MS COCO datasets. The model had two GANs stacked on top of each other, where the first GAN added primary characteristic features like shape and color, and the second one added more high-level features. This model utilized the original GAN, which led to issues like mode collapse and training instability. Xu *et al.* [5] introduced Attn-GAN, a newer architecture within the GAN framework that used the same dataset as StackGAN and followed a multiple-stage approach. In every stage, a few significant keywords from the input prompt were extracted and used to synthesize an image of low resolution, which was finally combined using the word contexts developed in the previous stages. The performance of this model deteriorated as the description got longer because the attention map became more complicated to train. Nasir *et al.* [22] proposed Text2FaceGan, where the main contribution lay in creating an algorithm to add captions to the images of the CelebA dataset, describing the attributes that were present in them. This was achieved using a skip-thought encoder and was a significant contribution to this field, as earlier facial image-text datasets were scarce.

Sabae *et al.* [13] introduced StyleT2F using DistilBert to extract facial features from a text description, which were transformed into a final latent vector passed to the StyleGAN2 generator. Feature directions were used to navigate the StyleGAN2 latent space and reach the required latent vector. This model had multiple problems due to entangled feature directions. Following a similar approach, Ayanthi and Munasinghe [12] proposed a framework that exhibited a similarity of over 50% to the original images. The model used BERT to extract the text encodings, which was given to a pre-trained GAN such as StyleGAN2. It was trained with the perceptual loss function and performed better than older GANs like AttnGAN [5], and StackGAN. However, the dataset used in this paper consisted of only 5685 image-text pairs and hence did not generalize well, leading to overfitting. Todmal *et al.* [23] proposed two methods that used CLIP, StyleGAN, and the pixel2style2pixel image encoder [24], which projects images into the extended latent space of StyleGAN. The first method mapped a prompt to the extended latent space of StyleGAN, while the second mapped it to the initial latent space of StyleGAN. The first method resulted in facial images that were true to the description but less lifelike, while the second had more realistic images but less control over the attributes. Reddy *et al.* [18] proposed an algorithm that trained a StyleGAN inverter to encode a given image and obtain its intermediate latent code and utilized a looped network for text-to-image generation. Initially, a random latent code is created and passed to the StyleGAN generator to synthesize a random image. The synthesized image and input caption are compared with CLIP Loss. The trained StyleGAN inverter finds the latent code of the synthesized image, updates it with the loss, and regenerates an image. This process occurs for a fixed number of steps as the loss value decreases. In the branch of text-to-face generation in multiple languages, Li *et al.* [25] proposed a model that used transfer learning and made use of neural machine translation. It had two approaches and tested the results on the CUB and COCO-CN datasets with StyleGAN2. It analyzed and evaluated how their cross-lingual transfer methods compare to other transfer methods.

3. METHODS AND COMPONENTS

3.1. StyleGAN

StyleGAN, built on top of ProGAN [8], is the state-of-the-art model in the field of generativeAI using GANs, that is capable of producing high-quality, realistic images. Unlike its predecessors, which used only one latent space called the Z space to sample attributes from, StyleGAN proposed to use an intermediate latent space called the W space. The Z space consists of random vectors (noise vectors) that control image generation. In contrast, the W space, which we get by passing the input through an 8-layer MLP mapping network, enables smoother interpolation between latent vectors. It ensured more disentanglement between different features in the W space, which provides more control over the individual attributes. StyleGAN was succeeded by StyleGAN2 and StyleGAN3 [10], which had architectural changes to combat the limitations posed by the original StyleGAN. These modifications fixed earlier issues like phase artifacts, the water-droplet effect, and texture sticking, which generated even more realistic images. Another major development was the StyleGAN2-ADA model, which handled the problem of limited datasets. Most GAN-based models require 50,000 - 100,000 images for training and use augmentation to work with small datasets, which results in overfitting. This challenge, addressed by the StyleGAN2-ADA model, produced outputs similar to StyleGAN2 while using a significantly smaller dataset, which assisted in overcoming data scarcity challenges.

3.2. Contrastive language-image pre-training

CLIP is a model, designed to link a set of pictures to text descriptions, trained on 400 million pairs of image-text pairs on ImageNet. The CLIP framework consists of a text encoder that utilizes transformers and an image encoder that utilizes ResNet [26] or image transformers [27]. It employs a metric to measure the likeness of a given caption to a picture called cosine similarity. Both the encoders yield uniform-sized embeddings positioned in a joint embedding space. It is possible to find how far apart the encoded caption and picture are in this space. CLIP is famous for categorizing an image with labels not encountered during training, called zero-shot learning.

