

BanglaCLIP: Bengali Text to Image Generation Using Few Shot Diffusion Model

by

**Mansuba Tabassum (192014014)
Mohammad Majhar Ali (173014029)
Sadia Sultana Satu (201014079)
Md. Shihan (193014059)**

*Project/Thesis report (CSE499) submitted in partial fulfillment of the
requirements for the degree of*

Bachelor of Science in Computer Science and Engineering

Under the supervision of

Dr. Mohammed Ashikur Rahman



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
UNIVERSITY OF LIBERAL ARTS BANGLADESH**

FALL 2024

© *Mansuba Tabassum, Mohammad Majhar Ali, Sadia Sultana Satu, Md. Shihan*
All rights reserved

DECLARATION

Project Title	BanglaCLIP: Bengali Text to Image Generation Using Few Shot Diffusion Model
Authors	<i>Mansuba Tabassum, Mohammad Majhar Ali, Sadia Sultana Satu, and Md. Shihan</i>
Student IDs	<i>192014014, 173014029, 201014079, and 193014059</i>
Supervisor	Dr. Mohammed Ashikur Rahman

We declare that this project/thesis report entitled *BanglaCLIP: Bengali Text to Image Generation Using Few Shot Diffusion Model* is the result of our own work except as cited in the references. The project/thesis report has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Mansuba Tabassum
192014014

Mohammad Majhar Ali
173014029

Sadia Sultana Satu
201014079

Md. Shihan
193014059

Department of Computer Science and Engineering
University of Liberal Arts Bangladesh

Date: January 30, 2025



Department of Computer Science and Engineering
University of Liberal Arts Bangladesh
Mohammadpur, Dhaka - 1207

CERTIFICATE

This is to certify that the project/thesis report entitled **BanglaCLIP: Bengali Text to Image Generation Using Few Shot Diffusion Model**, submitted by **Man-suba Tabassum** (Student ID: 192014014), **Mohammad Majhar Ali** (Student ID: 173014029), **Sadia Sultana Satu** (Student ID: 201014079), **Md. Shihan** (Student ID: 193014059) is a undergraduate student of the **Department of Computer Science and Engineering** has been examined. Upon recommendation by the examination committee, we hereby accord our approval of it as the presented work and submitted report fulfill the requirements for its acceptance in partial fulfillment for the degree of *Bachelor of Science in Computer Science and Engineering*.

Dr. Mohammed Ashikur Rahman, Assistant Professor
Dept. of CSE, University of Liberal Arts Bangladesh

Dr. Mohammad Golam Kibria, Professor and Head (acting)
Dept. of CSE, University of Liberal Arts Bangladesh

Suravi Akhter, Lecturer
Dept. of CSE, University of Liberal Arts Bangladesh

Md. Nazmul Abdal, Lecturer
Dept. of CSE, University of Liberal Arts Bangladesh

Place: Dhaka

Date: January 30, 2025

ACKNOWLEDGEMENTS

We would like to express our deep and sincere gratitude to our research supervisor, *Dr. Mohammed Ashikur Rahman*, for giving us the opportunity to conduct research and providing invaluable guidance throughout this work. His dynamism, vision, sincerity and motivation have deeply inspired us. He has taught us the methodology to carry out the work and to present the works as clearly as possible. It was a great privilege and honor to work and study under his guidance.

We are greatly indebted to our honorable teachers of the Department of Computer Science and Engineering at the University of Liberal Arts Bangladesh who taught us during the course of our study. Without any doubt, their teaching and guidance have completely transformed us to the persons that we are today.

We are extremely thankful to our parents for their unconditional love, endless prayers and caring, and immense sacrifices for educating and preparing us for our future. We would like to say thanks to our friends and relatives for their kind support and care.

Finally, we would like to thank all the people who have supported us to complete the project work directly or indirectly.

Mansuba Tabassum, Mohammad Majhar Ali, Sadia Sultana Satu, and Md. Shihaan

University of Liberal Arts Bangladesh

Date: January 30, 2025

Dedicated to
my loving parents
a good soul.

– *Mansuba Tabassum*

Dedicated to
ALLAH and
my loving mother.

– *Mohammad Majhar Ali*

Dedicated to
my loving parents.

– *Sadia Sultana Satu*

Dedicated to
my loving parents.

– *Md. Shihan*

ABSTRACT

Despite the increasing interest in vision-language models, there is still a wide gap when it comes to research around low-resource languages-such as Bengali. Creating realistic images from text finds diverse applications in the creative as well as semantic domains. Current vision-language models, such as CLIP (Contrastive Language - Image Pretraining) and its variants, have so far been targeting only high-resource languages such as English and have therefore paid no attention to the unique challenges posed by low-resource lingual settings. The existing multimodal retrieval systems have found it hard to bridge the gap between Bengali text and their corresponding images, especially because of lack of bangla text and image pair training dataset size, poorly developed language-specific models, and the non-availability of a satisfying CLIP adapter specifically designed for Bengali. On these grounds then we propose BanglaCLIP (Bangla Contrastive Language - Image Pretraining), an BanglaCLIP adapter integrated with the Stable Diffusion framework for the Bangla text-to-image generation. BanglaClip is based on text encoder BanglaBERT and image encoder ResNet50 with few-shot learning and transfer learning techniques, training which yields high-quality results possible with smaller datasets-such as BanglaLekha, Flickr8, and Bnature. And the similarity score of Banglaclip for different prompts was always above 95 percent. Incorporation of BanglaClip into the framework of the Stable Diffusion model gives superiority in the generation of images filled with syntactic and semantic richness of Bengali. Thus this development tries to bridge the knowledge gap by furthering research on Bengali in the field of vision-language, thereby creating more authentic and useful AI tools for Bengali-speaking communities. With BanglaClip, creative industry sectors and enterprises can drive such a wave of innovation as the rich linguistic and cultural heritage of Bengali finds its way through. The making of this BanglaClip adapter sets the spotlight on the need to fill in such intricate situations for low-resourced languages so that technology created using AI can be an initiative towards inclusivity for the betterment of society at large.

Keywords: Bangla, text-to-image, BanglaCLIP, Stable Diffusion, Few Shot.

Contents

1	Introduction	1
1.1	Problem statement	4
1.2	Aims, objectives and Motivation	4
1.2.1	Aims	4
1.2.2	Objective	5
1.2.3	Motivation	5
2	Literature Review	7
2.1	Vision-Language Models and CLIP	7
2.2	CLIP's Adaptation for Bangla Language and Few-shot Learning . . .	9
2.3	CLIP's Role in Stable Diffusion and Image Generation	10
2.4	Few-Shot Learning and Transfer Learning	10
2.5	Diffusion Models and Stable Diffusion	11
2.6	Hugging Face and Gradio for AI Development	11
2.7	Summary	11
3	Methodology	13
3.1	Data Collection:	13
3.2	Data Preprocessing and Data Augmentation:	17
3.3	Training Procedure	18
3.4	System Architecture	19
3.4.1	BanglaCLIP:	19
3.4.2	Stable Diffusion	21
3.5	Evaluation Metrics:	30
3.5.1	Recall Score:	31
3.5.2	Frechet Inception Distance (FID) Score:	32

3.5.3	Fréchet Score for Features (FSF):	33
4	Result Analysis	35
4.1	BanglaCLIP performance:	35
4.2	Training and Validation Loss:	35
4.3	Image-to-Text Recall:	36
4.4	Cosine Similarities:	37
4.5	Learning Rate Dynamics:	38
4.6	Similarity Analysis of Query Prompt Comparison:	38
4.6.1	Results	39
4.7	Performance of Bangla text-to-image generation model with Bangla Clip:	42
4.8	Performance of Bangla Text to image generation model without Bangla Clip:	45
4.9	Comparison of Results:	45
4.9.1	Performance of Bangla text-to-image generation model with Bangla Clip and integration with a Translation Algorithm: . .	46
5	Project Consideration	49
5.1	Project Overview	49
5.2	Project scope	50
5.2.1	Bengali Text to Image Generation:	50
5.2.2	Model Development and Integration:	51
5.2.3	Dataset and Preprocessing:	51
5.2.4	Model Evaluation:	51
5.2.5	User Interface Development:	51
5.2.6	Practical Applications:	52
5.3	Timeline	52
5.4	Development Tools	52
5.4.1	Hardware Requirements:	52
5.5	Software Requirements	53
5.6	Project Budget	53

6	Conclusion	54
6.1	Impact on Society	55
6.2	Impact on Environment	55
6.3	Ethical aspects	56
6.4	Sustainability Plan	57
6.5	Brief Summary	58
6.6	Future Scope:	59
	References	61

List of Figures

3.1	Scenario of Dataset- Flickr8k, BNature, BanglaLekha	13
3.2	Dataset of BNature	14
3.3	Dataset of BNature	14
3.4	Dataset of BanglaLekha	15
3.5	Dataset of BanglaLekha	15
3.6	Dataset of Flickr8k	16
3.7	Dataset of Flickr8k	16
3.8	Dataset of Hugging Face	17
3.9	BanglaCLIP	20
3.10	Stable Diffusion	24
3.11	U-Net architecture	25
3.12	Banglabert Architecture	26
3.13	Resnet50	27
4.1	Training and Validation loss	36
4.2	Validation Image-to-Text Recall	36
4.3	Training Image-to-Text Recall	37
4.4	Cosine Similarities	37
4.5	Learning Rate Progress	38
4.6	Comparison of raw and normalized average similarity scores	39
4.7	Similarity Score of result	40
4.8	Similarity Score of result	40
4.9	Similarity Score of result	41
4.10	Similarity Score of result	41
4.11	Similarity Score of result	42
4.12	Similarity Score of result	42
4.13	BanglaCLIP FID Scores	43

4.14 Feature Semantic Similarities score	44
4.15 FID and FIS scores	45
4.16 Translation Algorithm	47
4.17 Bangla Text to image generation result	47
4.18 Generated image	48
4.19 Bangla Text to image generation result	48

List of Tables

5.1	Project Budget	53
-----	--------------------------	----

Chapter 1

Introduction

In recent years, the exciting field of text-to-image generation aims to generate realistic and meaningful images from natural language descriptions. Existing models like CLIP and Stable Diffusion have worked quite successfully, but development has been predominantly around high-resource languages like English. There are still few works which focused on low resource languages like Bengali, Bengali is the fifth most populated language in the world, with more than 230 million speakers [Chen \(2023\)](#). These tools are limited in their processing capacity and are rather unavailable for Bengali speakers due to limited data and appropriate Clip adapter. Also The growing amount of digital data in the form of text, images, and videos has created a need for more advanced search systems that go beyond traditional keyword-based methods. Multimodal retrieval systems enable users to search using different types of inputs, like images or text, providing a more natural and intuitive way to access information. However, these systems have primarily been developed for high-resource languages, leaving low-resource languages like Bengali largely unexplored in this area. There is still a need for an encoder like BanglaClip that can effectively connect Bengali text and images, enabling various applications such as text-to-image generation, object detection with textual descriptions. Such an encoder would bridge the gap between textual and visual modalities, making multimodal search and image generation more accessible for Bengali users. By integrating natural language processing with digital text-to-image generation, a new and rapidly growing sector has emerged, significantly influencing research areas. This paper focuses on low resource languages like Bengali and employs the Bengali based CLIP using the Few Shot Diffusion Model to generate high-resolution images with limited data. The Bengali based CLIP addresses this disparity by

leveraging the principles of LowCLIP, which successfully adapted the CLIP model for low resource language like Azerbaijani [Asgarov \(2024\)](#). BanglaCLIP’s architecture is optimized for low-resource language like Bengali, to enable efficient and scalable multimodal retrieval for Bengali. Our proposed architecture BanglaClip also outperforms more than 95 percent similarity score for different prompts.

In this scenario, we develop Bengali based Clip-Adapter applied within the framework of stable diffusion for text-to-image generation with Bangla one. The other aims to extend developments that involve top-notch state-of-the-art methods in text and image processing to narrow the gap between Bangla text and image generation. In other words, for text processing these CLIP uses the BanglaBERT model as text encoder, which is an attention transformer-based model specifically trained on the Bengali corpus and which provides a good linguistic and semantic application for the Bengali-based text. Using BanglaBERT enables the model to learn complex sentence structures and the subtle meanings to be captured therein- for example, certain nuances in Bengali [Kowsher \(2022\)](#). For image generation, Bengali based Clip used ResNet50 as image encoder, a deep convolutional neural network, for encoding the image features and making the visual output fit well with the textual input. In the end, these elements are tied up to form a robust framework competent enough to generate detailed semantically accurate images from Bengali descriptions. One of the root causes for the difficulty in developing a text-to-image system for Bengali citizens is that large-scale, annotated datasets are often scarce. To overcome the challenge, bengali language based CLIP use few shot as transfer learning methodologies-to be reasonably effective within limited data. By utilizing datasets such as BanglaLekha, Flickr8, and Bnature, this CLIP learns to align text and images with minimum data. These datasets give a multiplicity of diverse cultural, natural, and daily sequences, making the images relatable and meaningful for Bengali-speaking individuals. The addition of this CLIP into this framework of Stable Diffusion equips the system with the potential to effectively deal with complex Bengali text inputs while producing high-resolution, contextually relevant images. The text processing pipeline ensures the accurate representation of linguistic intricacies, while the image processing pipeline focuses on producing visually coherent outputs with fine-grained details. This dual approach makes such CLIP an essential tool for addressing the gaps in AI accessibility for Bengali-speaking communities. With an estimated 230 millions population covered around the globe,

the Bengali people, through developments like Bengali based CLIP, could benefit richly. From the applications of the model over creative industries such as media, advertisement, and design to the world of education and communication, this CLIP allows Bengali speakers to harness the power of AI in the manner befitting their language and their culture. Further, in resolving the unique set of challenges that low-resource languages tend to face, Clip based on bengali language showcases how innovations in text and image processing could charge and empower under-resourced spoken languages. Within this paper, we explain the architecture, methodology, and training approaches employed, while processing text inputs of Bengali and its performance within the Stable Diffusion framework towards image generation. Bengali based CLIP, thus stands to become such a useful instrument in taking AI tools one step closer to getting inclusivity and efficacy among many Bengali-speaking communities.

Our Main Contributions:

- **Development of Bengali based CLIP for Connecting Bengali Text and Images:** This work introduces BanglaClip, a novel Bengali-specific CLIP model designed to bridge the gap between Bengali text and images, enabling accurate and contextually rich text-to-image generation. By integrating BanglaClip within the Stable Diffusion framework, we enhance the alignment between Bengali descriptions and generated visuals, ensuring semantically meaningful and high-resolution outputs. Additionally, BanglaClip can also be utilized for object detection and multimodal search engines, allowing users to retrieve images based on textual descriptions, making it a versatile tool for Bengali-language AI applications.
- **Advancing Few-Shot Learning and Text-to-Image Generation for Low-Resource Languages:** Our work contributes to the growing field of few-shot learning and text-to-image generation, specifically for low-resource languages like Bengali. By addressing challenges related to generating diverse images with varying structures, appearances, and levels of detail from Bengali descriptions, our system significantly improves realism and contextual accuracy. This development is particularly beneficial for industries such as media, entertainment, and education, where high-quality, language-specific image generation is essential.

