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-to-Image Generation using Pretrained LMs and Generative Architectures
Category	Multimodal Generative Models
Supervisor	Amir Hossein Mobasheri

Objective Design and implement a *multimodal* generative model that takes text descriptions as input and generates corresponding images. For language processing, a pretrained LM, e.g., BERT [1] or RoBERTa [3], is used. The designed multimodal model should integrate this pretrained LM into a generative architecture such as a VAE, GAN, or diffusion model.

Motivation. Multimodal generation is an active research area at the intersection of language and vision. This project provides hands-on experience in using pretrained LMs as building blocks for generative models, exploring how language representations can *condition* visual outputs.

Requirements The final submission should address the following requirements while the details can be freely decided by the group members.

1. Use a pretrained LM to extract textual context. The LM can be *kept frozen* or *fine-tuned*.
2. *Choose* or *construct* a small-scale image-text dataset; potential options include:
   - Fashion-MNIST with synthetic descriptions
   - CIFAR-10 with category-based captions
   - A sub-sampled version of MS-COCO [2] or any open image-text dataset
3. Implement a generative model, e.g., a conditional GAN, VAE, or a simplified diffusion model, conditioned on the LM-generated text embeddings.
4. Evaluate the quality of the generated images and the relevance to input text using suitable metrics, e.g., qualitative inspection, FID, or CLIP similarity.

Milestones The following milestones are to be accomplished through semester.

1. *Literature review and dataset preparation.* Identify the pretrained LM and finalize the dataset choice and preprocess it to get in a usable format.
2. *Build the pipeline to extract text embeddings.* Use the chosen LM to extract textual embeddings and analyze their suitability, e.g., dimensionality, semantic relevance.
3. *Implementation.* Implement the generative model, i.e., GAN, VAE, or diffusion, and integrate the text embeddings as conditioning input. Train the model on the selected dataset.
4. *Evaluation and analysis.* Evaluate model output using benchmark metrics. Perform ablation experiments to understand the role of conditioning.
5. *Final report.* Prepare the final report and present the results, highlighting key design decisions and limitations.

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] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proc. 13th European Conference on Computer Vision (ECCV)*, pages 740–755, Zürich, Switzerland, 2014. Springer.

[3] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.