3.3. Metadata-curated language-image pre-training

Facebook Research developed MetaCLIP, which highlights the data collection method of CLIP since it suggests that insight into this process can reveal the factors that made it so successful. Since CLIP does not disclose how it collects data, the paper observes what quantifies good quality data and discusses a technique to reveal CLIP's selection process. MetaCLIP has been trained with 1 billion images present in CommonCrawl and beats CLIP's performance in image classification with unknown labels, called zero-shot learning.

3.4. Multilingual-curated language-image pre-training

Most vision language networks are centered on English because of the sparse number of image-text datasets available in other languages. Multilingual-CLIP bridges this gap by utilizing transfer learning to train a novice network to produce a matching embedding to the pre-trained CLIP teacher model. Multilingual-CLIP focuses on using a student text encoder but with the same vision encoder. The trainee encoder utilizes a BERT transformer model, pre-trained in a non-English language, with mean squared error (MSE) loss to align its embeddings with the trainer network. Multilingual-CLIP is very useful in multilingual vision-language tasks since it was trained in numerous languages.

3.5. System architecture

3.5.1. Dataset description

This research utilized two datasets, the FairFace dataset and the multi-modal-CelebA-HQ dataset in which the first was used to train the StyleGAN2-ADA model and the second to test the proposed algorithm. The FairFace dataset has 108,501 facial images with a balanced representation across seven ethnic groups. This dataset was cleaned and preprocessed by ignoring the low-quality data and resizing the remaining to 256×256 pixels, after which a smaller subset of 10,500 images was created by selecting an equal proportion of pictures from each of the seven ethnicities. Since StyleGAN2-ADA is known for working with smaller datasets, 10,500 images were sufficient for training.

The multi-modal-CelebA-HQ dataset [16] contains 30,000 images, which are divided into 24,000 training and 6,000 testing data, where each image has a caption formed from a set of 40 attributes, such as "wavy hair", "mustache", "wearing earrings", and more. Since the proposed model does not have a training phase, both the training and testing image-caption pairs were used for evaluation. The captions in the multi-modal-CelebA-HQ dataset were translated to Tamil using the googletrans-py package to evaluate the multilingual-CLIP model.

3.5.2. System architecture

The overall architecture in Figure 1 contains these key components: StyleGAN, a CLIP variant, the loss function, and two loops where the first loop generates the starting image latent code, and the second uses this latent to produce the final image. The first loop begins by taking a random tensor in the Z space and passing it to the StyleGAN generator to synthesize an image. The generated image and the input text are then encoded using the respective image and text encoders of CLIP or its variants. A loss is calculated between this encoded image and text and recorded along with the intermediate W latent. This cycle occurs 10 times, after which the image with the lowest loss's W latent code is selected as the initial image latent in the next loop.

The following loop takes the selected W latent code, initializes it with an optimizer, and then generates an image by passing it through StyleGAN's synthesis network [9], [11]. The encoded text and image are utilized to calculate a loss, which is backpropagated to update the W latent code in each iteration. The final image is outputted after 100 iterations and has the highest accuracy to the given input prompt. This training-free method effectively produces an image that closely matches the input caption prompt using two loops.

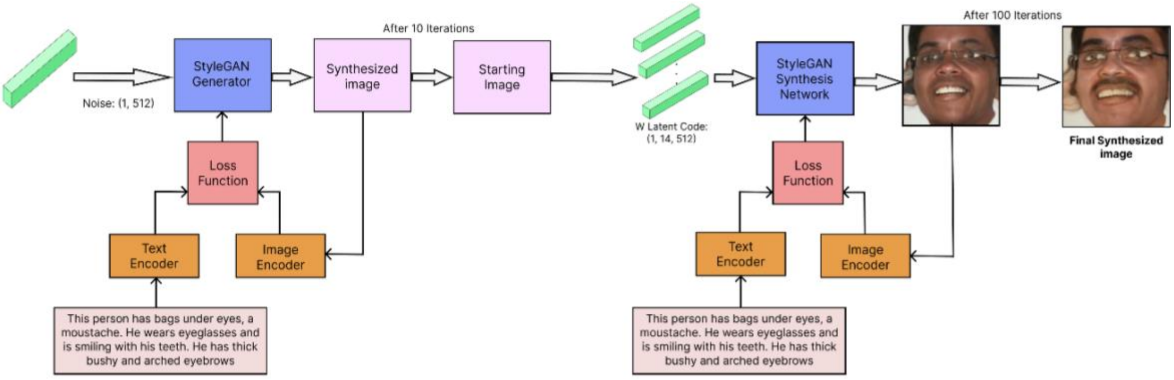

Noise: (1, 512)
StyleGAN Generator
After 10 Iterations
Synthesized image
Starting Image
Loss Function
Text Encoder
Image Encoder
This person has bags under eyes, a moustache. He wears eyeglasses and is smiling with his teeth. He has thick bushy and arched eyebrows
W Latent Code: (1, 14, 512)
StyleGAN Synthesis Network
After 100 Iterations
Final Synthesized image
Loss Function
Text Encoder
Image Encoder
This person has bags under eyes, a moustache. He wears eyeglasses and is smiling with his teeth. He has thick bushy and arched eyebrows