- **Enhancement of Diffusion Models with Bengali CLIP Integration:** Diffusion models have demonstrated significant potential in generating high-quality text-to-image outputs. However, the lack of Bengali CLIP integration has often resulted in suboptimal performance, leading to less accurate and visually incoherent images. BanglaClip enhances the diffusion model by incorporating Bengali-specific text-image alignment, improving the quality of generated images while also extending its applications to object detection and multimodal search systems. This integration ensures that the diffusion model can effectively map Bengali textual input to corresponding visual outputs, generating high-resolution, culturally relevant, and semantically rich images. This work lays the foundation for future advancements in multimodal AI for Bengali, addressing key challenges in AI accessibility for low-resource languages while expanding the capabilities of text-to-image generation, object detection, and search engine applications.

1.1 Problem statement

Most of the CLIP(Constructive Language Pretraining) model are biased towards high resource language like English. Ignoring the low resource language like Bengali due to lack Bangla text and image pair dataset. So Bengali Specific Clip adapter can be developed with few shot method to solve this issue , which can be used for bengali text image generation or as search engine work.

1.2 Aims, objectives and Motivation

1.2.1 Aims

The project aims to develop Bengali based CLIP, an advanced text-to-image generation approach tailored to generating images based on text prompts in the Bengali language by integrating it with a stable diffusion Model with limited data. Making a user friendly web application for text to image generation. Aim of this project is to reduce the gap between bengali text and image developing a Bengali language based CLIP and make AI tools accessible for low resource languages like bengali for image generation or as search engine.

1.2.2 Objective

- Create a Bengali-specific CLIP adapter to bridge the gap between Bengali text and images, enabling high-quality text-to-image generation and multimodal applications.
- Making advanced AI tools accessible for Bengali users like for search engines, and image generation accessible and culturally relevant for Bengali-speaking communities.
- Demonstrate effective text-to-image generation and multimodal retrieval using limited datasets, paving the way for similar advancements in other low-resource languages.
- Making a user friendly web application for text to image generation.

1.2.3 Motivation

It was an inspiration to the existing gap in AI tools, considering the low-resource languages, especially Bengali, being the fifth most spoken in the world with more than 230 million speakers. Most of the pre-existing vision-linguistic models, including CLIP and Stable Diffusion, have focused mostly on high-resource languages, like English, totally ignoring the peculiarity in linguistic and cultural needs of Bengal-speaking communities. The Bengali language is morphologically heavy, with complex grammar; hence, the CLIP based on bengali language aims to solve these issues with the help of BanglaBERT-a fine-tuned Bengali text encoder. It also requires a visual encoder using ResNet50 to generate high-quality images from the text description. Another motivation for the project is the lack of an annotated dataset that connects Bengali text to an image; hence, training models in a conventional way was difficult, since such data would be a few in number. Few-shot learning and transfer learning methods used in BanglaClip design the model to work with limited data. This will have it that this CLIP, featuring Hugging Face in its development, will be well-placed to exploit a solid model training, fine-tuning, and deployment ecosystem, while Gradio provides easy interaction of the system with speakers of Bengali on very user-friendly interface grounds. In return, this will, for the first time develop in a culturally inclusive and critical AI solution

for the Bengali speakers and allows better integration of AI tools in the creative sector, education, and communication. By solving such challenges for low-resource languages, opportunities are set toward more equal reach within AI technologies, joined by encouraging multimodal systems' developments for those who have received historical neglect.

Chapter 2

Literature Review

The project Bengali specific CLIP based Stable Diffusion leverages existing work on vision-language models, contrastive learning, diffusion models, and low-resource language processing. This section reviews key related work and technologies that form the basis of our work.

2.1 Vision-Language Models and CLIP

Large vision-language models have recently attracted wide attention for their excellent performance in aligning textual and visual modalities. The CLIP model, proposed by [Radford \(2021\)](#), showed the efficiency of contrastive learning in connecting images with natural language descriptions without requiring extensive task-specific fine-tuning [Radford \(2021\)](#). This zero-shot learning capability of CLIP generalizes across multiple tasks, making it a very powerful tool for image-text retrieval and multimodal understanding. Vision-language models have grown to become one of the important building blocks of artificial intelligence, allowing smooth communication between textual and visual data. Among such models, probably the most notable one is Contrastive Language-Image Pretraining (CLIP), which was introduced by Radford et al. in Learning Transferable Visual Models From Natural Language Supervision [Radford \(2021\)](#). Contrastively trained, CLIP aligns the text and image embeddings by making them distinguishable from others within a common latent space and performs impressively for zero-shot classification, image retrieval, and multimodal reasoning with limited or no task-specific fine-tuning. Large vision-language models have transformed artificial intelligence by seamlessly integrating textual and visual modalities. Among the most relevant

advances in this domain is the Contrastive Language-Image Pretraining (CLIP) model, introduced by Radford (2021). CLIP employs a contrastive learning mechanism to align representations of text and image in a shared embedding space, thus allowing CLIP’s zero-shot learning as well as image retrieval and multimodal reasoning without task-specific optimization. Unlike traditional models, which rely on a labeled dataset, CLIP generalizes across tasks by training on vast image-text pairs sourced from the internet. It has two key components: a text encoder and an image encoder, jointly trained to maximize the similarity of correct text-image pairs while minimizing incorrect associations through a contrastive loss function. This allows CLIP to associate textual labels with visual representations in an effective way.

Building on top of CLIP, Huang (2025). proposed the Image2Text2Image framework, thereby improving the evaluation of image-to-text models. Diffusion models of image generation allow comparison against original images Huang (2025). Besides, it was proposed by Huang (2025)., the Image2Text2Image framework in Image2Text2Image: A Novel Framework for Label-Free Evaluation of Image-to-Text Generation with Text-to-Image Diffusion Models Huang (2025). In related work, Cui (2025). introduced CAMScore -cyclic reference free evaluation metric of image captioning, comparing the determined captions against reference images by means of raster-generated copies Cui (2025). It provides a new evaluation methodology for image captioning models by generating textual descriptions of images and further reconstructing images from those descriptions using text-to-image diffusion models. The similarity between the original and generated images provides a quantitative measure of how well the captioning model captures the image’s content. This label-free evaluation method underlines the importance of diffusion models in refining vision-language tasks, which is in tune with the objectives of BanglaCLIP’s integration into Stable Diffusion.

Recent studies have focused on advancing vision-language models with multilinguality aspects and fine-tuning the evaluation methodologies. Mu et al. proposed the multilingual prompting in text-to-image generation; they showed that giving the model a prompt in multiple bilingual languages significantly improved its performance Mu (2025). Further work of Cao (2024) included the development of VLEU (Visual Language Evaluation Understudy) which judges semantic qualities of images generated via text-to-image models across various prompts Cao (2024).

Another key development in this domain is the research of vision-language pretraining models such as BLIP-2, doing a good job of building lightweight model architectures that reduce the required computational overhead with no loss of performance [Li \(2024\)](#). Similarly, [Kim \(2022\)](#) studied cross-modal alignment strategies that would improve the generality of large multimodal models for practical applications [Kim \(2022\)](#).

2.2 CLIP’s Adaptation for Bangla Language and Few-shot Learning

There has been limited research on CLIP-type approaches for low-resource languages; although high-resource languages like English have been investigated, low-resource languages, such as Bengali, remain underexplored. One of the biggest challenges is perhaps the scarcity of large-scale Bengali image-text datasets and dedicated Bangla language encoders. Natural Language Processing (NLP) for Bengali has historically been underdeveloped due to the limited availability of large-scale datasets. The introduction of BanglaBERT by Bhattacharjee et al. addressed this issue by providing a transformer-based pre-trained model for Bengali text processing [Kowsher \(2022\)](#). BanglaBERT has demonstrated improvements in various NLP tasks, including text classification, named entity recognition, and sentiment analysis. Leveraging BanglaBERT as the text encoder in Bengali specific CLIP ensures that our model effectively captures the linguistic nuances and syntactic complexity of the Bangla language. And this CLIP uses ResNet50 as the image encoder, and the datasets such as Bangla Lekha, Flickr8, and Bnature [Kowsher \(2022\)](#).

Using few-shot learning in Bangla text-to-image generation has the added benefit of models generalizing from limited training data. The research done by [Zhou \(2022\)](#). shows that diffusion models can learn good from few-shot examples which also has a lower dependence on large-scale annotated datasets [Zhou \(2022\)](#). On the other hand, also investigated meta-learning techniques for vision-language tasks that allow transformer-based models to quickly adapt to new domains using significantly less data [Hu \(2023\)](#).

2.3 CLIP’s Role in Stable Diffusion and Image Generation

Incorporating CLIP modules within Stable Diffusion model architecture enhances the quality and contextual accuracy of the generated images. [Dalva & Yanardag \(2022\)](#). investigated the role of contrastive learning on diffusion models and showed that embedding text-image alignment mechanisms improves performance in text-to-image generation significantly [Dalva & Yanardag \(2022\)](#). Furthermore, [H. Ye & Ji \(2021\)](#). explored the CLIP-based embeddings in GAN models to refine image synthesis and boost visual coherence [H. Ye & Ji \(2021\)](#).

Another important aspect of vision-language modeling is automated evaluation metrics. [P. Grimal & Tourille \(2024\)](#) propose a fusion of Fréchet-Inception-Distance (FID) and CLIP similarity scores termed FID-CLIP to assess better the quality of image generation [P. Grimal & Tourille \(2024\)](#). At the same time, [Bent \(2024\)](#) proposed a scoring system called semantic consistency scoring, which checks alignment between textual descriptions and images generated from those descriptions [Bent \(2024\)](#).

With the constant developments in these domain-specific vision-language models, ethical considerations, such as the bias in text-image generations and conscious AI implementation, remain to be handled. Bias mitigation techniques in multimodal learning from [I. Alabdulmohsin & Zhai \(2024\)](#). centered around the importance of culturally sensitive datasets and fairness-aware contrastive learning approaches [I. Alabdulmohsin & Zhai \(2024\)](#). Furthermore, Wu et al. touched upon the environmental impact of large-scale text-to-image models and suggested improved training approaches that may reduce computation costs and carbon footprints [Wu \(2021\)](#).

2.4 Few-Shot Learning and Transfer Learning

Few-shot learning has been extensively explored in vision and NLP tasks to address challenges related to data scarcity. [Brown \(2020\)](#). introduced GPT-3, which demonstrated how large-scale transformer models can generalize well with minimal examples [Brown \(2020\)](#). Similarly, in computer vision, [F. Sung & Hospedales \(2017\)](#). proposed prototypical networks for few-shot classification, demonstrating

effective transfer learning techniques [F. Sung & Hospedales \(2017\)](#). The integration of few-shot learning in CLIP based on Bengali language allows the model to perform well even with limited training data, making it efficient and adaptable for Bangla text-to-image generation.

2.5 Diffusion Models and Stable Diffusion

Diffusion models have revolutionized generative AI by progressively refining images from noise. Ho et al. introduced Denoising Diffusion Probabilistic Models (DDPMs), which have since been widely adopted in text-to-image generation tasks [J. Ho & Abbeel \(2020\)](#). Stable Diffusion, an optimized implementation of DDPMs, was later developed by [Wolf \(2019\)](#). to allow for efficient, high-quality image synthesis with reduced computational requirements [Wolf \(2019\)](#). By incorporating bengali based CLIP into the Stable Diffusion framework, we ensure that Bangla text prompts generate contextually accurate and visually coherent images.

2.6 Hugging Face and Gradio for AI Development

Open-source AI platforms like Hugging Face and Gradio have significantly contributed to accessible model development and deployment. [A. Abid & Zou \(2019\)](#). introduced the Transformers library, which provides pre-trained models and fine-tuning tools for NLP tasks, making it easier to develop and share AI models [A. Abid & Zou \(2019\)](#). Similarly, Abid et al. developed Gradio, an open-source framework for building interactive machine learning applications [O. Ronneberger & Brox \(2015\)](#). By utilizing Hugging Face for model implementation and Gradio for deployment, we ensure that bengali based CLIP remains scalable, interactive, and accessible to researchers and developers.

2.7 Summary

Although CLIP has performed amazingly well for high-resource languages such as English [Radford \(2021\)](#), its application on low-resource languages like Bangla is still at large. First, large-scale Bangla image-text datasets do not exist in Bangla. Second, it lacks any kind of Bangla-specific encoders that prevent direct application

on CLIP Bengali text-to-image tasks. It surmounts most of the key limitations by adapting CLIP architecture in bengali based CLIP, using BanglaBERT as the text encoder and ResNet50 as an image encoder, finetuned on Bangla-specific datasets like BanglaLekha, Flickr8, and Bnature. This allows it to grasp even the nuances of Bangla with all the major advantages of contrastive learning preserved. Therefore, the CLIP based on bengali language brings in much better performance in generating text-to-image generation in bridging Bengali textual descriptions with generated visuals, therefore extending the potential use in creative industries, education, and content creation among Bengali-speaking communities through the Stable Diffusion framework. By using few-shot learning techniques, bengali language based CLIP improves performance even with limited training data. By extending vision-language models like CLIP to low-resource languages through adaptations such as bengali based CLIP, the gap in Bengali vision-language research is significantly reduced. However, CLIP pretraining is highly dominated by high-resource languages like English, making representations of low-resource languages scarce. The above-mentioned limitation leads us to the need for BanglaCLIP: a vision-language model specific to Bengali text-image alignment. Future research should focus on expanding Bangla-specific datasets, improving the model's generalization capabilities, and integrating more advanced architectures to further enhance multimodal performance. The research should increase Bangla-specific datasets, enhance BanglaCLIP generalization, and formulate fresh multimodal architectures for low-resource languages. Improvements in few-shot learning and contrastive learning strategies will unblock the path for AI-based vision-language applications to become real tools for Bengali-speaking societies.

Chapter 3

Methodology

3.1 Data Collection:

The dataset is adopted from BanglaLekha, Flickr8, and Bnature, consisting of a well-rounded supply of image-text pairs. To support Bengali, captions from existing datasets (e.g., MS COCO) were translated into Bengali using a machine translation algorithm. These datasets contained the kinds of scenes, objects, and daily lives depicted in texts in the Bengali language and, therefore, add much to the training of BanglaCLIP by ensuring the dataset realistically presents different scenarios in line with Bengali speakers.

