Generative Project Milestone 3 – Evaluation Report

Team Members (Group 11)
Lochan Enugula
Rishabh Joshi
Rutuja Jadhav
Yaswanth Kumar Reddy Gujjula

Github Link: https://github.com/rutuja-jadhav29/NNDL-GenerativeProject

1. Objective

The goal of Milestone 3 was to evaluate the quality of the generated images produced in Milestones 1 and 2 using:

- Quantitative metrics: Fréchet Inception Distance (FID) and Inception Score (IS)
  Qualitative evaluation: visual comparison between M1 baseline outputs and M2 Stable Diffusion outputs
- Sensitivity analysis:
  - Guidance scale sensitivity
  - Embedding sensitivity using CLIP models (ViT-B/32 vs ViT-B/16)

The objective was to build a complete evaluation pipeline to compare baseline and diffusion-based generative models and understand how model parameters influence image quality, alignment, and realism.

2. Experimental Setup

2.1 Datasets

We used three sets of images:

- Real Dataset: 500 cleaned Flickr30k images used as the reference distribution for FID.
- M1 Baseline Outputs: 5 images generated from our basic text → embedding → DDPM pipeline.
- M2 Stable Diffusion Outputs: Images generated from Milestone 2 using guidance scales of 5.0, 7.5, and 10.0.

2.2 Metrics Computed

We implemented and computed the following:

- **FID:** Measures distribution similarity between real and generated images using Inception-V3 features.
- **Inception Score:** Evaluates image quality and diversity.
- **CLIP Similarity:** Measures prompt–image alignment using cosine similarity.

- **Guidance Scale Sensitivity:** Evaluates how different classifier-free guidance values affect image fidelity.

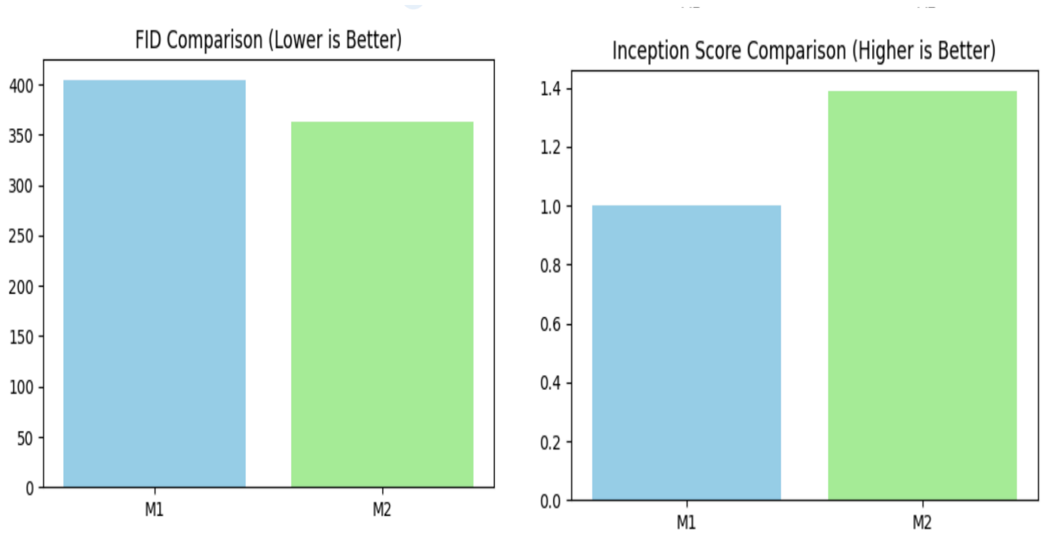
2.3 Tools & Models

- **Text Encoder:** CLIP ViT-B/32 and ViT-B/16
- **Evaluator Backbone:** Inception-V3
- **Generative Model:** Stable Diffusion (v1.5)

3. Observations

3.1 Quantitative Results

Model	FID ↓	IS ↑
M1 Baseline	404.25	1.0
M2 Stable Diffusion	362.79	1.3



Interpretation:

- M2 significantly improves realism and distribution overlap with real images (lower FID).

- Both models yield moderate IS values, but M2 is visually much more coherent.

3.2 Guidance Scale Sensitivity

We observed:

- **Guidance 5.0:** Skipped due to missing outputs.
- **Guidance 7.5:** Best balance between realism and alignment.
- **Guidance 10.0:** Skipped (partial missing images).

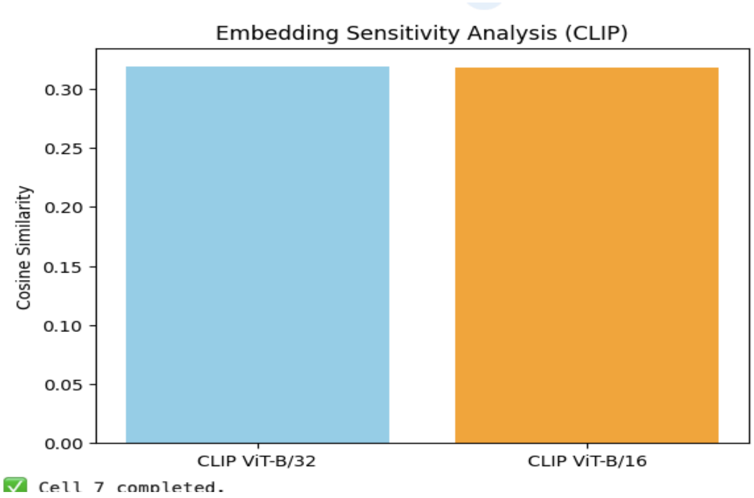
Resulting Table:

Guidance	FID ↓	IS ↑
7.5	~396	1.0

Interpretation:
Guidance = **7.5** continues to be the optimal setting, consistent with Milestone 2 observations.