Figure 1. Overall architecture of the proposed model

3.5.3. Loss function

This loss function calculates how accurately the attributes in the caption are depicted in the image by finding the distance between their embeddings in the shared embedding space. There are two steps in finding the loss. First, the embeddings are normalized to the unit sphere, guaranteeing that the direction is constant and its magnitude is scaled to one. An arcsin function is used in the second step to determine the spherical distance loss between the image and text embeddings. It considers spherical geometry to find the distance between the embeddings on a hypersphere. Curvature is ignored in traditional Euclidean distance, making it unsuitable for points on a unit sphere like normalized vectors [28]. Therefore, spherical distance metrics are employed to calculate these distances accurately.

3.5.4. Algorithm

The proposed algorithm generates a suitable starting image latent code and then optimizes it to align closely with the input prompt. It begins by initializing the following key variables: the input text prompt X; an array of loss values La; and an array of W latent codes Wa, both initially empty. L is the loss function, G is the StyleGAN generator [9], [11], and C refers to CLIP or its variants. Ct is CLIP's text encoder, and Ci is its image encoder. In the first phase, the algorithm encodes the input text prompt and generates an initial latent code. Initially, X is encoded with Ct to get T. Z is assigned to a random tensor, size [1,512], and is mapped to W, the intermediate latent code, by passing it through G's mapping network. W is then appended to Wa, after which its image is synthesized by G's synthesis network and encoded with Ci. The encoded image is used along with T to compute the loss, with the loss function L, and this loss is appended to La. The image with the least loss is taken, after 10 iterations, and its intermediate latent code W is used in the next phase. The next phase uses the latent code with the lowest loss and improves the synthesized image. An optimizer is assigned W. An image F is synthesized from W and then encoded with Ci, after which its loss with T is back-propagated, prompting the optimizer to take a step. After S cycles, the final image F is produced, which most accurately represents the characteristics in the input text prompt.

Algorithm 1 Synthesizing an image from a text description

```
Input: $X$: text description, $C_t$: CLIP variant text encoder, $C_i$: CLIP variant image encoder,
$G$: StyleGAN generator, $S$: Number of steps, $L$: Loss function
Output: $F$: Final output image
    $T \leftarrow C_t(X)$
    $La \leftarrow [\ ]$
    $Wa \leftarrow [\ ]$
    For $i$ = 1 → 10 do
            $Z \leftarrow$ Create random tensor of size [1, 512]
```

```
        W ← G.mapping(Z)
        Wa ← Wa∪{W}
        Image ← G.synthesis(W)
        EncodedImage ← $C_i$(Image)
        loss ← L(EncodedImage,T)
        La ← La ∪ {loss}
end for
i←index(min(La))
W←$Wa_i$
Opt←Optimizer initialized on W with learning rate 0.03
for i =1 → S do
        F ← G.synthesis(W)
        EncodedImage ← $C_i$(F)
        Loss ← L(EncodedImage,T)
        Loss.backward()
        Opt.step()
end for
```

3.5.5. Evaluation metrics

To evaluate the quality and realism of generated images, several common evaluation metrics are used to compare them with real images. Two metrics used are Fréchet inception distance (FID) and learned perceptual image patch similarity (LPIPS).

- Fréchet inception distance: a commonly used metric to measure the realism and diversity of synthesized pictures is FID [29]. In contrast to the inception score, which solely focuses on the generated images, FID analyses the distribution of authentic and synthesized image sets. When using an InceptionV3 model, the features of the real and fake sets are extracted. Each pixel is changed to a numerical vector for the edges and lines. Fréchet distance between the embeddings is computed to measure the similarity of distributions. Higher image quality and realistic looks are marked by lower FID scores.
- Learned perceptual image patch similarity: perceptual loss is a metric that finds how structurally alike two high-dimensional images are. It uses a deep convolutional neural network to obtain intricate characteristics of images and determine how near the image patches' activations are to each other. Several layers in these deep CNNs effectively capture abstract visual representations. LPIPS [30] is useful in comparing the resemblance of a real image and its corresponding generated image. LPIPS is known to be comparable to human perception.