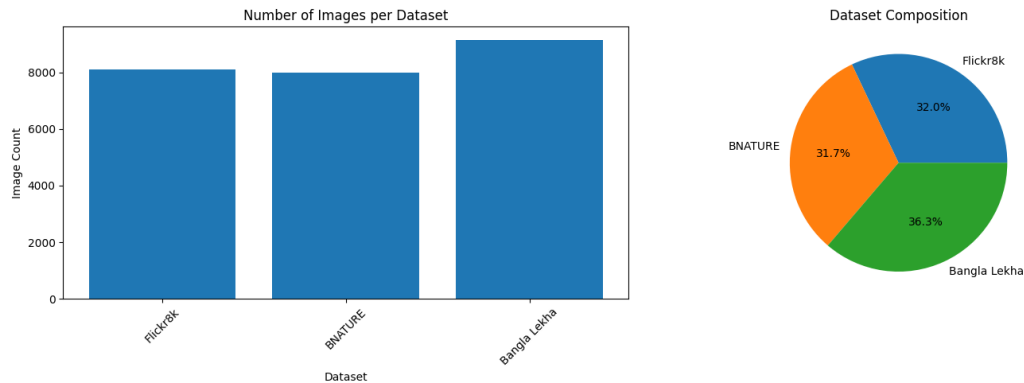


Figure 3.1: Scenario of Dataset- Flickr8k, BNature, BanglaLekha

The resulting dataset consists of:

1. **BNature:** 39105 image caption



Dataset: BNATURE
Total Entries: 39105
Image Count: 8000
Sample Captions:

- {'caption_id': '1.jpg', 'bengali_caption': 'গ্রামে হাঁটা দুই শিশু।'}
- {'caption_id': '1.jpg', 'bengali_caption': 'দুই শিশু গ্রামে হাঁটছে।'}
- {'caption_id': '1.jpg', 'bengali_caption': 'একটা নীল জামা পড়া শিশু।'}

Figure 3.2: Dataset of BNature
A. M. Faruk & Morol (2020)

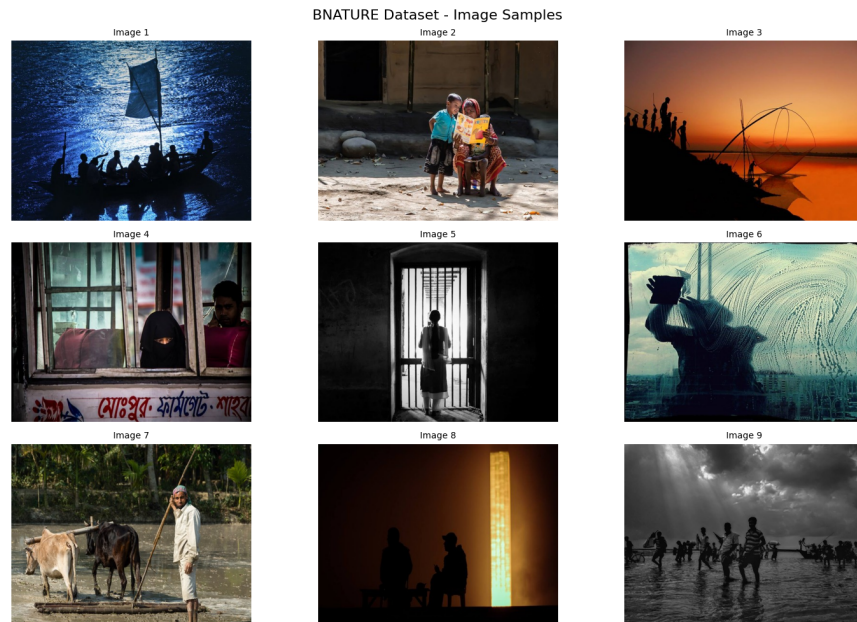


Figure 3.3: Dataset of BNature
A. M. Faruk & Morol (2020)

2. BanglaLekha: 9154 image caption

```

Dataset: Bangla Lekha
Total Entries: 9154
Image Count: 9154
Sample Captions:
- {'filename': '1.png', 'caption': ['তিন জন মেয়ে মানুষ আছে। এক জন দাড়িয়ে আছে আর দুই জন বসে আছে।', 'একটি হলুদ জামা পায়জামা পরা মহিলা দাড়িয়ে হাতে একটি বে']}
- {'filename': '2.png', 'caption': ['অনেক মেয়ে মানুষ বসে আছে।', 'একটি নীল জামা পরা মহিলা একটি নীল ল্যাপটপ এর দিকে তাকিয়ে আছে এবং পিছনে তার দিকে বসে শারি']}
- {'filename': '3.png', 'caption': ['অনেক মানুষ একসাথে বসে কাজ করছে।', '২ টি ছোট ছেলে একজন শর্ট প্যান্ট দাড়িয়ে চেয়ে আছে আরেকজন বসে গ্যামেস্টস এ কাজ করছে']}

```

Figure 3.4: Dataset of BanglaLekha
Mansoor (2019)

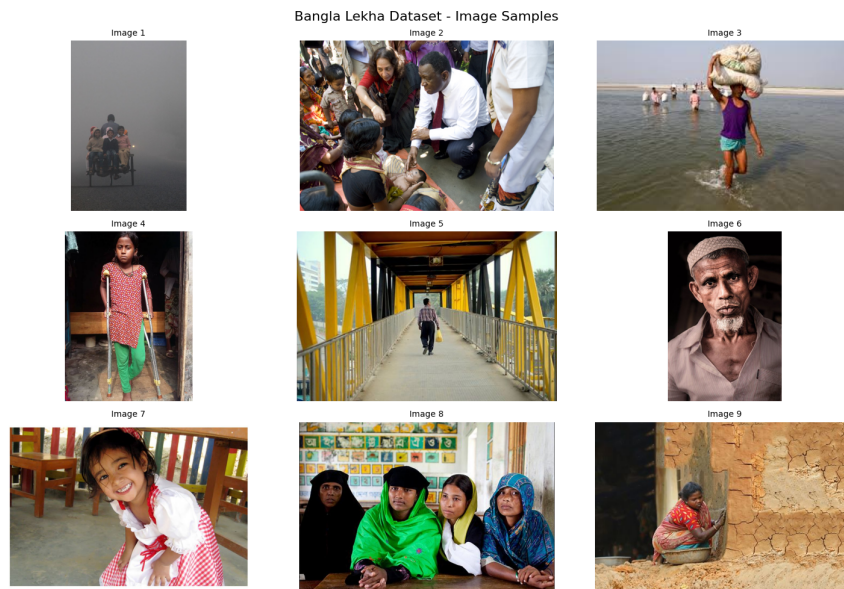


Figure 3.5: Dataset of BanglaLekha
Mansoor (2019)

3. **Ban CAP:** 40455 image caption .

```

Dataset: Flickr8k
Total Entries: 40455
Image Count: 8091
Sample Captions:
- ('caption_id': '1000260201_693b08c0ee.jpg#0', 'english_caption': 'A child in a pink dress is climbing up a set of stairs in an entry way .', 'bengali_caption': 'একটি সোপা
- ('caption_id': '1000260201_693b08c0ee.jpg#1', 'english_caption': 'A girl going into a wooden building .', 'bengali_caption': 'একটি মেয়ে শিশু একটি কাঠের বাড়িতে ঢুকছে।
- ('caption_id': '1000260201_693b08c0ee.jpg#2', 'english_caption': 'A little girl climbing into a wooden playhouse .', 'bengali_caption': 'একটি ছোট মেয়ে কাঠের খেলঘরে উঠছে

```

Figure 3.6: Dataset of Flickr8k
Khan et al. (2022)

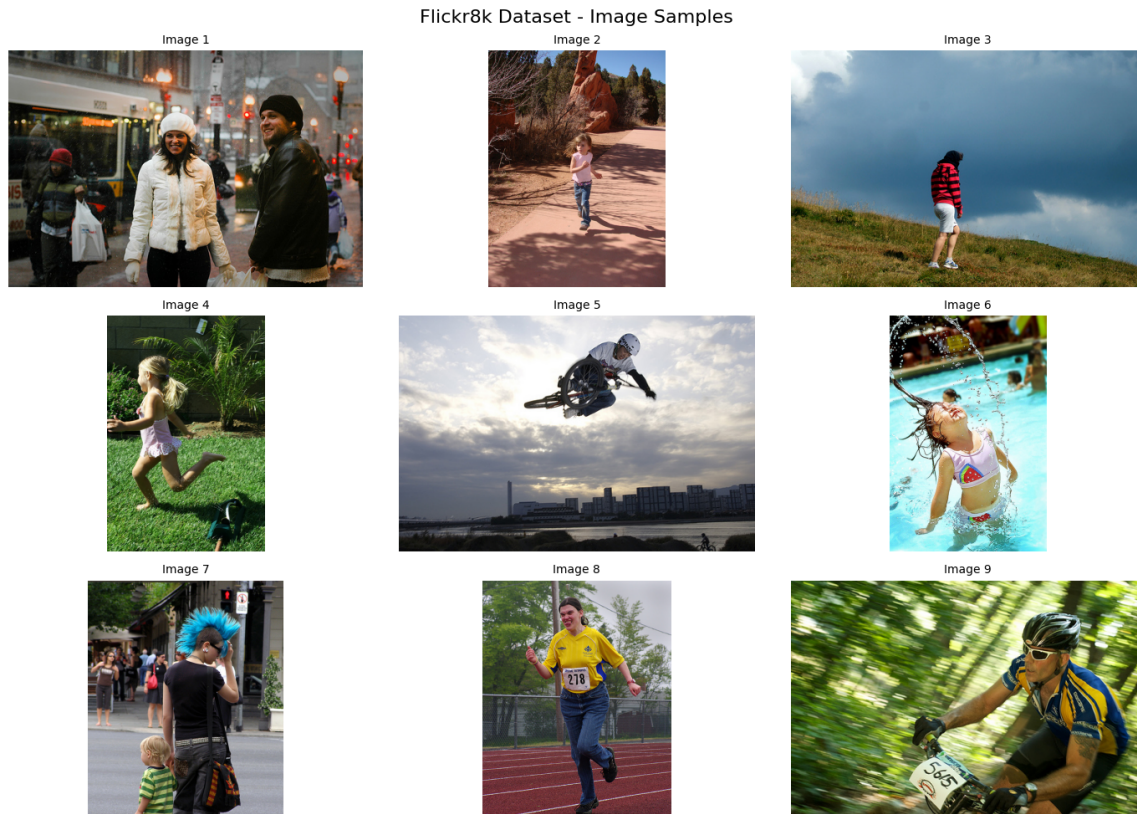


Figure 3.7: Dataset of Flickr8k
Khan et al. (2022)

Dataset was pre processed and then combinedly uploaded in huggingface hub. [Link: Tabassum \(2025\)](#)

image image · width (px)	caption string · lengths
158~679 94.6%	93~124 5.5%
	['একজন মানুষ ধান ছাড়ছে।', 'একটি লোক ধান ভাংছে।']
	['একজন নারী ও কয়েকজন পুরুষ আছে।', 'একটি লোক দুটো পোস্টার ধরে সবাইকে দেখাচ্ছে।']
	['একটি শিশু একজনের হাত ধরে আছে।', 'একটি ছোট ছেলে কাপড় পড়ে তার বাবাকে ধরে রাখছে।']
	['একজন পুরুষ ক্যামেরা দিয়ে ছবি তুলছে।', 'একটি লোক ডিজিটাল ক্যামেরা দিয়ে ছবি তুলছে।']
	['একজন পুরুষ ও দুইজন বালক আছে।', 'সমুদ্রের পাশের কাছে কয়েকজন ছেলে ক্রিকেট খেলছে।']

Figure 3.8: Dataset of Hugging Face

3.2 Data Preprocessing and Data Augmentation:

Cleaning data involves, among other things, normalizing the text, tokenizing the data, and converting everything into embeddings using BanglaBERT so that the model could read inputs written in the Bengali language accurately. The image denoising involves resizing, normalizing, and feeding the images into the latent space via ResNet50. These methods assure that both the text and image modalities remain compatible and consistent with each other.

Different kinds of augmentations are employed to enhance and generalize those images. For instance, images undergo random cropping, rotation, flipping, and contrast adjustments. This achieves a variation analogous to how the model validates a clear understanding of varying data types through this bulk of other cases.

BanglaBERT basically gives text inputs semantic embeddings, while deep visual features are derived from the images with the help of ResNet50. It aligns text and visual features in a shared embedding space using contrastive learning, enabling proper associations between Bengali text and corresponding images.

The preprocessed and augmented dataset trains the Stable Diffusion model, going through a systematic stepwise denoising diffusion process (DDPM). Latent

representations are compressed through image VAE encoding and allow the decoder to generate high-resolution outputs, ensuring proper contextual alignment with the Bengali text prompts.

Training and fine-tuning are performatively conducted by Hugging Face and Gradio, a suitable environment for Faced development and interactive testing. They allow for iterative and continuous improvement of the entire pipeline in ensuring that BanglaCLIP generates good quality most contextually aligned images from the Bengali text inputs.

3.3 Training Procedure

The CLIP training procedure for Bangla image-text pairs utilizes a dual-encoder architecture, combining ResNet50 for image encoding and BanglaBERT for text encoding. Projection heads refine the embeddings into a common 256-dimensional space, ensuring meaningful alignment. The training process employs contrastive learning, where image-text pairs are compared using a normalized temperature-scaled cosine similarity metric. The InfoNCE loss function encourages the model to maximize similarity between matched pairs while minimizing it for mismatched ones. Optimization is performed using AdamW with gradient clipping and weight decay, alongside a cosine decay learning rate schedule for stable convergence. Training metrics such as loss, similarity score, and recall@k are tracked, with validation-based checkpointing to ensure optimal performance.

Algorithm 1 CLIP Training Procedure for Bangla Image-Text Pairs

- 1: **Input:** Training dataset D containing paired images and text
 - 2: **Parameters:** Temperature parameter $\tau = 0.07$, Batch size N , Learning rate η , Number of epochs K
 - 3: **Initialization:** Initialize ResNet50, BanglaBERT, projection heads g_I, g_T , and AdamW optimizer
 - 4: **for** $k = 0$ to K **do**
 - 5: **for** each batch B in D **do**
 - 6: Extract features z_1 (image), z_2 (text)
 - 7: Compute embeddings $E_k \leftarrow g_I(\text{ResNet}(z_1)) + g_T(\text{BERT}(z_2))$
 - 8: Compute similarity scores: $S_k = \frac{\cos(E_k^{\text{img}}, E_k^{\text{text}})}{\tau}$
 - 9: Compute contrastive loss: $L(S_k) = -\log \frac{\exp(S_{k_{ii}})}{\sum \exp(S_{k_{ij}})}$
 - 10: Update parameters using AdamW
 - 11: Compute training metrics: train_loss, similarity, recall@k
 - 12: Update learning rate using cosine decay schedule
 - 13: **end for**
 - 14: Validate and save checkpoint if improved
 - 15: **end for**
-

The dataset is then used in the training of a U-Net-based denoising model during the Stable Diffusion phase, where images are noised and then denoised to learn the reverse process of diffusion. Further, it uses the VAE encoder to compress images for lower dimensionality and complexity while the scheduler optimizes the process of noise removal at generation time. The training is done on Hugging Face and Gradio, taking advantage of pre-trained models, fine-tuning them on Bengali datasets so that they perform well under low-resource conditions. The structured process of training has ensured that BanglaCLIP generates high-resolution images from Bengali text prompts with semantic accuracy and contextual relevance.

