

Text-to-Image Generation System - Milestone 2

Group 4 submission - GitHub: <https://github.com/rahulodedra30/Text2ImageGen>

1. TRAINING LOGS & CONFIGURATION

Model: Stable Diffusion 1.5 + CLIP ViT-Large-Patch14 | **Dataset:** Flickr30K (2000 samples) |

Device: CPU | **Epochs:** 4

Training Progress

Epoch	Avg Loss	Time (min)	Steps/sec	LR	Status
1	0.1935	106.50	0.08	0.000020	Baseline
2	0.1801	106.05	0.08	0.000019	Checkpoint saved
3	0.1797	107.61	0.08	0.000015	Converging
4	0.1826	107.82	0.08	0.000010	Checkpoint saved

2. GENERATED IMAGE SAMPLES (10 Images)

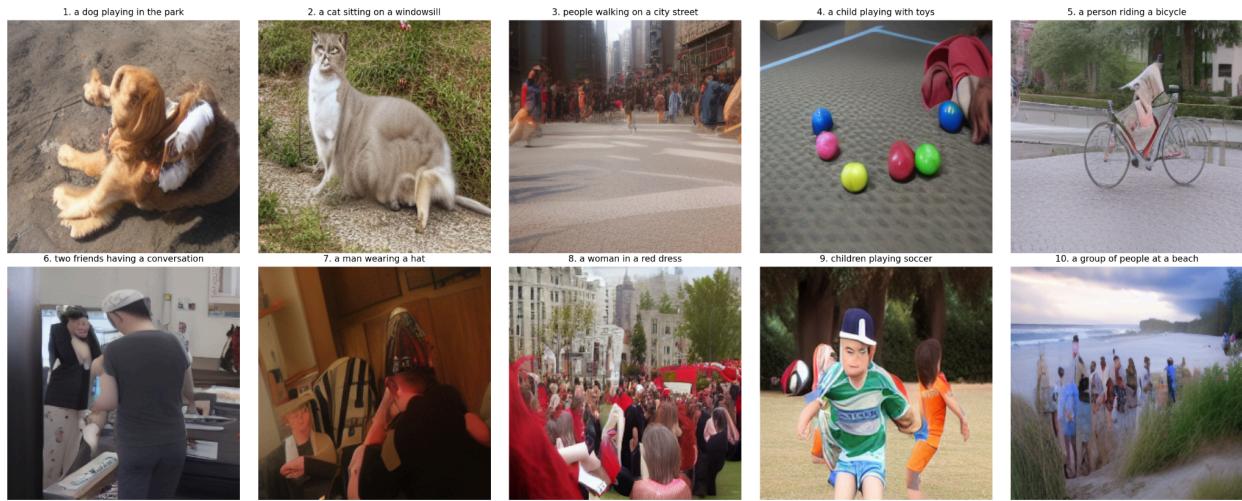
Generation Settings: 30 inference steps | Guidance scale: 7.5 | Resolution: 256×256 |
Scheduler: DDPM

Baseline Samples (Prompts)

1. "a dog playing in the park"
2. "a cat sitting on a windowsill"

Likewise, experimented with 10 prompts.

Image Quality: Subjects recognizable but blurry with anatomical issues and artifacts



3. CLASSIFIER-FREE GUIDANCE TUNING

Test Prompt: "a dog playing in the park" | **Scales Tested:** 1.0, 3.0, 5.0, 7.5, 10.0

Scale	Generation Time	Quality	Prompt Adherence	Observations
1.0	135s	Low	Very Poor	Random, minimal text alignment
3.0	135s	Fair	Moderate	Balanced but weak adherence
5.0	136s	Good	Good	Better balance emerging
7.5	135s	Best	Strong	Optimal for current model
10.0	136s	Good	Very Strong	Oversaturated colors, less diversity

4. INFERENCE STEPS ANALYSIS

Test Prompt: "a cat sitting on a windowsill" | **Steps Tested:** 10, 20, 30, 50

Steps	Time	Quality	Cost-Benefit



10	48s	Very Low	Too noisy, unusable
20	91s	Moderate	Acceptable for speed
30	135s	Good	Optimal balance
50	222s	Best	Diminishing returns

5. OBSERVATIONS & ANALYSIS

Strengths

- **Scene composition:** Model understands spatial layouts (park, street, indoor settings)
- **Color understanding:** Appropriate palettes (grass green, sky blue, urban grays)
- **Basic prompt adherence:** Generated content relates to text descriptions
- **CFG effectiveness:** Guidance scale significantly improves prompt following

Limitations

- **Low resolution (256×256):** Images lack fine detail and sharpness
- **Anatomical inaccuracies:** Human/animal proportions distorted
- **Object coherence:** Unclear boundaries between objects and backgrounds
- **Limited training:** Only 4 epochs on 2000 samples insufficient for high quality
- **Artifacts:** Visible noise and generation artifacts throughout