3.6. Experimenting with CLIP and its variants

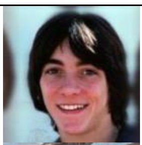
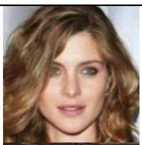
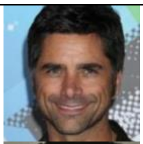
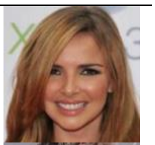
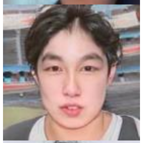
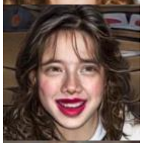
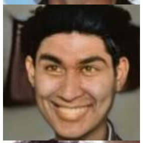
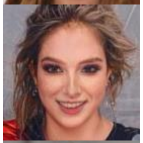
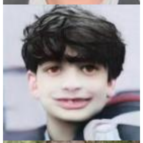
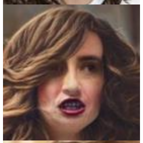
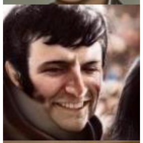
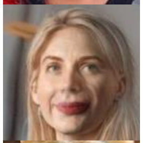
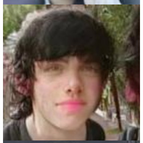
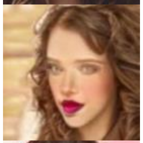
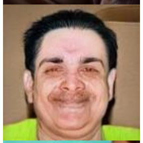
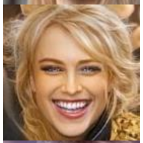
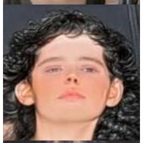
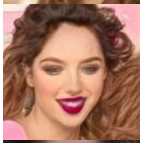
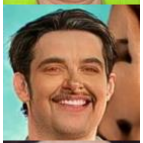
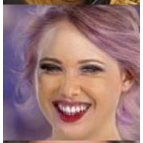
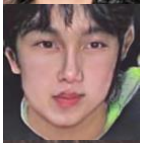
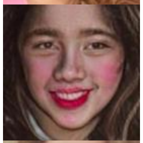
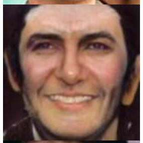
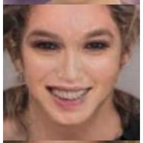
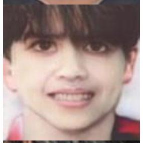
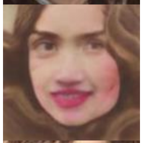
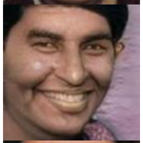
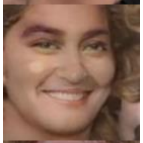
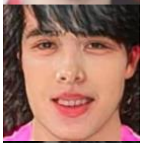
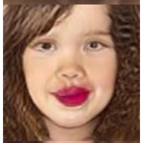
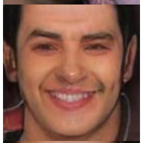
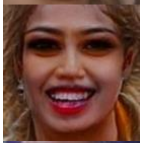
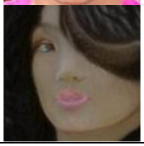
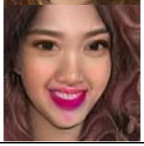
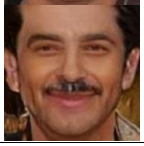
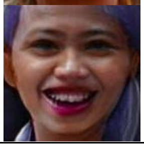
This study aimed to investigate the abilities of CLIP and its variants in various experiments and datasets. CLIP, MetaCLIP [17], and multilingual-CLIP were tested on a subset of 2,100 captions from the CelebA dataset. Due to the CLIP model's context length limitation, only the two longest sentences in each caption are used for evaluation. Analysis of the custom StyleGAN2-ADA model trained on a subset of 10,500 images from the Fairface dataset and the pre-trained FFHQ dataset StyleGAN2 model was another goal of this research. The system used a constant learning rate of 0.03 and 100 steps for CLIP and its variants in testing. To review the strengths of the CLIP variants, the ViT-B vision model, a vision transformer pre-trained with BERT, was utilized. For CLIP, the ViT-B/32 model, pre-trained on 400 million pairs of image and text was selected, MetaCLIP, the ViT-B-32-quickgelu and metaclip_fullcc were chosen to assess the MetaCLIP model trained on 1 billion image-text pairs, and in multilingual-CLIP, the ViT-B/32 vision encoder was used with the M-CLIP/XLM-Roberta-Large-Vit-B-32 text encoder. The multilingual-CLIP text encoder utilized RoBERTa, primarily designed for cross-lingual language processing tasks and pre-trained in 109 languages, which produces embeddings of size 512, making it a suitable fit for this analysis.