3.4 System Architecture

3.4.1 BanglaCLIP:

BanglaCLIP's system architecture incorporates a varied set of enhanced AI components to optimize Bengali text-to-image generation. BanglaCLIP is ultimately derived from a contrastive learning mechanism that associates Bengali text descriptions with images. The system starts when the user inputs a Bengali text description, by which the BanglaBERT encoder—a pre-trained model specifically designed for

Bengali-processes this text. BanglaBERT makes an effort to encode the complex syntactic and semantic features of the Bengali language into the dense vector representation of the input text. This textual representation is, in turn, passed to the Stable Diffusion model, which is a generative model that generates a high-resolution image from random noise by iteratively refining it, guided by the semantic content captured by BanglaBERT in the feature representation. This generated image once

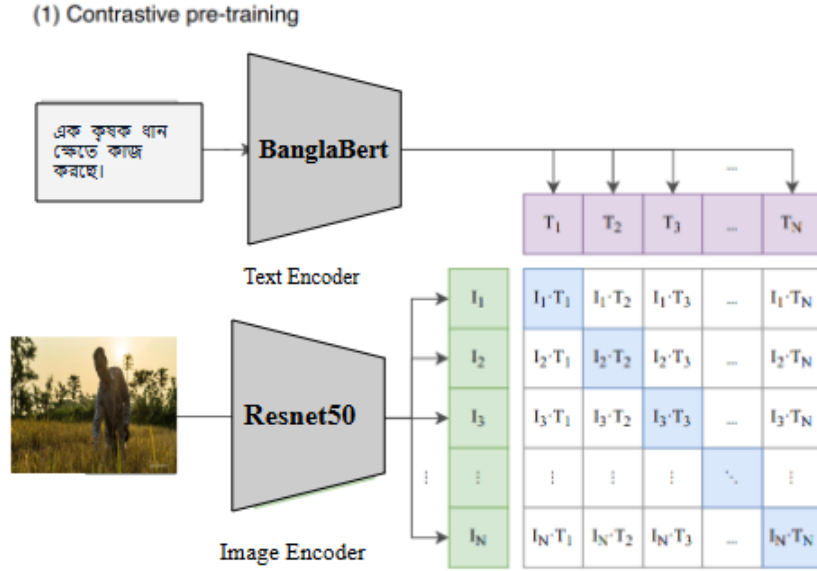


Figure 3.9: BanglaCLIP

again comes into play when we pass it through the ResNet50 encoder, which encodes high-level visual features to ensure that the image generated corresponds to the input text. Limited annotated datasets of Bengali text-to-image present challenges, as few-shot learning and transfer learning come to the rescue. With a few labeled samples, these techniques allow the model to generalize well even with them. This training is performed on BanglaLekha, Flickr8, and Bnature. The images are thus generated with contextuality and cultural relevance. The whole system is developed and tuned through the Hugging Face platform that provides the necessary tools for model integration and optimization. The user interacts with the system through a Gradio interface, making it easy to input Bengali text and display the generated images. The architecture thus integrates BanglaClip, BanglaBERT,

ResNet50, Stable Diffusion, and few-shot learning to render specific strength to this Bengali text-to-image-generation and facilitate the Bengali speakers with a highly capable and ready-to-use AI tool for creative and professional use.

BanglaCLIP development comprises advanced models and technologies for the possible development of the image-generation system, which uses the Bengali language. Such work would be directed at bridging the gap in translation between Bengali text and image generation through the integration of the latest in technology and improving existing ones-in view of fostering further improvement.

3.4.2 Stable Diffusion

1. **Text to image generation with Stable Diffusion Model:** Diffusion models, such as Stable Diffusion, generate high-quality images by progressively removing noise from a noisy image over multiple steps through forward diffusion, reverse diffusion, and conditional generation. In forward diffusion, noise is gradually added to a clean image x_0 over timesteps TTT , transforming it into pure Gaussian noise, following Equation (1):

$$x_t = \alpha_t x_0 + \sqrt{1 - \alpha_t} \epsilon$$

Where x_0 is the clean image, $\epsilon \sim \mathcal{N}(0, I)$ is random noise sampled from a Gaussian distribution, and α_t is a noise schedule parameter that controls the amount of noise added at timestep t . The noise schedule α_t ensures that the image becomes progressively noisier until it resembles pure noise at the final timestep T . In reverse diffusion, the process starts with random noise and gradually removes noise over multiple timesteps to generate a clean image, guided by a U-Net, which predicts the noise to be removed at each step. Mathematically, at each timestep t , the U-Net predicts the noise $\epsilon_\theta(x_t, t)$ in the noisy image x_t , and the predicted noise is subtracted from x_t to refine the image using Equation (2):

$$x_{t-1} = \frac{1}{\alpha_t} \left(x_t - \frac{\sqrt{1 - \alpha_t}}{1 - \alpha_t} \epsilon_\theta(x_t, t) \right)$$

where x_{t-1} is the refined image at the next timestep, and $\epsilon_\theta(x_t, t)$ is the noise predicted by the U-Net, conditioned on the timestep t . This process is repeated iteratively for T timesteps, gradually transforming random noise into a coherent image. In text-to-image generation, the reverse diffusion process is guided by a textual prompt (e.g., “a sunset over a river”), embedded into the U-Net’s architecture, ensuring that the generated image aligns with the textual description. The text prompt is first encoded into a text embedding using a language model (e.g., CLIP or BanglaCLIP), capturing its semantic meaning. The text embedding is then injected into the U-Net using cross-attention layers, allowing the U-Net to condition its noise predictions on the text prompt. The conditional noise prediction follows Equation (3):

$$\epsilon_\theta(x_t, t, c) = \text{UNet}(x_t, t, c)$$

where c is the text embedding. The scheduler plays a crucial role in controlling the noise removal process, ensuring stability and gradual refinement. The scheduler determines the scaling factor for the predicted noise before it is subtracted from the noisy image, preventing artifacts or instability in the generated image. In the training phase, forward diffusion is used to add noise to clean images over multiple timesteps, and the U-Net is trained to predict the noise added at each timestep. The model minimizes the difference between the predicted noise and the actual noise using a mean squared error (MSE) loss, given in Equation (4):

$$L = E_{x_0, \epsilon, t} [\|\epsilon - \epsilon_\theta(x_t, t, c)\|^2]$$

where x_0 is the clean image, ϵ is the actual noise, and $\epsilon_\theta(x_t, t, c)$ is the predicted noise conditioned on the text prompt c . During the generation phase, the process begins with random noise sampled from a Gaussian distribution, and the U-Net iteratively predicts and removes noise at each timestep, guided by the text prompt. After T timesteps, the refined image x_0 is generated. The complete process can be summarized mathematically with forward diffusion using Equation (1), reverse diffusion using Equation (2), conditional noise prediction using Equation (3), and the loss function using Equation (4). By

combining forward diffusion, reverse diffusion, and text conditioning, diffusion models like Stable Diffusion can generate high-quality images from textual descriptions, leveraging the power of iterative noise removal and semantic alignment.

2. **BanglaCLIP in Diffusion Model:** BanglaCLIP is a special variant of CLIP which is fine-tuned for the Bengali language. CLIP generates text-image pairs with a contrastive pre-training method. In the Diffusion model, BanglaCLIP plays an important role as a bridge between the textual and visual modalities. The Diffusion model is an iterative process that generates images by progressive denoising of random noise. BanglaCLIP plays a vital role in keeping the image generation process in tune with the meaning and semantics of the Bengali input text. Trained on a large dataset of text-image pairs in Bengali, BanglaCLIP learns to associate textual descriptions with images even in low-resource settings, allowing the model to generate high-quality, relevant images from Bengali text. The input is a Bengali text description. The input is embedded by BanglaCLIP into a dense, contextually rich vector space that captures the content of the Bengali description. A Diffusion model accepts this vectorized representation and begins to generate an image from random noise. The Diffusion model, in an iterative process, keeps refining the image based on the guidance provided by BanglaCLIP until it produces the final output. While it does this noise reduction, the semantic information from the BanglaCLIP supports making sure the resulting image correctly reflects both the content and context of what is described in the text. This yields a high-resolution image that is pertinent to the generation based on the Bengali description.
3. **BanglaCLIP in Stable Diffusion:** Stable Diffusion is a deep generative model that creates high-quality images from text prompts. This model generates images by progressively refining random noise through a series of denoising steps until it reaches a coherent image. The Stable Diffusion framework is used as the base for integration with BanglaCLIP, hence providing semantic understanding in Bengali text. A given Bengali text input gets embedded into vector representations with the help of BanglaCLIP, which serves as a guiding force for the Stable Diffusion model in order to generate an image.

This enables the Stable Diffusion model to generate images which are not only visually correct but contextually relevant with respect to the provided Bangla description. Integration with BanglaCLIP allows the system to do text-to-image generation for Bengali-speaking people and addresses the resource gap in low-resource languages.

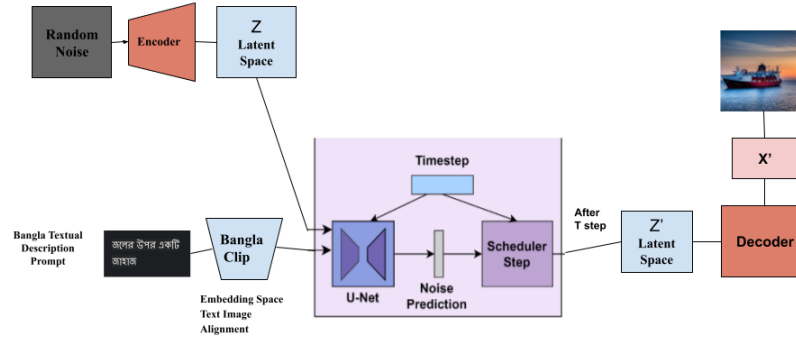


Figure 3.10: Stable Diffusion

When a user inputs a Bengali text description into the system, BanglaCLIP first encodes the text into a high-dimensional vector representation that captures the meaning of the text. This is then passed to the Stable Diffusion model. The Stable Diffusion model operates by first generating a random noise image and then refining it through multiple steps. While regressing at every step, it cleans up the incoherent images to look realistic. A vector generated through BanglaCLIP guides that refinement process, keeping the generated image aligned with what was depicted through Bengali text. Stable Diffusion, coupled with BanglaCLIP, produces a final image that is high-resolution and contextually correct with respect to the input text through successive refinements.

4. **UNet:** UNet is a deep learning architecture that originally found its adaptation in image segmentation tasks and currently finds applications in generative models like Stable Diffusion. It is known as UNet due to the U-shaped architecture comprising one usually contracting path normally referred to

as the encoder, and one expanding path, which is usually referred to as the decoder. The encoder captures the high-level features from the input data while the decoder reconstructs the image from these features. This important UNet characteristic is usually constituted from its encoder to its decoder, joined using the famous "skip connections". For the iterative refinement to refine the results gradually into cleaner, high-quality pictures starting with random noisy frames, one of the very prominent UNet architectures applies to this stable diffusion in creating stunning texts-to-images that maintain their minutest features and keep generating quite smooth texts without disturbance.

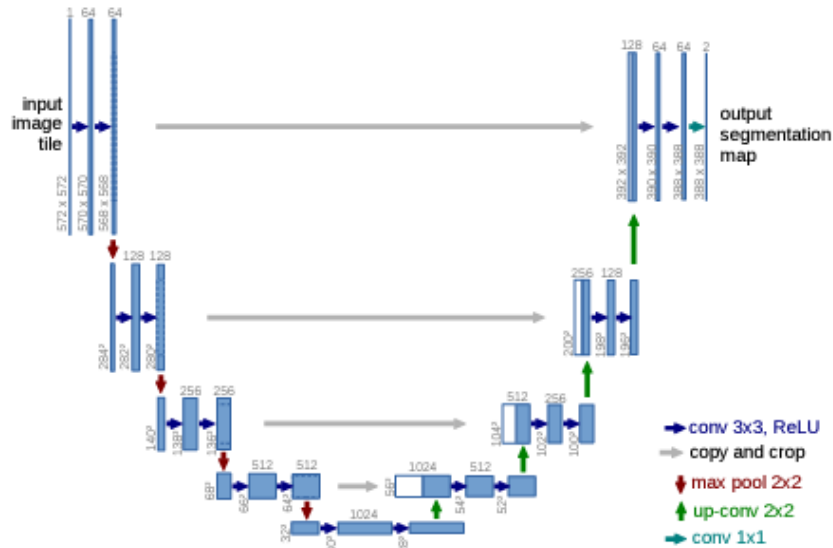


Figure 3.11: U-Net architecture
O. Ronneberger & Brox (2015)

In Stable Diffusion, the UNet model plays an important role in this refinement process. In this model, UNet progressively transforms a noisy image into a clear, high-quality image. So, in essence, the encoder part of UNet takes the features extracted from the inputting image or even noise, while at decoding, it stretches these out into reconstructed images. So, skip connections would basically help the encoder part of the UNet ensure that important information with regard to spatial specifics is preserved along the transformation phase. At every step of diffusing, UNet cleans away the noise progressively, enhancing

this toward where this generated image better fits visually toward what Bengali text was describing. The UNet aids in generating fine details and maintaining the overall coherence of the image, while the convergence of the Stable Diffusion model takes place to a final image output.

5. **Text Encoding with BanglaBERT:** The BanglaBERT model is used as the text encoder in BanglaClip. Since this model is particularly created with consideration to syntax and subtleties-semantic and others-of Bengali, the process of training this highly sensitive Bengali semantic analyzer, BanglaBERT, aided in encoding proper representation from Bengali text into representative that could carry confluences into an imagery pipeline. That is how everything begins, with assigning such encoding in text for furthering a pipeline for semantic, well-structured representation through image generation.

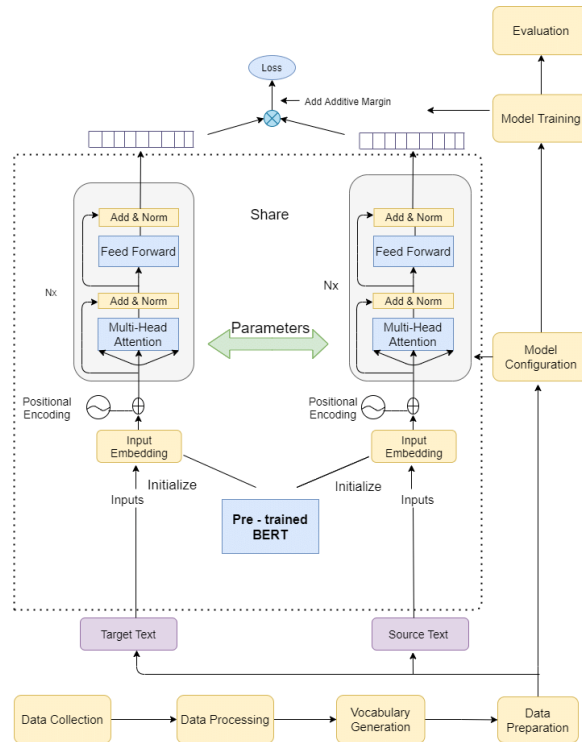


Figure 3.12: Banglabert Architecture
Huang (2025)

6. **Image Encoding with ResNet50:** This image encoding unit will implement a deep CNN structure called ResNet50 that will translate Bengali text descrip-

tions to their corresponding images. ResNet50 is a very well-known model that captures important features of the image and at the same time keeps those of the picture while also maintaining vision. BenglaCLIP is the one in this work that has made ResNet50 the image encoder so it will process and align the generated image with the content and context of the input in Bengali. This process will enable the model to create images that do not only correspond to the semantic meaning of the presented text but also are very attractive and clear.