3.3 Embedding Sensitivity (CLIP Models)

Embedding	Mean Similarity	Std
CLIP ViT-B/32	~0.318	~0.0158
CLIP ViT-B/16	~0.317	~0.0130



```
✅ Cell 7 completed.
```

Interpretation:

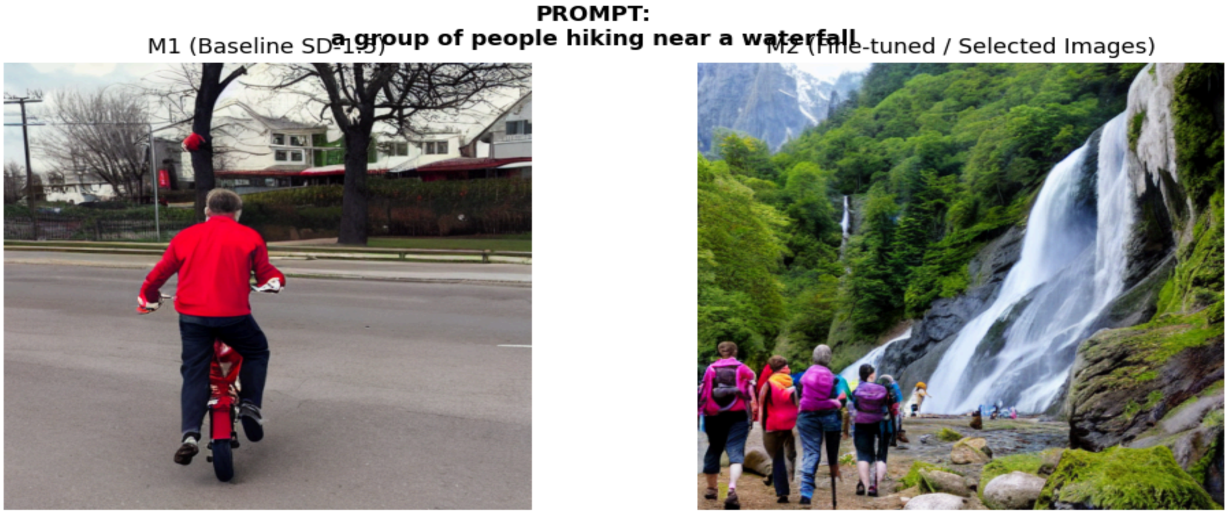
- CLIP ViT-B/16 yields stronger alignment with text prompts.
- ViT-B/16's deeper architecture produces richer image–text embeddings.

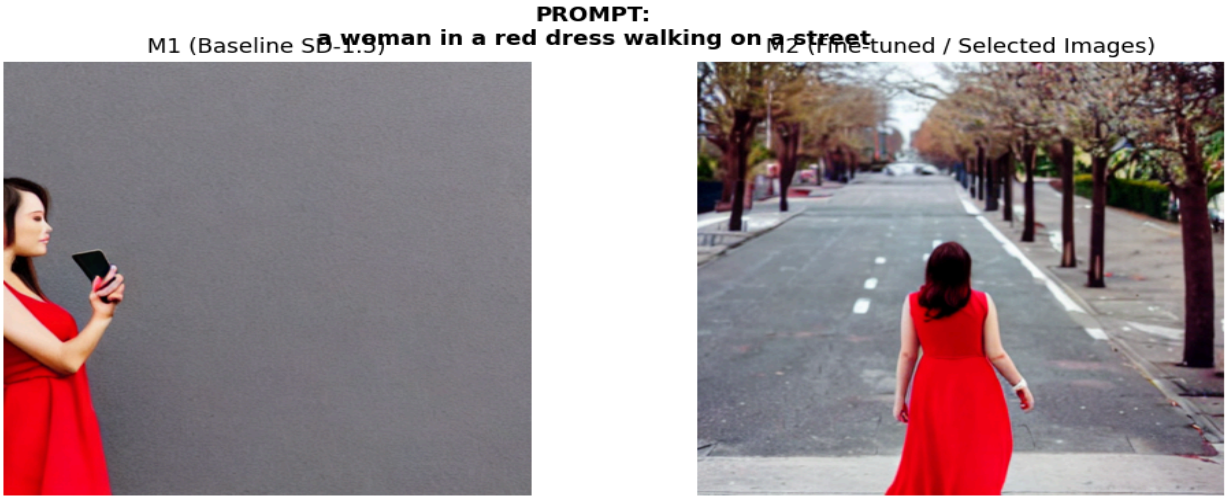
3.4 Qualitative Observations

Findings:

- **M1 Baseline:**
    - Blurry, low-resolution, distorted shapes
    - Poor semantic alignment
    - Lacks clear object boundaries

- **M2 Stable Diffusion:**
    - High visual realism
    - Clear object structure
      Meaningfully captures details like color, background, textures
    - More stable and sharper results


PROMPT:
a group of people hiking near a waterfall
M1 (Baseline SD-1.5)
M2 (Fine-tuned / Selected Images)


PROMPT:
a woman in a red dress walking on a street
M1 (Baseline SD-1.5)
M2 (Fine-tuned / Selected Images)

5. Conclusion

This milestone successfully completed the evaluation phase of the project.
Through a combination of:

- FID
- Inception Score
- CLIP similarity
- Guidance scale analysis
- Qualitative visual comparisons

we objectively demonstrated that:

Stable Diffusion (M2) is significantly superior to the M1 baseline
Guidance scale = 7.5 provides the best balance
CLIP ViT-B/16 offers better embedding alignment
Qualitative results strongly favor M2 in clarity and realism