To evaluate CLIP, MetaCLIP, and multilingual-CLIP in generating realistic and accurate images, the LPIPS score, CLIP variant arcsin loss, and FID score [29] were used. The LPIPS score finds the perceptual similarity of the synthesized image to the original. The arcsin loss computes the distance between the generated image and the input description in the embedding space to see how well the image represents the characteristics in the caption. The FID score [29] measures how realistic the synthesized images are compared to the originals. A minimum of 2048 images, resized to 299×299 pixels, are required to calculate a valid FID score. The results of testing the algorithm with the different StyleGAN and CLIP variants are shown in Tables 1 and 2, where Table 1 exhibits the evaluation metrics and Table 2 depicts the images generated for four different text captions. An experiment was conducted to analyze the performance of each CLIP variant in generating an image from an input text and optimizing it to match the description better. The loss trajectory of CLIP and its variants, applied to the StyleGAN2-ADA model trained on the FairFace Dataset and the StyleGAN2 model pre-trained on the FFHQ dataset, are depicted in Figures 2 and 3. The X-axis in the graphs represents every 10th step of the 100 iterations and the Y-axis displays the corresponding loss values, showcasing each model's performance in generating an image over time. The algorithm ran for 100 steps with a learning rate of 0.03 and used the caption, "This person has bags under his eyes and a mustache. He has brown eyes. He wears eyeglasses and is smiling with his teeth. He has thick bushy, and arched eyebrows".

Table 1. Performance of different models in the algorithm

Model	CLIP	MetaCLIP	Multilingual-CLIP		CLIP	MetaCLIP	Multilingual-CLIP	
			English	Tamil			English	Tamil
	Pretrained StyleGAN2 -FFHQ				StyleGAN2-ADA-Fairface			
FID	39.82	40.95	46.30	50.44	63.39	72.88	73.40	86.54
Average LPIPS	0.54	0.53	0.55	0.57	0.62	0.61	0.63	0.64
CLIP variant loss	0.66	0.64	0.68	0.68	0.66	0.66	0.69	0.69

Table 2. Images generated with CLIP and its variants on the same text description

Source	Method	Column 1	Column 2	Column 3	Column 4
Original images from multi-modal-CelebA-HQ	Ground truth				
StyleGAN2-FFHQ	CLIP				
	MetaCLIP				
	Multilingual-CLIP English				
	Multilingual CLIP Tamil				
StyleGAN2-ADA FairFace	CLIP				
	MetaCLIP				
	Multilingual CLIP English				
	Multlingual CLIP Tamil				

Column 1: He is young and has mouth slightly open, bags under eyes, rosy cheeks, bangs, high cheekbones, double chin, wavy hair, and black hair.
Column 2: She has wavy hair, mouth slightly open, brown hair, and rosy cheeks and wears lipstick.
Column 3: This man is smiling and has pointy nose, black hair, bags under eyes, big nose, sideburns, and bushy eyebrows.
Column 4: The person is smiling, and young and has blond hair, narrow eyes, mouth slightly open, and high cheekbones. This person is wearing heavy makeup and wavy hair

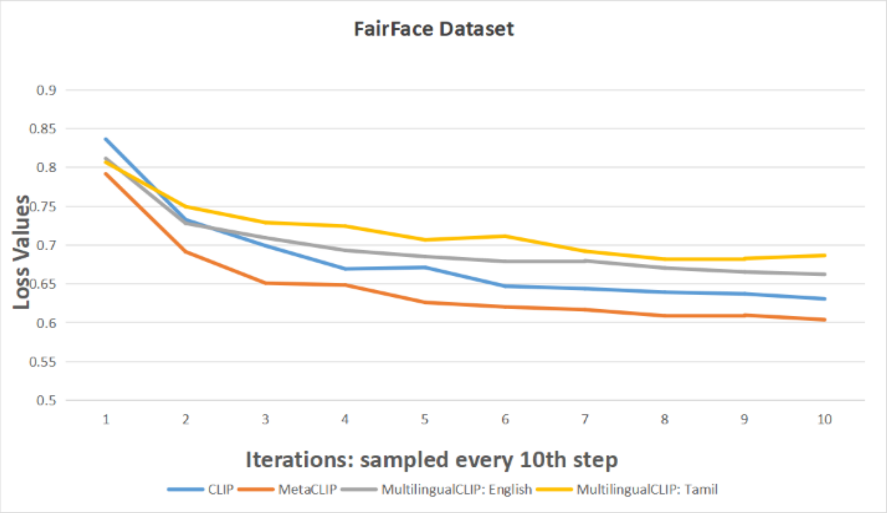


Figure 2. Graph comparing the loss values CLIP, MetaCLIP, multilingual-CLIP English, and multilingual Tamil for the FairFace dataset