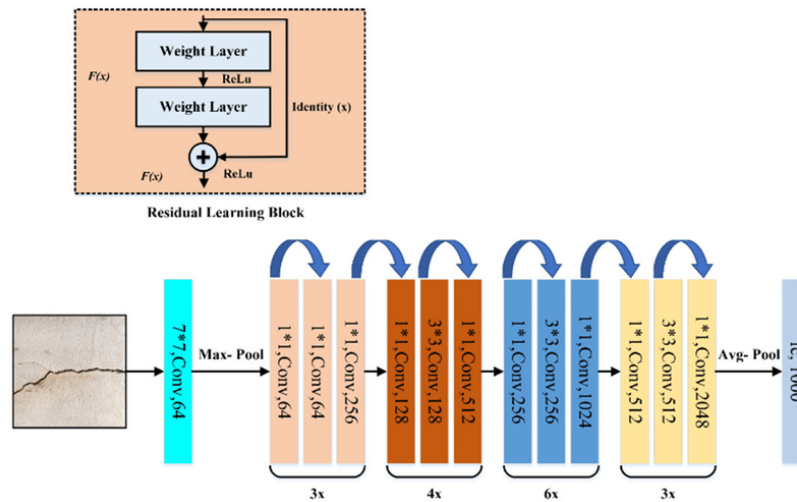


Figure 3.13: Resnet50

L. Ali & Serhani (2021)

7. **Scheduler and Variational Autoencoder (VAE):** The scheduler and Variational Autoencoder (VAE) are some of the important components involved in generating high-quality images from Bengali text prompts using BanglaCLIP in Stable Diffusion. The scheduler controls the step-by-step noise removal process, ensuring a smooth transition from random noise to a meaningful image. Different schedulers, such as DDIM, DDPM, PNDM, and LMS, have different impacts on the speed and quality of image generation, balancing efficiency and accuracy. Meanwhile, VAE compresses it so the model can work in this latent space efficiently and does not need to handle high-dimensional images directly. First, the VAE Encoder compresses images to a low dimension where the diffusion model will act iteratively and refine the details. Once the noise reduction is complete, the VAE Decoder will rebuild a clear and coherent final

high-resolution image from that latent space. The scheduler and VAE together help in optimizing the diffusion process, which makes BanglaCLIP in Stable Diffusion an effective solution for Bangla text-to-image generation without losing either computational efficiency or high-quality image synthesis.

8. **Denoising Diffusion Probabilistic Model (DDPM):** In BanglaCLIP-based Stable Diffusion, the Denoising Diffusion Probabilistic Model (DDPM) is used for generating high-quality images from Bengali text prompts through progressive refinement of noise into meaningful visuals. DDPM is a generative model that learns to reverse a noise-adding process where an image is progressively transformed into pure noise during training. Starting with random noise and conditioned on the BanglaCLIP text encoder, this model iteratively removes noise in small steps to make sure that the synthesized image represents the given Bangla text. The model forms a Markov chain of denoising steps to make the generation stable and realistic. While DDPM can generate high-fidelity images, its major drawback is being computationally expensive due to its many steps of inference. Although the model lacks capturing fine-grained details, which is necessary to preserve semantic consistency in Bangla text-to-image generation, it forms a very integral part of the project.
9. **Few-Shot Learning and Transfer Learning:** However, because the large-scale, annotated Bengali text-to-image datasets are not much available, the model training may have to mix some few-shot learning and transfer learning to overcome a data scarce issue. Few-shot learning, a less supervised learning strategy, is a method for the model to learn atleast one of the many out of only a couple of example images. In the meantime, transfer learning would allow the model to learn from known information derived from bigger, more universal datasets for the tasks related to Bengali picture generating. When we find out to work with it, the fine-tuning process of the pre-trained models using only Bangla language will be general enough to validate either with the smallest dataset.
10. **Model Building with Hugging Face and user Interaction with Gradio** Hugging Face acts as a powerful and dynamic platform, where smooth coordination with the current state of the art is easily enabled for model development

and training. The libraries of Hugging Face, Transformers and Diffusers, help to fine-tune the BanglaBERT and ResNet50 techniques to suit text-to-image generation. Gradio has been employed to provide easy public access to BanglaCLIP by all Bengali speakers. It is a web interface that makes it as simple as possible for users to interact with the model. The model generated an image as output based on Bengali text, and the Gradio interface made it easy for the Bengali-speaking community to find value in areas such as creative industries, education, and communication. With Gradio, users can experiment in real-time with ongoing feedback so as to effect refinements and changes based on user suggestions.

11. **Evaluation and Optimization** Upon completion of the model, extensive evaluation of the quality, accuracy, and cultural relevance of the images generated is carried out. The system is tested with datasets like BanglaLekha, Flickr8, and Bnature, with outputs compared against existing benchmarks related to text-to-image generation. Great care is taken to ensure proper alignment of generated images with Bengali text based on visual fidelity and semantic consistency. Feedback is gathered from users via a Gradio interface to further refine model fine-tuning and generated outputs.
12. **Text Understanding and Embedding Alignment:** During this process, BanglaCLIP returns a feature map while the CLIP model processes the input Bangla text prompt and encodes it into a high-dimensional vector representation in a shared embedding space. Furthermore, this message space is enriched with image embeddings to enable semantic alignment of the visual and textual modalities. In fact, it is basically a contrastive learning-based mechanism where the model learns to maximize the similarity amongst the matched text-image pairs, minimizing it when there is no matching between the pairs. In the Bangla language, for generating good text embeddings, one has to deal with the linguistic nuances of such attributes like .accuracy in text embedding generation for Bangla requires the consideration of following linguistic nuances :

- Bangla Morphology and Complex Syntax

- Word constructions in which are polymorphemic and require smart tokenization policies.
- Context-based variations in word usage and meaning.

We have seen how this directly correlates with the capability of the encoder to successfully capture the nuances of Bangla text and render an output with semantic validity. Morphology and syntax complex, unique to Bangla; polymorphemic word structures that require sophisticated tokenization strategies; context-based variations in word usage and meaning. This already encodes all the important aspects of Bangla text correctly, which determines the coherence and relevance of the generated images.

Multimodal Alignment and Diffusion Guidance The CLIP Bangla text embeddings serve as a guiding signal for the diffusion process in the Stable Diffusion pipeline. Diffusion models generate images by progressively refining random noise into coherent images aligned with the input text.

13. During this process:

- **Embedding Integration:** The embeddings for Bangla text generated by the CLIP encoder are added as conditions to guide the process of diffusion. These embeddings make sure semantic contexts and constraints maintain faithfulness to the textual prompt in the generated images.
- **Semantic Coherence:** The alignment that CLIP ensures in such a way that the visual elements of the generated images will correspond to the specific entities, actions, or settings described in the Bangla prompt. That is essential for keeping the contextual accuracy intact. At the same time, being an intermediary translator between textual and visual modalities, CLIP keeps the diffusion process on track by making the generated output relevant not only visually appealing but also contextually meaningful.

3.5 Evaluation Metrics:

Evaluation metrics are the tools that show how well a machine learning model is performing and operating. They provide information about the quality of a model's

performance, predictive accuracy, and about the observance of specific goals in solving the task. The type of the deep learning challenge may differ-classification, regression, clustering, etc.-which influences the selection of evaluation metrics.

3.5.1 Recall Score:

Recall is also known as Sensitivity or True Positive Rate. Being a performance measure in binary classification tasks, including the sentiment analysis for which it is to be applied to, it's a measurement for generated output diversity because it calculates the level at which the model covers the real image distribution. A high recall score means that the model can generate a wide range of images that reflect the diversity of possible outputs for the given text. It is designed to measure the ability of a model to correctly identify true positives from all actual positive instances in the dataset. Recall is calculated using the formula:

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

TP represents the number of positive cases that the model correctly identified, while FN represents positive cases that were mistakenly labeled as negative. A high Recall value means that the model is efficiently detecting positive cases. This characteristic is especially critical when working with imbalanced data or cases in which missing positive cases are consequential. However, it is important to note that sometimes a high Recall would have a lower Precision-intrinsically, the measure of actual true positives from the total number of positives that have been predicted-and in machine learning models, it usually involves some trade-off in getting an optimal Recall with Precision. This is balanced practically to ensure that in striving for perfect precision, one doesn't end up with very realistic but similar results, and vice versa-that with a high recall, results will indeed be varied but less realistic. Both these metrics together offer a comprehensive view of a model's ability to produce varied and high-quality images in applications like text-to-image. The CLIP Score aims to estimate text-image alignment through the semantic similarity of the generated image and the input text via a pre-trained CLIP model. The better the alignment, the higher the scores. Reconstruction Accuracy metrics include Mean Squared Error-MSE and Peak Signal-to-Noise Ratio-PSNR, which have been used to measure the efficacy in terms of reconstructing the noisy

input. Human evaluations provide a qualitative perspective in terms of perceived realism, relevance, and creativity of generated outputs. Last but not least, metrics with respect to the inference time and model size can be used as a test on the computational efficiency and practicality of the model when it comes to real-world use. These various metrics together represent comprehensive quality, diversity, alignment, and efficiency checks over text-to-image diffusion models.

3.5.2 Frechet Inception Distance (FID) Score:

Performance metrics quantify the model’s performance in terms of producing realistic, diverse, and text-centric images. FID has been one of the most used metrics for quantifying realism, comparing feature distributions of generated images against those of real images; the lower the FID score, the better the quality. The Inception Score is one metric that captures both clarity and diversity of the images; the higher the IS score, the more distinct and relevant the output will be.

$$FID = ||\mu_r - \mu_g||^2 + TR(\sum_r + \sum_g - 2(\sum_r + \sum_g)^{1/2})$$

FID assesses how closely the distribution of generated images aligns with that of real images. It merges both feature quality (mean differences) and diversity (covariance differences).

$$\mu_r, \mu_g :$$

These represent the means of the features for real and generated images.

$$||\mu_r - \mu_g||^2 :$$

This measures the discrepancy between the average features of real and generated images. When means are closer together, it indicates that the generated images better resemble real ones.

$$TR(\sum_r + \sum_g - 2(\sum_r + \sum_g)^{1/2}) :$$

This part evaluates the comparison of diversity (variation) in both real and generated features. Close spreads suggest that the model generates diverse images akin to real data.

So, FID consolidates both realism (the quality of individual images) and diversity (how varied the generated images are). A lower FID score indicates that the generated images are similar to real images in both aspects of quality and variety. FID combines quality and diversity into a single metric.

3.5.3 Fréchet Score for Features (FSF):

Fréchet Score for Features (FSF) and Fréchet Score for Variance (FSV) are metrics used to evaluate the model's capability to produce high-quality and diverse images. FSF gauges the alignment of features from generated images with those from real images by analyzing the similarity of feature distributions. It computes the Fréchet distance between the mean and covariance of features derived from both real and generated images, utilizing a pretrained model like CLIP or Inception. A lower FSF value signifies a better match between the real and generated datasets in terms of overall quality and realism.

$$FID = ||\mu_r - \mu_g||^2 + TR(\sum_r + \sum_g - 2(\sum_r + \sum_g)^{1/2})$$

FSF evaluates the features (such as colors, shapes, and textures) of real images compared to generated ones. It assesses whether the generated images resemble real ones in terms of these features.

$$\mu_r, \mu_g :$$

These represent the means of the features for real and generated images.

$$||\mu_r - \mu_g||^2 :$$

This component quantifies the distance between the averages of the features from real and generated images. The lower this value, the more similar they are.

$$\Sigma_r, \Sigma_g :$$

These denote the covariances (variation or diversity) of the features for both real and generated images.

$$TR(\sum_r + \sum_g - 2(\sum_r + \sum_g)^{1/2}) :$$

: This part assesses the differences in diversity between the two sets of images. FSF represents a combination of the distance between averages and the differences in diversity. A lower FSF indicates that the generated images are more akin to real ones in terms of both quality and variation. FSF evaluates both quality and diversity while focusing on general feature distributions.

FSF represents a combination of the distance between averages and the differences in diversity. A lower FSF indicates that the generated images are more akin to real ones in terms of both quality and variation. FSF evaluates both quality and diversity while focusing on general feature distributions. Conversely, FSV zeroes in on diversity by examining the extent to which the generated output reflects the variance found in the actual image distribution. It emphasizes the model's ability to generate a broad spectrum of unique and varied outputs that align with the input text. Together, FSF and FSV offer a complete evaluation framework that balances the aspects of realism and diversity in images produced by the diffusion model, which is essential for achieving both high-quality and versatile results in text-to-image generation tasks.

Chapter 4

Result Analysis

4.1 BanglaCLIP performance:

The proposed Bangla CLIP encoder was monitored, and the training process regarding evaluation metrics is shown for eight epochs. The results from the graph of training progress unmistakably show model performance improvement in a smooth and rising manner.

4.2 Training and Validation Loss:

The training loss has gone down steadily from about 7.0 to below 1.0 by the last epoch. Such consistent reduction surely indicates that this model has learned some meaningful patterns from the training data. The validation loss rapidly went down during the first few epochs and then mostly stabilized at approximately 3.0. This stabilization indicates that the model has maintained generalization capabilities without overfitting, as supported by the gap between training and validation loss remaining manageable throughout.

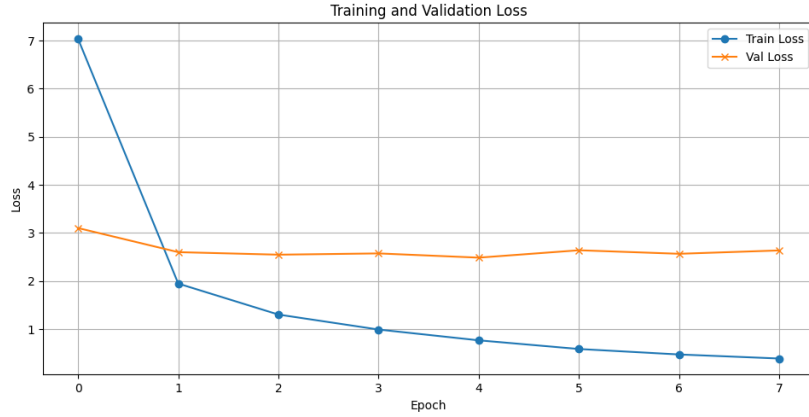


Figure 4.1: Training and Validation loss

4.3 Image-to-Text Recall:

R@1, R@5, R@10 are improving steadily with the growth of all three evaluated thresholds. The best results were at Recall at Rank 10, reaching up to about 0.25, though R@5 and R@1 also grew throughout the advancement of epochs but at a slower pace. With these, the gradual rise in these metrics underlines an increased capability of the model in finding a better alignment to retrieve text-image pairs relevantly. The relatively lower R@1 illustrates the potential room for further optimization, especially in high-precision retrieval tasks.

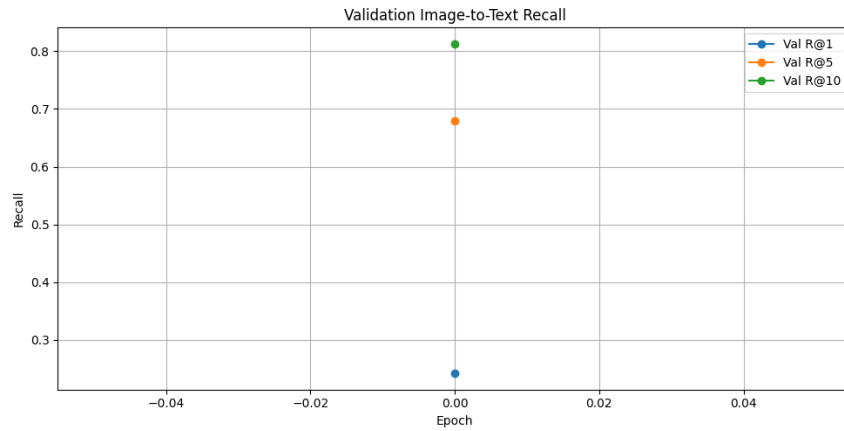


Figure 4.2: Validation Image-to-Text Recall



Figure 4.3: Training Image-to-Text Recall

4.4 Cosine Similarities:

The cosine similarities between text and image embeddings increased linearly for both training and validation datasets. In the last epoch, the validation similarity reached around 0.4, slightly outperforming the training similarity at about (0.37). Such an increase in alignment indicates that the model learns semantically meaningful representations for both modalities. A minimal difference in training and validation similarities underlines that the model's learning is well generalized and robust.

