Course Project

Code of Honor. All external resources used in the project, including research papers, open-source repositories, datasets, and any content or code generated using AI tools, e.g., ChatGPT, GitHub Copilot, Claude, Gemini, must be *clearly cited* in the final submission. The final report must also include *a clear breakdown of individual group member contributions.* Any lack of transparency in the use of external resources or in reporting group contributions will be considered academic dishonesty and will significantly impact the final evaluation.

Topic	Text-guided Image Editing through Latent Modification in VAEs
Category	Tiny AI Products
Supervisor	Likun Cai

Objective Design and implement an image editing pipeline that modifies visual content based on a given textual prompt. The goal is to build a lightweight but effective multimodal editing system that uses text embeddings to guide latent modifications in a VAE-based generative model.

Motivation This project aims to demonstrate how pretrained vision-language models, e.g. CLIP [1], can be used to interpret text into visual transformations, and how generative models can realize such tasks.

Requirements The final system should enable editing of an input image based on a user-provided text instruction, e.g., `change the shirt to red` or `make the background green`. The following components should be implemented or integrated.

1. **Dataset.** Use a small and low-resolution dataset suitable for image autoencoding and editing; recommended options are
   - CelebA (cropped to 64×64), and
   - Fashion-MNIST with synthetic text prompts.

   Even simpler datasets can be used.
2. **Image embedding and reconstruction.** Train a lightweight VAE to learn the image latent space. Pretrained VAEs can be used to save time.
3. **Text-Image embedding alignment.** Use a pretrained model such as CLIP [1] to obtain embeddings of both input images and target text prompts. Implement a method to compute or approximate a latent shift in the image representation that aligns better with the prompt embedding.
   Hint: Pretrained CLIP models can be used to avoid GPU strain.
4. **Latent modification.** Design a method to update the latent vector of the image using the guidance from text, e.g., via interpolation, linear shift, or lightweight learned mapping. Keep this step *simple and interpretable*.
5. **Reconstruction and analysis.** Decode the modified latent and compare the original and edited images.

Milestones

1. *Literature and dataset study.* Select a vision-language models and review its details. Select a dataset, and prepare the data in a usable format with feasible complexity depending on your computation resources.
2. *Image encoder-decoder training.* Train a VAE on the chosen dataset and analyze reconstruction quality.
3. *Embed text and design latent shift method.* Align the image and text embeddings, and design a strategy to compute latent edits.
4. *Editing loop.* Apply latent modifications, decode, and visualize the results. Perform refinement if needed.
5. *Evaluation and reporting.* Evaluate alignment using similarity metrics. Prepare qualitative examples. Elaborate limitations and failure cases.

Submission Guidelines The main body of work is submitted through Git. In addition, each group submits a final paper and gives a presentation. In this respect, please follow these steps.

- Each group must maintain a Git repository, e.g., GitHub or GitLab, for the project. By the time of final submission, the repository should have
  - Well-documented codebase
  - Clear `README.md` with setup and usage instructions
  - A `requirements.txt` file listing all required packages or an `environment.yaml` file with a reproducible environment setup
  - Demo script or notebook showing sample input-output
  - *If applicable,* a `/doc` folder with extended documentation
- A final report (maximum *5 pages*) must be submitted in a PDF format. The report should be written in the provided formal style, including an abstract, introduction, method, experiments, results, and conclusion.
  Important: Submissions that do not use template are considered *incomplete.*
- A 5-minute presentation (maximum *5 slides including the title slide*) is given on the internal seminar on Week 14, i.e., *Aug 4 to Aug 8,* by the group. For presentation, any template can be used.

Final Notes While planning for the milestones please consider the following points.

1. You are encouraged to explore innovative approaches to conditioning or generation as long as the core objectives are met.
2. While computational resources are limited, carefully chosen datasets and training setups can make even diffusion models feasible. Trade-offs, e.g., resolution, training steps, are expected and should be justified.
3. Teams are expected to manage their computing needs and are advised to perform early tests to estimate runtime and training feasibility. As graduate students, team members can use facilities provided by the university, e.g., ECE Facility. Teams are expected to inform themselves about the limitations of the available computing resources and design the model accordingly.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021.