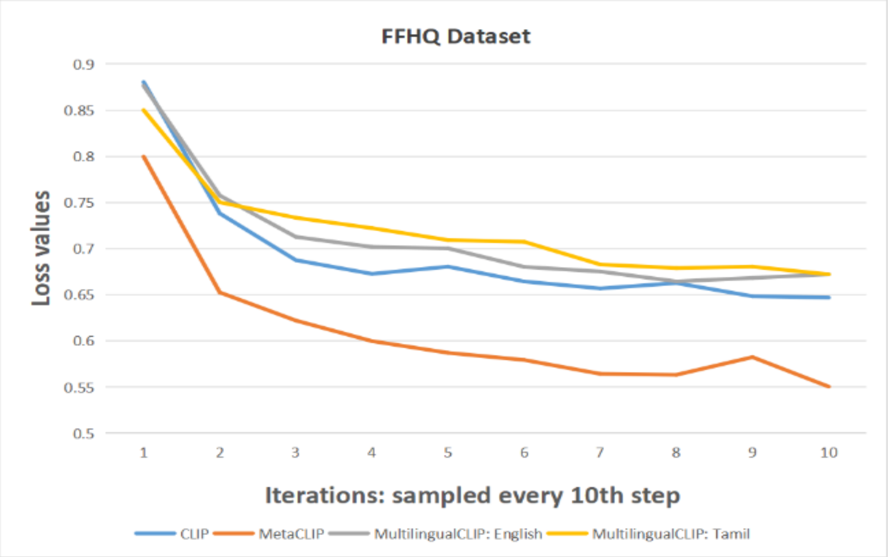


Figure 3. Graph comparing the loss values CLIP, MetaCLIP, multilingual-CLIP English, and multilingual Tamil for the FFHQ dataset multilingual Tamil for the FairFace dataset

4. RESULTS AND DISCUSSION

4.1. Results

Table 1 highlights the difference between CLIP and its variants in generating realistic and accurate images. The StyleGAN2 model pre-trained on the FFHQ dataset shows CLIP slightly outperforming MetaCLIP in FID scores [29], with both achieving low scores that display the high realism of the images. MetaCLIP exhibits better attribute representation than CLIP, indicated by its lower LPIPS score [30] and CLIP variant loss. Multilingual-CLIP [20] shows different results based on the language that it uses. In English, it has a higher FID score than CLIP and MetaCLIP, generating good-quality images but struggling to represent fine-grained attributes, characterized by its LPIPS score and CLIP variant loss. In Tamil, multilingual-CLIP has inferior performance, with higher LPIPS scores and CLIP variant loss, having significant errors in representing attributes such as hair and lip color, e.g. the input text blonde hair generates purple hair in Table 2.

The performance of the custom StyleGAN2-ADA model trained on the Fairface dataset showcases higher FID scores, which indicate less realistic images than the StyleGAN2 model pre-trained on the FFHQ dataset. For StyleGAN2-ADA, CLIP is much better at generating realistic images, compared to MetaCLIP, but both underperform compared to their counterparts using the StyleGAN2 FFHQ model. Multilingual-

CLIP falls behind CLIP and MetaCLIP in FID, LPIPS, and CLIP variant loss scores and has issues representing certain attributes in the text captions, particularly in Tamil.

Figures 2 and 3, discuss the loss curves and convergence of the CLIP variants and StyleGAN models. In Figure 2, MetaCLIP’s loss decreases quickly initially, slowing down around the 50th step. CLIP starts with a higher loss than the other CLIP variants but follows a similar trend to MetaCLIP. The multilingual-CLIP model shows a steady decrease in loss values, with the English model performing slightly better than the Tamil one. In contrast, the FFHQ StyleGAN2 model, in Figure 3, displays better performance for all the CLIP models than the custom FairFace StyleGAN2-ADA model. MetaCLIP achieves the best performance with a steeper and lower loss curve, while CLIP and multilingual-CLIP have similar courses, with an initially sharp decrease in the loss function that then gradually stabilizes.

4.2. Discussion

This paper analyses the performance of different StyleGAN models [9], [11] and CLIP variants in a training-free text-to-face generation model. Earlier research examined text-to-face generation with CLIP, but there has been little investigation of MetaCLIP, a recent development trained on one billion image-text pairs. Multilingual-CLIP has had inadequate research and can assist in multilingual text-to-face generation. This research aims to understand the advantages and disadvantages of CLIP and its variants by conducting a comparative study. Our investigation found that the StyleGAN2 model pre-trained on the FFHQ dataset surpasses the custom StyleGAN2-ADA model in realism, as indicated by the lower FID scores, however, the StyleGAN2-ADA models produced more diverse and fair images. The FFHQ model, trained on 70,000 images, showed better feature representation than the FairFace model [21], trained on only 10,500 images. The images synthesized by the pre-trained StyleGAN2 model are of superior quality and clarity, which aided CLIP and its variants in associating the extracted features of the synthesized images to their text descriptions, leading to better convergence. MetaCLIP synthesizes images that are the most accurate to the description, whereas CLIP creates the most realistic images. MetaCLIP's low LPIPS score could be due to its diverse and easily traversable embedding space. This is emphasized by MetaCLIP's sharp and low curve shown in Figure 3. CLIP displays the best realism, denoted by its FID scores and better convergence than the multilingual-CLIP model. Multilingual-CLIP’s Tamil and English performance was satisfactory in bridging the gap in multilingual text-to-image generation. Finally, testing durations for CLIP and multilingual-CLIP were shorter than MetaCLIP due to the large amount of information that MetaCLIP processes while testing.