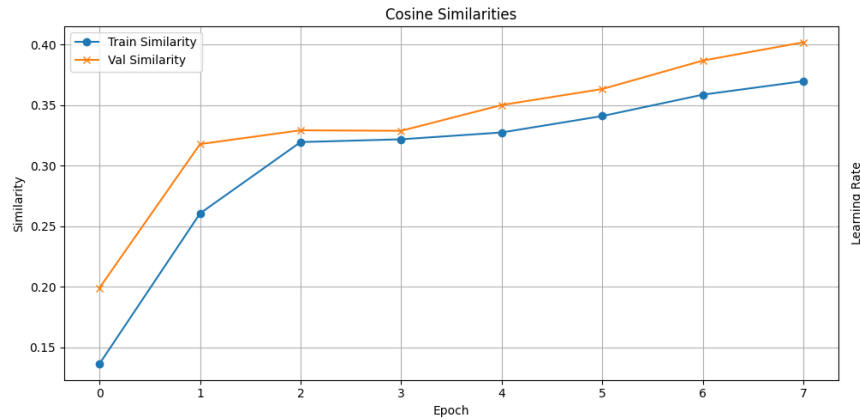


Figure 4.4: Cosine Similarities

4.5 Learning Rate Dynamics:

The learning rate has a warm-up schedule and, therefore, it sees its maximum value at epoch 3 while subsequently decaying. This might provide an optimal setting that allowed for decent escaping of early local minima and model stabilizing later in the process of fine-tuning.

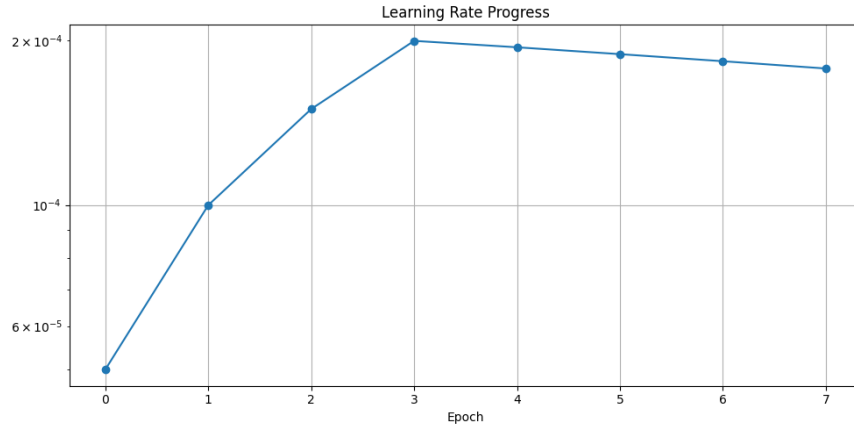


Figure 4.5: Learning Rate Progress

4.6 Similarity Analysis of Query Prompt Comparison:

A comparison of raw and normalized average similarity scores for different queries demonstrates the effectiveness of the model in capturing meaningful relationships between text and image embeddings. In the query similarity graph, the orange line for normalized scores reaches higher values closer to 1.0 for maximum alignment across all query prompts. On the other hand, raw scores (in blue) have lower but rather constant trends across various queries. In fact, this reflects the robustness of the model to maintain similar performances regardless of the complexity of queries.

The comparison of raw vs normalized average similarity scores across various queries also shows how well the model connects Bangla text and images. From the graph, we can see two sets of scores:

- **Raw Scores (Blue Bars):** These are direct model outputs that are not adjusted at all. They are slightly lower than the normalized scores but represent consistent performance across queries. This points to a great consistency in performance

by the model in terms of understanding text-image relationships, be the queries simple or complex.

- **Normalized Scores (Orange Bars):** Orange bars in the graph show the Normalized scores that post-processing techniques increase the raw score to give more life-like semantic closeness. Note that, in most cases, values of the normal scores will be close to 1.0 or higher to give perfect alignment of texts with their image descriptions.

4.6.1 Results

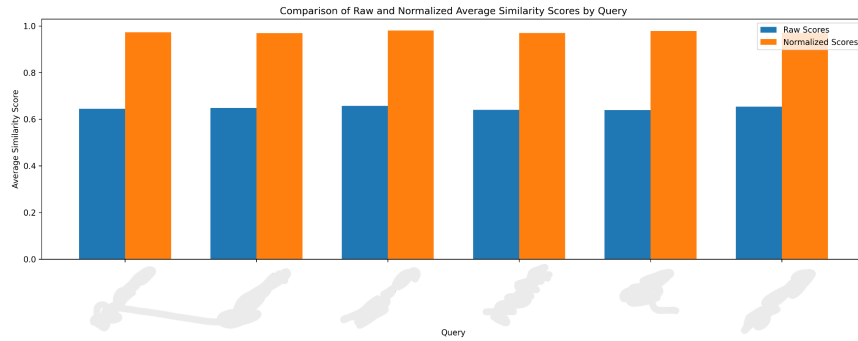


Figure 4.6: Comparison of raw and normalized average similarity scores

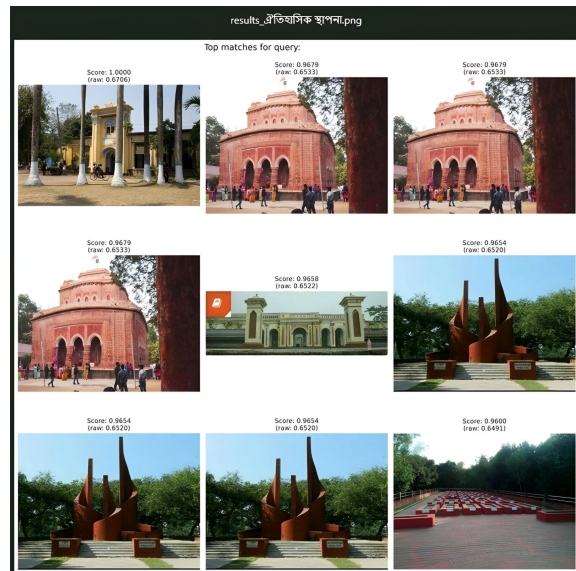


Figure 4.7: Similarity Score of result

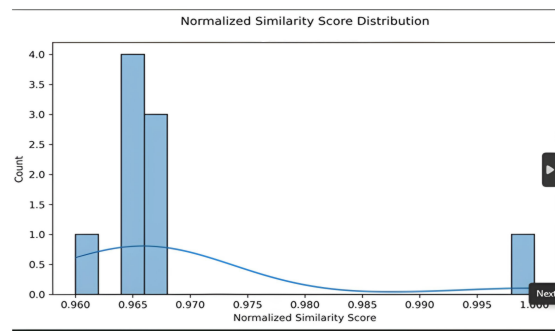


Figure : Similarity Score of ঐতিহাসিক স্থাপনা

Figure 4.8: Similarity Score of result

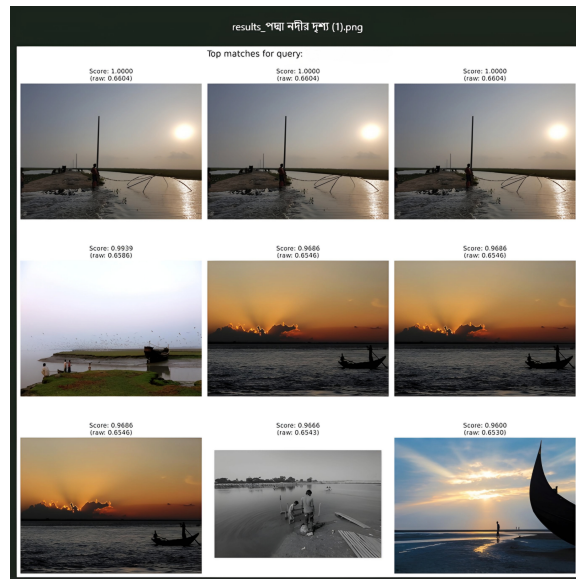


Figure 4.9: Similarity Score of result

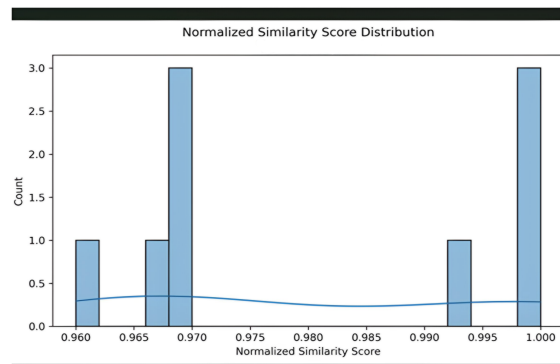


Figure: Similarity Score of পদ্মা নদীর দৃশ্য

Figure 4.10: Similarity Score of result

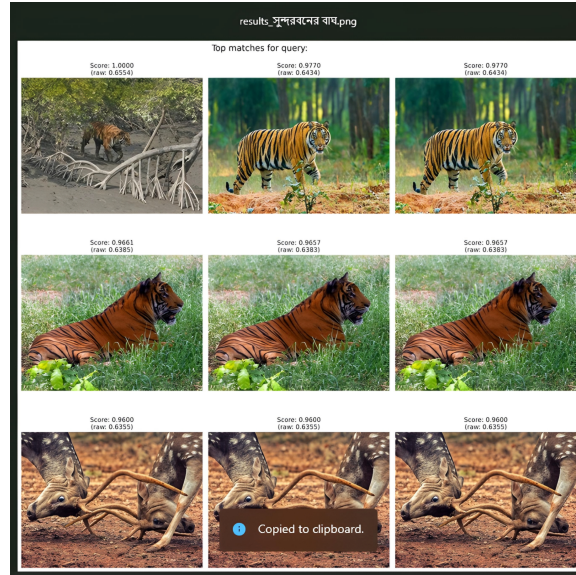


Figure 4.11: Similarity Score of result

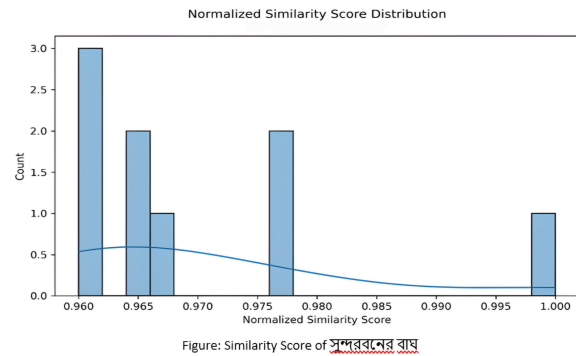


Figure 4.12: Similarity Score of result

4.7 Performance of Bangla text-to-image generation model with Bangla Clip:

The main performance assessment of a Bangla text-to-image generation model is in terms of FID-Fréchet Inception Distance-and Feature Semantic Similarity scores measured with BanglaCLIP. The FID score is a measure of quality and diversity for the generated images. It compares the statistics of the generation distribution with real images. A low FID score means high quality with more diversity, and

vice versa. In the left plot, the FID scores are relatively high in all generations; this means that the generated images may not be realistic and diverse enough compared to the real-world image dataset. This could also mean that the model would further improve with more varied training data, higher image resolution, or overall better quality of generation to make generated images closer to real ones.

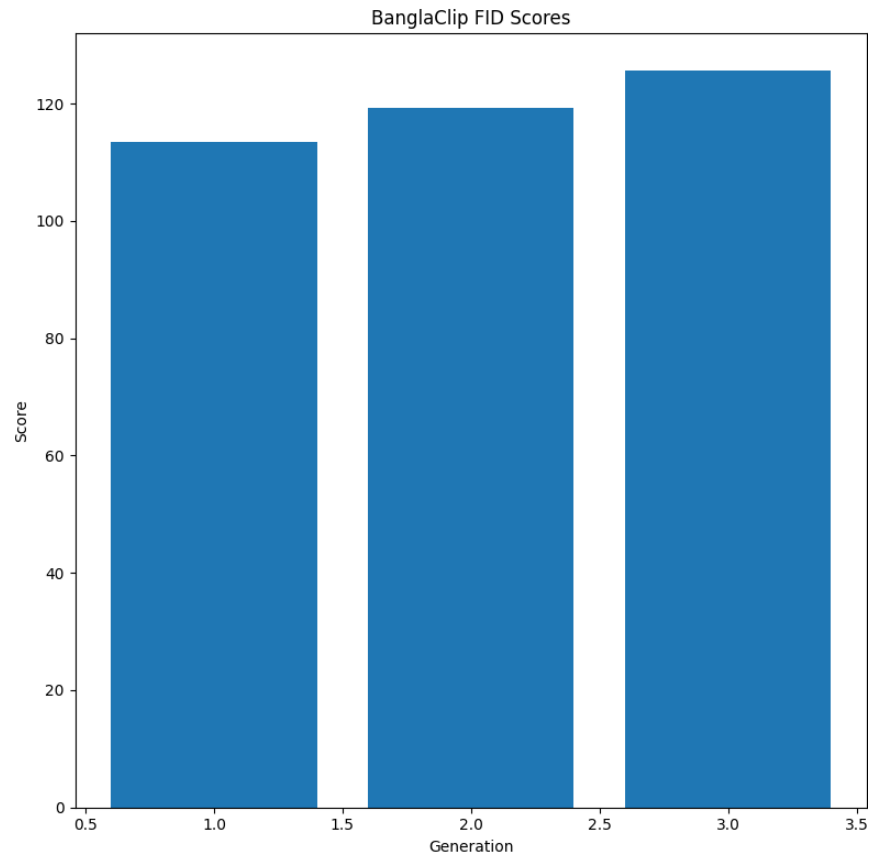


Figure 4.13: BanglaCLIP FID Scores

- Feature Semantic Similarities score: The Feature Semantic Similarities score is plotted on the right, which quantifies how well the generated images semantically align with their respective Bangla text descriptions. The FSS score indicates a perfect alignment of the text prompt with its resulting visual output; here, the scores are steadily upwards of 0.5 for all generations, which suggests that the semantic relationship capture has been considerable. While this indeed shows some level of semantic generation capability in the model,

there is indeed quite a large room for further improvement which will yield higher alignment and better representation of the textual prompts in generated visuals.

