Generative Project Milestone 2

Team Members (Group 11):

Lochan Enugula
Rishabh Joshi
Rutuja Jadhav
Yaswanth Kumar Reddy Gujjula

1. Objective:

The goal of this milestone was to connect the text-encoding and diffusion components to create a working text-to-image generation pipeline. We used the pretrained Stable Diffusion v1-5 model with its built-in CLIP text encoder to generate images from natural-language prompts. The main focus was to observe how changing parameters such as guidance scale and number of inference steps affected the quality and alignment of generated images.

2. Experimental Setup:

We selected five sample prompts from the cleaned Flickr30k dataset that included both human and object-based scenes. Each prompt was tested with three different guidance scales (5.0, 7.5, and 10.0). We also compared three inference step values (25, 50, and 75) to understand how the noise-removal process influences the final image. All generations were executed on a GPU runtime using the Hugging Face Diffusers library.

3. Observations:

Effect of Guidance Scale:

- At 5.0 → Images were more diverse but not always faithful to the prompt.
- At 7.5 → Best balance between visual realism and semantic accuracy.
- At 10.0 → Sharper visuals but occasionally over-saturated or stiff.

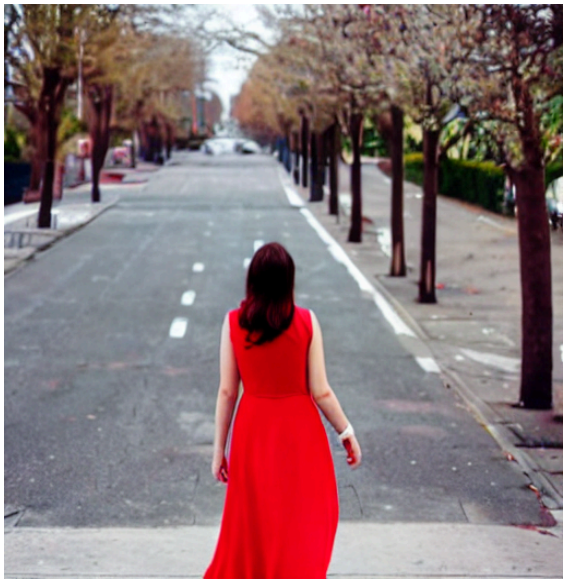
Effect of Inference Steps:

- 25 steps produced faster but slightly noisy images.
- 50 steps gave clearer and well-balanced results.
- 75 steps offered the highest detail but took almost twice the time.

These patterns show the usual trade-off between generation time, image diversity, and visual fidelity.

4. Results and Examples:

The generated samples accurately represented the meaning of their captions.
For instance, prompts like “a dog playing in the snow” and “a woman in a red dress walking on a street” produced visually distinct and contextually appropriate outputs.
Across all prompts, guidance = 7.5 and 50 inference steps yielded the most consistent results.



5. Appendix A - Training Log:

The table below shows the recorded prompts, guidance scales, and corresponding output filenames for baseline generation.

Prompt	Guidance	Steps	Filename
a man riding a red bicycle	5.0	50	outputs/milestone2/a_man_riding_a_red_bicycle_guid5.0.png
a man riding a red bicycle	7.5	50	outputs/milestone2/a_man_riding_a_red_bicycle_guid7.5.png
a man riding a red bicycle	10.0	50	outputs/milestone2/a_man_riding_a_red_bicycle_guid10.0.png
a dog playing in the snow	5.0	50	outputs/milestone2/a_dog_playing_in_the_snow_guid5.0.png
a dog playing in the snow	7.5	50	outputs/milestone2/a_dog_playing_in_the_snow_guid7.5.png
a dog playing in the snow	10.0	50	outputs/milestone2/a_dog_playing_in_the_snow_guid10.0.png
a woman in a red dress walking on a street	5.0	50	outputs/milestone2/a_woman_in_a_red_dress_walking_on_a_street_guid5.0.png
a woman in a red dress walking on a street	7.5	50	outputs/milestone2/a_woman_in_a_red_dress_walking_on_a_street_guid7.5.png
a woman in a red dress walking on a street	10.0	50	outputs/milestone2/a_woman_in_a_red_dress_walking_on_a_street_guid10.0.png
a group of people hiking near a waterfall	5.0	50	outputs/milestone2/a_group_of_people_hiking_near_a_waterfall_guid5.0.png
a group of people hiking near a waterfall	7.5	50	outputs/milestone2/a_group_of_people_hiking_near_a_waterfall_guid7.5.png
a group of people hiking near a waterfall	10.0	50	outputs/milestone2/a_group_of_people_hiking_near_a_waterfall_guid10.0.png
a blue car parked beside a lake	5.0	50	outputs/milestone2/a_blue_car_parked_beside_a_lake_guid5.0.png
a blue car parked beside a lake	7.5	50	outputs/milestone2/a_blue_car_parked_beside_a_lake_guid7.5.png
a blue car parked beside a lake	10.0	50	outputs/milestone2/a_blue_car_parked_beside_a_lake_guid10.0.png

6. Conclusion:

This milestone successfully completed the model integration and baseline image generation phase. Through the implementation of Stable Diffusion, we explored how different parameters — particularly the guidance scale and number of inference steps — affect image quality, prompt accuracy, and overall visual realism. The generated outputs clearly demonstrated that higher guidance values tend to improve concept alignment but can introduce an artistic or less realistic style, while moderate settings maintain a balance between clarity and creativity.

This stage established a strong foundation for our upcoming milestone, where we will focus on quantitative evaluation using metrics such as Fréchet Inception Distance (FID) and Inception Score (IS). These metrics will help us objectively measure visual fidelity and model performance, enabling a deeper understanding of how diffusion-based generation can be optimized for both quality and efficiency.