According to our outcomes, it can be concluded that MetaCLIP's massive amount of high-quality training data contributes to the powerful and efficient navigation of its shared embedding space. This study reinforces MetaCLIP's superior performance in visual-linguistic tasks compared to CLIP. Multilingual-CLIP, which utilized RoBERTa, and trained on a massive body of data on CommonCrawl, had inaccuracies in face generation. The Tamil multilingual-CLIP has a less clearly defined embedding space and converges slower than the other CLIP models, as indicated by its scores. This could be due to the translation inconsistencies of certain words in the captions, which it did not encounter in training, which shows the need for further research on multilingual models. Our research had the following limitations. There were hardware and computational limitations that confined the training of the StyleGAN2-ADA model. Increasing the size of the FairFace training dataset and improving the quality of the data could enrich the realism of the model. The system presented has a reliance on pre-trained models, which may limit its nuance in text-to-face synthesis. Recent observations established MetaCLIP's ability to generate images closely aligned to the text and CLIP's ability to produce highly realistic outputs. Future research can explore the feasibility of the different configurations of CLIP's vision transformers, text encoders, and other loss functions. StyleGAN was used in our system to examine the CLIP variants, but diffusion models [2] can also be investigated. Multilingual-CLIP's performance in other languages can be improved by fine-tuning it with better captions. Extending this study to other vision-language models and newer CLIP variants can give us insight into other text-to-image generation approaches.

5. CONCLUSION

In this paper, we have successfully examined the performance of the visual-linguistic models, CLIP, MetaCLIP, and multilingual-CLIP in text-to-face generation using StyleGAN. This study utilized a training-free algorithm that iteratively optimized a latent code to generate a final image and compared the pre-trained StyleGAN2 model with a custom StyleGAN2-ADA model. Qualitative analysis was conducted with evaluation metrics like FID score and LPIPS and a loss function that finds the similarity of an input caption to an image. Since the algorithm is training-free it is highly dependent on the CLIP variants text-encoder. Our findings help us conclude that MetaCLIP outdoes CLIP and multilingual-CLIP in CLIP variant loss, which indicates that it closely matches the attributes in the given text. CLIP, on the other hand, synthesized the most realistic faces and had the lowest FID score. Despite multilingual-CLIP exhibiting lesser accuracy and realism, it provides a choice in the input text language, which creates more accessibility in text-to-face generation globally. However, it struggles to represent certain common features and exhibits entanglement between attributes. Future research can aim to improve these limitations and explore newer visual-linguistic models.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Priyadharsini Ravisankar	✓				✓	✓	✓	✓		✓	✓	✓	✓	
Shruthi Dhanvanth	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓			
Vaishnave Jenane Padmanabhan	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis

I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting

Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[2] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," *32nd International Conference on Machine Learning, ICML 2015*, vol. 3, pp. 2246–2255, 2015.

[3] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv-Computer Science*, pp. 1-7, Nov. 2014.

[4] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *4th International Conference on Learning Representations - Conference Track Proceedings*, 2015, pp. 1-16.

[5] T. Xu *et al.*, "AttnGAN: fine-grained text to image generation with attentional generative adversarial networks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1316–1324, 2018, doi: 10.1109/CVPR.2018.00143.

[6] H. Zhang *et al.*, "StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5908–5916, 2017, doi: 10.1109/ICCV.2017.629.

[7] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 12, pp. 4217–4228, 2021, doi: 10.1109/TPAMI.2020.2970919.

[8] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018, pp. 1-26.

[9] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8107–8116, 2020, doi: 10.1109/CVPR42600.2020.00813.

[10] T. Karras *et al.*, "Alias-free generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 2, pp. 852–863, 2021.

[11] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," *Advances in Neural Information Processing Systems*, 2020, pp. 1-11.

[12] D. M. A. Ayanthi and S. Munasinghe, "Text-to-face generation with styleGAN2," *Computer Science & Information Technology*, pp. 49–64, 2022, doi: 10.5121/csit.2022.120805.

[13] M. S. Sabae, M. A. Dardir, R. T. Eskarous, and M. R. Ebbed, "StyleT2F: generating human faces from textual description using styleGAN2," *arXiv-Computer Science*, pp. 1-10, Apr. 2022.

[14] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," *Proceedings of Machine Learning Research*, vol. 139, pp. 8748–8763, 2021.