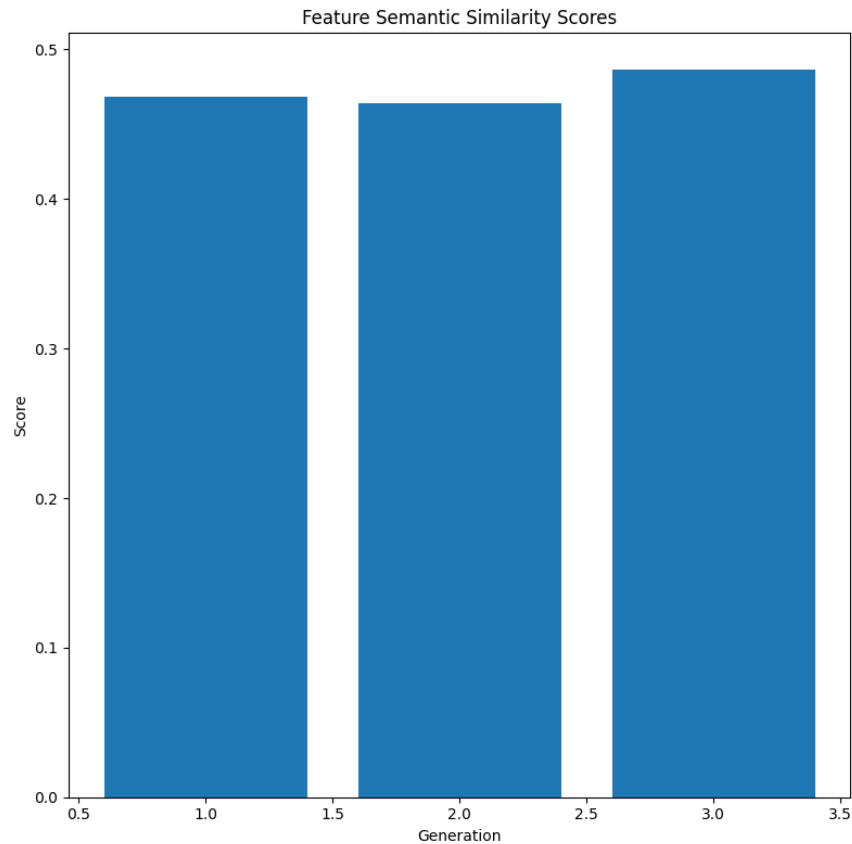


Figure 4.14: Feature Semantic Similarities score

The overall performance metrics are indicative that there is definitely promise in the model, but a lot of improvements are needed. High FID scores mean the generated images have to be drastically better in terms of quality and diversity; moderate FSS scores mean semantic understanding and alignment between Bangla text prompts and generated images need to be improved. Further improvements might be done on the architecture of the model, enhancing training strategies, or using more diverse and good-quality datasets to improve both metrics.

4.8 Performance of Bangla Text to image generation model without Bangla Clip:

Below is a picture that plots the performance for Bangla text-to-image generation without using BanglaCLIP. It follows two metrics: FID (Fréchet Inception Distance) and FSS (Feature Semantic Similarity) scores.

- FID Score: Left Plot First, the FID score, in the absence of BanglaCLIP, is way above 3000, which represents very poor quality and diversity against real images. This contrasts very strongly with results where BanglaCLIP had an FID score of about 120. That would imply that without BanglaCLIP, the model produces unreal outputs.
- FSS Score (Right Plot): Where there is semantic alignment-as extracted by the FSS score of text prompts vs. generated image-their scale is above 1000 when compared with scores with BanglaCLIP at approximately 0.5, pointing toward huge overbias towards generation of outputs on matching general feature criteria without much regards for image quality when no BanglaCLIP.

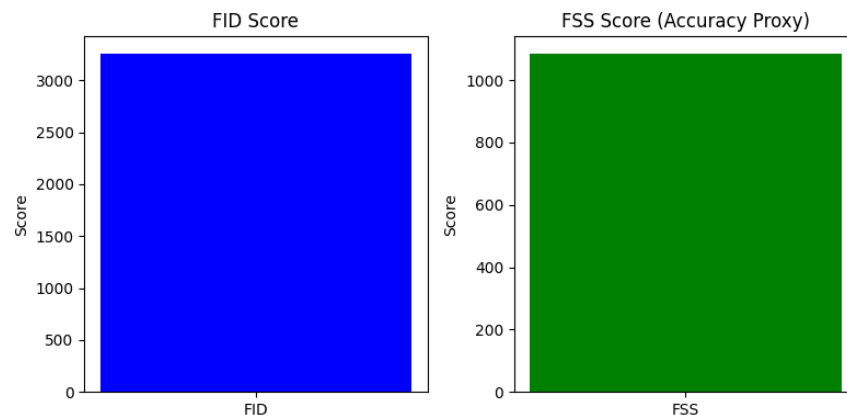


Figure 4.15: FID and FIS scores

4.9 Comparison of Results:

- FID Score With BanglaCLIP: FID scores were way lower, around 120, indicating that BanglaCLIP has helped the model generate images closer to real ones

in quality and diversity.

- **FID Score Without BanglaCLIP:** The FID score is alarmingly high, with a value greater than 3000, showing a significant drop in the quality and diversity of generated images.
- **FSS Score:** The resulting models have modest FSS scores (0.5) and are thus likely to be reasonably alignment semantic to the text prompts. Without BanglaCLIP, the scores are very high (> 1000) and indicative of the overfitting and alignment biases in those models where realism and quality are sacrificed at the cost of basic semantic matching. BanglaCLIP improves the model's capability for high-quality image generation, while it maintains diversity and strikes a reasonable balance with semantic alignment, as reflected in the case of lower FID scores and moderate FSS scores. Without BanglaCLIP, the model outputs are poor in quality and diversity with high FID scores, though semantic alignment seems inflated with high FSS scores. This underlines the importance of BanglaCLIP in balancing realism, diversity, and semantic relevance in Bangla text-to-image generation tasks.

4.9.1 Performance of Bangla text-to-image generation model with Bangla Clip and integration with a Translation Algorithm:

For better performance, we have integrated a translation algorithm into the pipeline alongside BanglaCLIP, improving the overall quality and semantic alignment of generated images. The FID (Fréchet Inception Distance) score, shown in the left plot, remains relatively high (27-28), indicating that the generated images still lack realism and diversity compared to real-world images. On the other hand, the FSS (Feature Semantic Similarity) score, shown in the right plot, is above 0.5, suggesting a moderate level of semantic alignment between text prompts and generated visuals. The translation algorithm plays a crucial role by refining Bangla text inputs, ensuring clearer encoding and reducing noise before passing them to BanglaCLIP. This enhances the model's ability to understand and generate images that better match the given prompts. While these enhancements improve semantic consistency, the relatively high FID scores suggest that further refinements are necessary, such as higher-resolution training data, optimized diffusion models, and more diverse

datasets. These improvements will help achieve a better balance between generating high-quality images and maintaining strong semantic alignment with Bangla text descriptions.

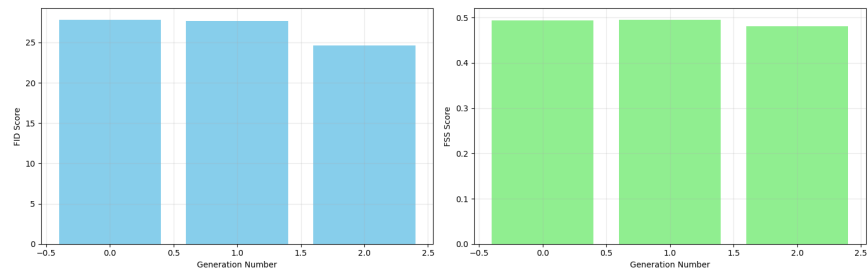


Figure 4.16: Translation Algorithm

বাংলা টেক্সট থেকে ছবি তৈরি

যেকোনো বাংলা টেক্সট দিয়ে উচ্চমানের ছবি তৈরি করুন

বাংলা টেক্সট লিখুন

জলের উপর একটি জাহাজ

ছবির সংখ্যা: 1

স্টেপস: 50

গাইডেন্স স্কেল: 7.5

সীড (ঐচ্ছিক): 0

Clear Submit

১) তৈরি করা ছবি



বাবাছত প্রদর্শন

Activate Windows
Go to Settings to activate Windows.

Figure 4.17: Bangla Text to image generation result



Figure 4.18: Generated image

বাংলা টেক্সট থেকে ছবি তৈরি

যেকোনো বাংলা টেক্সট দিয়ে উচ্চমানের ছবি তৈরি করুন

বাংলা টেক্সট লিখুন

এক কাপ চা টেবিলে রাখা আছে গুটি বিস্কুট

ছবির সংখ্যা 1 ৗ 4

স্টেপস 50 ৗ 100

পাইডেস ফোল 7.5 ৗ 20

সীড (ঐচ্ছিক) 0

ৗ তৈরি করা ছবি



Clear
Submit

ব্যবহৃত প্রম্পট

There's a cup of tea on the table 4 biscuits, professional photography, high resolution, 4k, detailed, realistic, beautiful composition, sharp focus, HDR, dramatic lighting, cinematic view

Figure 4.19: Bangla Text to image generation result

Chapter 5

Project Consideration

5.1 Project Overview

BanglaCLIP-Based Stable Diffusion has been a new perspective in Bangla text-to-image generation. The model aims at bridging the representation gap between the Bengali language and visual understanding by using contrastive learning with diffusion models. Whereas there has been rapid growth recently in the arena of vision-language models, most of the existing approaches have been tailored for high-resource languages like English, leaving widely spoken yet underrepresented languages like Bengali limited in AI-driven multimodal applications. This work introduces BanglaCLIP, a Bengali-specific encoder; a variant of the Stable Diffusion framework that effectively aligns textual and visual representations using BanglaBERT as a text encoder and ResNet50 as an image encoder. The BanglaCLIP encoder allows the model to map Bengali text descriptions to their relevant images by using contrastive learning such that the semantic meanings are preserved. Coupled with Few-Shot Learning, this work enables effective image generation even with limited training data and hence is suitable for low-resource environments. Image synthesis is done by a state-of-the-art generative model, Stable Diffusion, which progressively refines noisy images into quality images using Denoising Diffusion Probabilistic Models (DDPM). It compresses and reconstructs images by Variational Autoencoder (VAE) and a scheduler that manages diffusion for efficiency and coherence in the generated outputs. The training and testing dataset is a collection from BanglaLekha, Flickr8, and Bnature, representing a rich variety of real-world Bengali contexts in terms of image-text pairs. The pre-processing for data involves text normalization, tokenization using BanglaBERT, and feature

extraction using ResNet50. Data augmentation methods such as random cropping, rotation, and contrast adjustment were employed to facilitate generalization. Training consists of a structured pipeline where BanglaCLIP aligns the textual and visual features, while Stable Diffusion generates images that are semantically and contextually accurate. The testing phase considers the model performance with qualitative and quantitative measures. CLIP Score, Fréchet Inception Distance, and Structural Similarity Index ensure the accuracy and quality of the generated images. Ablation studies and user evaluations give further insight into the robustness of the models. Also, edge case testing refines the system's ability in handling rare words, complex phrases, and abstract Bengali descriptions. This application is developed in Hugging Face and Gradio and provides a more interactive interface in generating images from Bengali text in real time. The developed model would then be used to test repeatedly, tune further, and obtain user feedback, making it easier for practical use. By integrating BanglaCLIP into Stable Diffusion, this study will contribute toward research in Bengali AI, creative content generation, and digital applications and make the vision-language AI tools more inclusive and effective for Bengali speakers.

5.2 Project scope

Key components such as the generation of images from text, especially for Bengali language and its visual representation, constitute the major scopes of BanglaCLIP-based stable diffusion. This will be a very effective mover toward the establishment of a system that bridges the gap between Bengali text and corresponding visual content, thus allowing, in turn, Bengali-speaking people to interact with and reap the benefits of AI-powered image generation tools.

5.2.1 Bengali Text to Image Generation:

This work centres around the photo-realistic image being produced out of Bengali text descriptions. This opens up classic AI image generation techniques to Bengali-speaking users, who are going through a comparative shortage of usable software in their languages, especially concerning all activities that involve multimodality. These will open casinos of creative, educational, and business possibilities before

members of that community, thus allowing AI leverage in multiple applications for content creation.

5.2.2 Model Development and Integration:

This work is focused on building a BanglaCLIP model that integrates BanglaBERT as a text encoder and ResNet50 as an image encoder for text-image alignment. Further, this model will be integrated into the Stable Diffusion framework using Denoising Diffusion Probabilistic Models (DDPM) and Variational Autoencoders (VAE) that can synthesize high-quality images based on Bengali text inputs. The project will also explore the use of **Few-Shot Learning** to produce balanced images with insufficient data so as to make it advantageous in resource-constrained environments.

5.2.3 Dataset and Preprocessing:

The dataset will be used for training and testing, comprising BanglaLekha, Flickr8, and Bnature representing wide coverage of image-text pairs in real-world Bengali scenarios. This involves the pre-processing of data, such as normalization, tokenization, augmentation, and so on, to feed quality input to train the model.

5.2.4 Model Evaluation:

This scope includes qualitative and quantitative reviews of model performance, including quantitative parameters such as CLIP Score, Fréchet Inception Distance (FID), and Structural Similarity Index (SSIM) that measure the quality of generated images. Furthermore, user evaluation and edge case testing are conducted to provide confidence in the system with respect to being highly successful in providing useful responses when information related to ambiguous or complex Bengali text descriptions is involved.

5.2.5 User Interface Development:

The system is going to leverage Hugging Face and Gradio for real-time image generation from Bengali text as an application. Also, the scope shall ensure that users have input facilities to provide their Bengali descriptors in easy terms for

getting corresponding images of their choices, keeping a close consideration for user interaction and ease of use.

5.2.6 Practical Applications:

The ultimate goal of the project is to provide tools that will find application in a real-world scenario. This includes creative industries, education, marketing, and content creation for Bengali-speaking people. The model therefore aims to cater to an intermediary for people, firms, and organizations that want to benefit by means of AI-generated visual data based on the Bengali text.

5.3 Timeline

Our project on Bengali Text-to-Image needed to follow a timeline in order for it to be completed. By mid-March, the topic had been chosen; hence, that marked the beginning of our project. By the end of the month, the plan got finalized, with field research beginning almost immediately. Late April to early August: dataset collection and pre-processing in preparation for model training. Following the model selection and training for several days in August, advanced learning techniques were applied at the end of September and early October to enhance the performance. The backend coding was performed from the end of October into the first part of November, while design and development of the user interface were completed by mid-November. The system was tested and validated in the middle of December to ensure it was reliable and precise. Full documentation of the project report was produced during the weeks from late December to early January. In the early days of January, the project was finally thoroughly reviewed to guarantee an accurate high-quality outcome. This schedule ensured that every stage of the project was given the required time and attention in order for it to be carried out effectively.

5.4 Development Tools

5.4.1 Hardware Requirements:

The system demands the following hardware components to operate optimally:

- Processor: Dual-core or higher
- RAM: 8 GB or more
- Storage: Minimum 20 GB of free disk space
- Internet connectivity: High-speed internet for web application access
- Display: Monitor or screen for accessing the user interface
- Hardware accelerator : T4 GPU , A100 GPU , L4 GPU , v2-8 TPU , v5e1 TPU

5.5 Software Requirements

The software requirements necessary to develop and operate the system include:

1. Programming Language: Python 3.x for application logic and deep learning implementation.
2. Framework: PyTorch or TensorFlow for model training and development.
3. Libraries: text preprocessing and for image generation.
4. Dataset: Bengali text datasets paired with visual datasets for model training.
5. Application Framework: for creating a user-friendly web interface.
6. Version Control: GitHub for collaboration and version management.

