Generative Project - A Deep Learning Based Text to Image Generation

Milestone 1
Group 11

Lochan Enugula
Yaswanth Kumar Reddy Gujjula
Rishabh Joshi
Rutuja Jadhav

enugula.l@northeastern.edu
joshi.risha@northeastern.edu
jadhav.rutu@northeastern.edu
gujjula.y@northeastern.edu

Git Repo: https://github.com/rutuja-jadhav29/NNDL-GenerativeProject.git

Submission Date: 2nd September 2025

Overview

This milestone develops a text-to-image generation and validation pipeline using the Flickr30k dataset.
It combines Stable Diffusion v1-5... to examine how natural-language prompts can be transformed into realistic, semantically aligned images.
Objectives:

1. Curate and preprocess ~3 000 image–caption pairs from Flickr30k.
2. Validate text–image alignment using CLIP embeddings.
3. Generate novel, prompt-conditioned images through diffusion modeling.

Dataset & Preprocessing

The Flickr30k dataset (31 000 images, five captions each) was down-sampled to ≈ 3 000 records; 2 879 clean pairs remained after removing 120 corrupted files.
Captions were normalized (lowercased, trimmed, deduplicated), and images were resized to 224 × 224 px for model compatibility.
File integrity was verified through existence and consistency checks.
Outputs: cleaned_captions.csv and cleaned_images/ (≈ 2 879 files)
This step ensured uniform input quality and reduced processing cost by ≈ 85 %.

Model Implementation

1 · **CLIP Validation – Text** → Embedding Alignment

- Model: ViT-B/32 (pretrained on LAION-400M).
- Task: Compute cosine similarity between text and image embeddings.
- Result: Mean similarity ≈ 0.31 (range 0.25–0.38), indicating moderate semantic alignment.

2 · **Stable Diffusion Generation** – Prompt → Image

- Model: runwayml/stable-diffusion-v1-5 via Hugging Face Diffusers.
- Hardware: GPU runtime with attention slicing enabled for memory efficiency.
- Prompts - A man on a red bike," "A woman in a red dress," "A boy standing at a van door,"etc.
- Observation: All outputs accurately depicted prompt semantics and context.

Conclusion

This milestone demonstrates a fully reproducible generative deep-learning pipeline integrating data preprocessing, embedding validation, and text-to-image synthesis.
The system achieved consistent semantic alignment and prompt accuracy, establishing a foundation for Milestone 2, which will focus on fine-tuning Stable Diffusion on custom captions, expanding prompt diversity, and introducing quantitative generative evaluation metrics.

Outputs

a little boy standing in the doorway of a van

a man on a red bike popping a wheelie

a woman in a purple shirt playing an instrument

two people riding a log flume in a theme park

a woman in a red dress looking at her cellphone