[15] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, "StyleCLIP: text-driven manipulation of StyleGAN imagery," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2065–2074, 2021, doi: 10.1109/ICCV48922.2021.00209.

[16] W. Xia, Y. Yang, J. H. Xue, and B. Wu, "TediGAN: text-guided diverse face image generation and manipulation," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2256–2265, 2021, doi: 10.1109/CVPR46437.2021.00229.

[17] K. Frans, L. B. Soros, and O. Witkowski, “Clipdraw: exploring text-to-drawing synthesis through language-image encoders,” *Advances in Neural Information Processing Systems*, vol. 35, 2022.

[18] E. E. Reddy, M. M. M. Durga, M. J. Kishore, and V. Chaitanya, “Human facial image generation from textual descriptions using StyleGAN,” *Proceedings of the 2nd International Conference on Cognitive and Intelligent Computing*, pp. 131–141, 2022, doi: 10.1007/978-981-99-2742-5_14.

[19] H. Xu *et al.*, “Demystifying clip data,” *12th International Conference on Learning Representations, ICLR 2024*, 2024, pp. 1-20.

[20] F. Carlsson, P. Eisen, F. Rekathati, and M. Sahlgren, “Cross-lingual and multilingual CLIP,” *2022 Language Resources and Evaluation Conference, LREC 2022*, pp. 6848–6854, 2022.

[21] K. Karkkainen and J. Joo, “FairFace: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation,” *2021 IEEE Winter Conference on Applications of Computer Vision,* pp. 1547–1557, 2021, doi: 10.1109/WACV48630.2021.00159.

[22] O. R. Nasir, S. K. Jha, M. S. Grover, Y. Yu, A. Kumar, and R. R. Shah, “Text2FaceGAN: Face generation from fine-grained textual descriptions,” in *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, Sep. 2019, pp. 58–67, doi: 10.1109/BigMM.2019.00-42.

[23] S. Todmal, A. Mule, D. Bhagwat, T. Hazra, and B. Singh, “Human face generation from textual description via style mapping and manipulation,” *Multimedia Tools and Applications*, vol. 82, no. 9, pp. 13579–13594, 2023, doi: 10.1007/s11042-022-13899-5.

[24] E. Richardson *et al.*, “Encoding in style: a styleGAN encoder for image-to-image translation,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2287–2296, 2021, doi: 10.1109/CVPR46437.2021.00232.

[25] Y. Li, C. Y. Chang, S. Rawls, I. Vulić, and A. Korhonen, “Translation-enhanced multilingual text-to-image generation,” *61st Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 9174–9193, 2023, doi: 10.18653/v1/2023.acl-long.510.

[26] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.

[27] N. Parmar *et al.*, “Image transformer,” *35th International Conference on Machine Learning, ICML 2018*, vol. 9, pp. 6453–6462, 2018.

[28] K. Kobs, M. Steininger, and A. Hotho, “InDiReCT: language-guided zero-shot deep metric learning for images,” *2023 IEEE Winter Conference on Applications of Computer Vision, WACV 2023*, pp. 1063–1072, 2023, doi: 10.1109/WACV56688.2023.00112.

[29] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs trained by a two time-scale update rule converge to a local Nash equilibrium,” *Advances in Neural Information Processing Systems*, pp. 6627–6638, 2017, doi: 10.18034/ajase.v8i1.9.

[30] G. G. Pihlgren *et al.*, “A systematic performance analysis of deep perceptual loss networks: breaking transfer learning conventions,” *arXiv-Computer Science*, pp. 1-22, Jul. 2023.

BIOGRAPHIES OF AUTHORS

Priyadharsini Ravisankar is working as Professor in Department of Artificial Intelligence and Data Science, Rajalakshmi Engineering College, Tamil Nadu. She received her Ph.D. in Information and Communication Engineering from Anna University. She has more than 20 years of teaching experience. Her areas of research interest include underwater acoustic image processing, computer vision, and data analytics. She has published and presented her research works in reputed journals and conferences. She can be contacted at email: priyadharsini.r@rajalakshmi.edu.in.

Shruthi Dhanvanth has completed her undergraduate degree in B.E. Computer Science Engineering from Sri Sivasubramaniya Nadar College of Engineering, Tamil Nadu. Her interests are in software development and artificial intelligence. She has experience with machine learning and computer vision and has helped develop an interactive teaching model with MediaPipe. She can be contacted at email: shruthi2010113@ssn.edu.in.

Vaishnave Jenane Padmanabhan has completed her undergraduate degree in B.E. Computer Science Engineering from Sri Sivasubramaniya Nadar College of Engineering, Tamil Nadu. She is passionate about generative AI and software engineering. She has worked with LLMs to develop an in-house ChatBot with custom data. She can be contacted at email: vaishnave2010250@ssn.edu.in.