5.6 Project Budget

Item	Cost
GPU Cost	TK 15000

Table 5.1: Project Budget

Chapter 6

Conclusion

The BanglaCLIP-Based Stable Diffusion project represents a leap forward in the domain of text-to-image generation for low-resource languages, particularly Bengali. This project bridges the gap between Bengali textual descriptions and their corresponding visual content by embedding BanglaCLIP into the Stable Diffusion framework-a new tool for Bengali speakers to generate images from their native language. The current work uses powerful models for text encoding, such as BanglaBERT, for image feature extraction-ResNet50, and for high-quality image synthesis-Denoising Diffusion Probabilistic Models (DDPM). Therefore, it may be useful for several applications. This model has the potential to perform tasks with very few training data with Few-Shot Learning and thus opens a lot of opportunities in AI applications within Bengali-speaking communities where multimodal AI research is significantly lacking. The project will democratize access to quality, AI-powered image generation and thus enable creators, businesses, educators, and developers to achieve their desired output with much less friction, especially in resource-scarce environments. Besides, development with Hugging Face and Gradio has made interaction with the system highly user-friendly, while the thorough testing and evaluation process makes sure that the model provides accurate, reliable, and meaningful results. The contributions of this go to the advancement of not only Bengali text-to-image generation but also further improvements and adaptations of similar technologies for other low-resource languages. In the process, BanglaCLIP-based Stable Diffusion will mark a transformative leap toward creating inclusive AI tools that address Bengali-speaking needs, covering a wide array of areas including content creation, education, marketing, and digital media.

6.1 Impact on Society

The BanglaCLIP-Based Stable Diffusion project will be of utmost consequence in increasing access to AI-driven tools that support text-to-image generation among the Bengali-speaking communities. In addition, it closes the gap between the Bengali language and visual content whereby users, companies, and educational institutions can easily produce contextually accurate and high-quality images from the given text description in Bengali. This development not only enhances creativity in the realm of digital content creation but also assists small businesses and entrepreneurs in effectively developing cost-efficient marketing and branding materials. It allows in education for an interactive, personalized manner of learning, especially for students located within Bengali-speaking countries. The model even goes on to support digital inclusivity by allowing Bengali speakers to meaningfully engage with these emerging areas of AI and machine learning. The project also plays an instrumental role in the preservation and promotion of Bengali heritage through cultural expression using visual media. On a larger scale, this work contributes to AI research by improving technologies created for low-resource languages, thus opening up avenues for similar development for linguistic communities in need. The current project thus represents a substantial step toward developing more equitable, inclusive, and culturally relevant AI applications that will have an impact on various industries, ranging from education and business to digital arts and social media.

6.2 Impact on Environment

The BanglaCLIP-Based Stable Diffusion project, like any other AI-driven model, impacts the environment directly and indirectly. Deep learning models, especially diffusion models, involve intensive computational resource requirements during training and deployment and hence result in energy consumption and carbon emission. Large-scale model training, like Stable Diffusion with BanglaCLIP, requires running high-performance GPUs and TPUs for extended periods of time, adding to higher electricity usages and environmental footprint. However, this project follows a few-shot learning approach that reduces the need for large datasets and

computational power, hence minimizing energy consumption compared to conventional deep learning methods. This project also contributes to sustainability in the creation of digital content by letting Bengali speakers create images from text directly without having to go through traditional graphic design tools or physical resources, hence manually creating and printing images and generating lots of waste. Similarly, enhanced development on platforms such as Hugging Face and Gradio ensures deployment with efficiency, hence minimizing redundant computational overhead. This can be further mitigated in the long run, by integrating efficient AI techniques with sustainable computing practices, such as low-power AI hardware and cloud solutions running on renewable energy. For evolving AI models, green AI practices, such as optimizing the model size and inference efficiency, make the AI-driven innovation-like BanglaCLIP-Based Stable Diffusion more ecologically responsible while retaining its technological and societal benefits.

6.3 Ethical aspects

The project BanglaCLIP-Based Stable Diffusion opens up some very important discussions regarding ethical concerns in the field, especially in areas such as bias, misinformation, data privacy, and the responsible use of AI. First among these ethical concerns is that this model, which is generated with biased data or limited data, can end up producing erroneous and culturally insensitive images, thereby miseducating people on the Bengali culture. Such biases can only be reduced if fairness and inclusivity are ensured in the selection of the dataset and fine-tuning of the model. Other ethical issues relate to the misuse of AI-generated images, in that the technology can be used to create deepfakes, misinformation, or manipulated content-with implications for digital ethics and media authenticity. To avoid misuse, it is important to give security by watermarking AI-generated content and to establish rules of use. Moreover, there are data privacy and security to consider, ensuring that the collection or processing of text data with AI applications meets the ethical standards of AI and data protection. Further, accessibility of AI models has to be weighed against ethical considerations regarding automation replacing human creativity and must be positioned to empower and not replace human artistic expression. Following ethical AI principles, BanglaCLIP-Based Stable Diffusion can

be developed and deployed in a manner that maximizes benefits to society while minimizing potential risks and adverse consequences.

6.4 Sustainability Plan

The BanglaCLIP-Based Stable Diffusion project is designed to incorporate a long-term sustainability plan that will ensure its continuing development, accessibility, and ethical use. The plan for efficient use of resources, optimization of the model, and deployment of responsible AI with community involvement will help maintain its relevance and impact over time. Few-shot learning is applied in the project to address computational sustainability, reducing the need for large datasets and computational power by a great amount, hence reducing energy consumption. Besides, further environmental impact reduction can be achieved by using optimized training pipelines, cloud-based solutions, and energy-efficient hardware. The model will be more sustainable by adopting green AI practices such as improving the efficiency of inferences and using renewable energy-powered data centers. It will further be updated with newer datasets that contribute much to model performance in terms of generating content in a nondiscriminatory way for usability. Model fine-tuning and contributions from the community will keep changing as the needs of Bengali-speaking users are dynamic. Collaboration with open-source AI platforms such as Hugging Face and Gradio will be a great way for this model to reach out to developers, researchers, and end-users and develop innovation and knowledge sharing in the AI community. From the aspect of social sustainability, it is a jump that the project makes in delivering equitable access to AI for Bengali-speaking communities while empowering content developers, educators, and businesses alike with AI-enabled tools. Building ethical guidelines about responsible AI use will be applied by employing methods of reducing biases and techniques for ensuring authenticity so that the model functions in a non-opaque responsible manner. These would be sought with research institutions, AI organizations, and government bodies for long-term sustainability of effort and finance. The partnerships also hold the possibilities of financial endowments, infrastructural support, and even technical development. Besides this, selling API-based services to companies and researchers will create a self-sustaining ecosystem where the model is kept maintained and continuously improved without dependence on

exclusive foreign funding. BanglaCLIP-Based Stable Diffusion, by incorporating environmental, social, and economic sustainability strategies, would create a lasting impact on the accessibility of AI, representation of languages, and content creation with responsible use of AI.

6.5 Brief Summary

BanglaCLIP-Based Stable Diffusion focuses on enabling Bangla text-to-image generation using a few-shot diffusion model. Most of the traditional vision-language models, such as CLIP, are focused on high-resource languages like English and have left a big gap in low-resource languages like Bangla. To bridge the gap between Bangla text and image representations, a custom encoder has been developed, which is called BanglaCLIP. It employs BanglaBERT as the text encoder and ResNet50 as the image encoder for more precise and contextual image generation. The model is developed using the Stable Diffusion framework, a powerful generative AI model, and is integrated with Hugging Face and Gradio for development and accessibility. The project employs datasets such as BanglaLekha, Flickr8, and Bnature for diverse and high-quality training data. In this work, few-shot learning is followed to make the model optimized on limited data, thus making the model computationally efficient with high-quality generation. This architecture consists of a UNet-based denoising model, Variational Autoencoders, and Schedulers in order to refine the image generation process. The approach based on Denoising Diffusion Probabilistic Models has been followed towards the progressive generation of high-resolution images which are coherent with the Bangla text description. This ensures that BanglaCLIP effectively captures the linguistic and contextual intricacies of the Bangla language while aligning them with visual representations. Training and testing are done to reinforce the model's performance in terms of accuracy, generalization, and efficiency for real-world applications. The project bears great socio-economic significance for content creators, businesses, and educators alike by enabling AI-powered image creation in Bengali. The project focuses on responsible AI development, bias mitigation, and energy-efficient computing from the ethical and environmental point of view. A sustainability plan is proposed to ensure long-term model optimization, dataset updates, and ethical AI governance.

This work presents a significant leap toward making AI more accessible to Bengali-speaking users and will open up avenues for more inclusive and culturally relevant AI applications.

6.6 Future Scope:

Though the BanglaCLIP-based Stable Diffusion has initiated a breakthrough in Bangla text-to-image generation, there are a lot of potential areas that remain open for future developments and enhancements. Expanding the training dataset with more and varied Bangla text-image pairs would be one very important area, so that the model generalizes better and gives improved performance. The next versions can further explore multimodal learning through the use of audio, video, and other linguistic features to enhance semantic understanding of Bangla text. Another very important line of future work will be the optimization of the model to make it more efficient and accessible. Lowering computational costs by using lightweight architectures and knowledge distillation techniques will make the model more viable in low-resource environments. Real-time inference techniques will also be explored to enable a more interactive AI system. From a technical perspective, further refinement of the model's output will come from enhancing the BanglaCLIP encoder to handle complex sentence structures, regional dialects, and domain-specific terminology. Increasing the resolution and fidelity of the diffusion model's images through the use of advanced denoising and super-resolution techniques will contribute to more realistic and detailed image generation. This model can be deployed as a public API or open-source tool on platforms such as Hugging Face and Gradio to enable community-driven improvements and widespread adoption. Integration of user feedback loops will enable continuous learning and fine-tuning based on real-world use cases. Finally, ethical concerns in mitigation of bias, responsible usage of AI, and verification of authenticity of content shall be a very important area of research in times to come. This will make sure that the model is inclusive, unbiased, and ethically responsible by incorporating fairness-aware training techniques and AI transparency tools. This would be evolving the project BanglaCLIP-Based Stable Diffusion into more power and efficiency, gaining higher levels of adoptions to AI, immensely developing the ability for AI in reaching Bengali-speaking communities

worldwide and helping further to conduct low-resource language research all over the globe.

References

- A. Abid, A. A. D. K. A. A., A. Abdalla, & Zou, J. (2019). Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv.org*.
- A. M. Faruk, M. M. A. M. R. F., H. A. Faraby, & Morol, M. K. (2020). Image to bengali caption generation using deep cnn and bidirectional gated recurrent unit. *International Conference on Computer and Information Technology (ICCIT)*.
- Asgarov, . R. S., A. (2024). Lowclip: Adapting the clip model architecture for low-resource languages in multimodal image retrieval task. *arXiv:2408.13909*.
- Bent, B. (2024). Semantic approach to quantifying the consistency of diffusion model image generation. *arXiv.org*.
- Brown, M. B. R. N. S. M. K. J. D. P. N. A. S. P. S. G. A. A. A. S. H.-V. A. K. G. H. T. C. R. R. A. Z. D. M. W. J. W. C. . . . A. D., T. B. (2020). Language models are few-shot learners. *arXiv.org*.
- Cao, Z. Z. W. H. . W. K., J. (2024). Vleu: a method for automatic evaluation for generalizability of text-to-image models. *2021 Conference on Empirical Methods in Natural Language Processing*.
- Chen, H. L. C. Y. D. W. S. L. J. X. . . W. W., G. (2023). Mclip: Multilingual clip via cross-lingual transfer. *61st Annual Meeting of the Association for Computational Linguistics*.
- Cui, B. J. W. G. C. Q. X. Z. L. W. Z. K. . S. Y., T. (2025). Evaluating image caption via cycle-consistent text-to-image generation. *Computer Vision and Pattern Recognition*.
- Dalva, Y., & Yanardag, P. (2022). Noiseclr: A contrastive learning approach for unsupervised discovery of interpretable directions in diffusion models. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- F. Sung, L. Z. T. X. P. H. S. T., Y. Yang, & Hospedales, T. M. (2017). Learning to compare: Relation network for few-shot learning. *Computer Vision and Pattern Recognition*.

- Hu, . K. F., H. (2023). Meta-learning for vision-and-language cross-lingual transfer. *Proceedings of the the 3rd Workshop on Multi-lingual Representation Learning (MRL)*.
- Huang, Z. H. S. Y. R. S. . K. E., J. (2025). Image2text2image: A novel framework for label-free evaluation of image-to-text generation with text-to-image diffusion models. *Computer Vision and Pattern Recognition*.
- H. Ye, M. T. R. S., X. Yang, & Ji, S. (2021). Improving text-to-image synthesis using contrastive learning. *Cornell University: arxiv.org, <https://arxiv.org/abs/2107.02423>*.
- I. Alabdulmohsin, A. S. P. G. A. D., X. Wang, & Zhai, X. (2024). Clip the bias: How useful is balancing data in multimodal learning? *arXiv.org*.
- J. Ho, A. J., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *arXiv.org*.
- Khan, M. F., Shifath, S. S.-U.-R., Islam, e. . C. N., Md Saiful", Béchet, F., Blache, P., Choukri, K., ... Piperidis, S. (2022). Ban-cap: A multi-purpose english-bangla image descriptions dataset. *Thirteenth Language Resources and Evaluation Conference*.
- Kim, S. H. . Z. B., T. (2022). Cross-modal alignment learning of vision-language conceptual systems. *Computer Vision and Pattern Recognition*.
- Kowsher, S. A. A. P. N. J. A. M. S. D. P. K. . K. T., M. (2022). Bangla-bert: Transformer-based efficient model for transfer learning and language understanding. *IEEE Access*.
- L. Ali, H. A. J. M. G. W. K., F. Alnajjar, & Serhani, M. A. (2021). Performance evaluation of deep cnn-based crack detection and localization techniques for concrete structures. *MDPI*.
- Li, L. D. S. S. . H. S., J. (2024). Bootstrapping language-image pre-training with frozen image encoders and large language models. *Computer Vision and Pattern Recognition*.
- Mansoor, A. H. M. N. M. S. R. M. M., Nafees; Kamal. (2019). Banglalekhaimagecaptions. *Mendeley Data, V2, doi: 10.17632/rxxch9vw59.2*.
- Mu, L. H. W. J. Z. X. W. C. L. Y. H. Q. X. T. C. G. . Z. J., Y. (2025). Boosting text-to-image generation via multilingual prompting in large multimodal models. *Computation and Language*.
- O. Ronneberger, P. F., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *U-NET: Convolutional Networks for Biomedical Image Segmentation*.

- P. Grimal, O. F., H. L. Borgne, & Tourille, J. (2024). Tiam - a metric for evaluating alignment in text-to-image generation. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.
- Radford, K. J. W. H. C. R. A. G. G. A. S. S. G. A. A. M. P. C. J. K. G. S.-I., A. (2021). Learning transferable visual models from natural language supervision. *Computer Vision and Pattern Recognition*, 48, 1–10.
- Tabassum, A. M. S. S. S. M., M. (2025). *Banglaclip*. <https://huggingface.co/datasets/Mansuba/Banglafinal>. HuggingFace.
- Wolf, D. L. S. V. C. J. D. C. M. A. C. P. R. T. L. R. F. M. . B. J., T. (2019). Huggingface's transformers: State-of-the-art natural language processing. *arXiv.org*.
- Wu, R. R. G. U. A. B. A. N. M. K. C. G. B. F. A. H. J. B. C. G. M. G.-A. O. M. M. A. C. S. B. D. C. G. L. B. L. H. S. . . . H. K., C. (2021). Sustainable ai: environmental implications, challenges and opportunities. *arXiv (Cornell University)*.
- Zhou, L. C. C. C. G. J. . X. J., Y. (2022). Lafite2: Few-shot text-to-image generation. *Computer Vision and Pattern Recognition*.