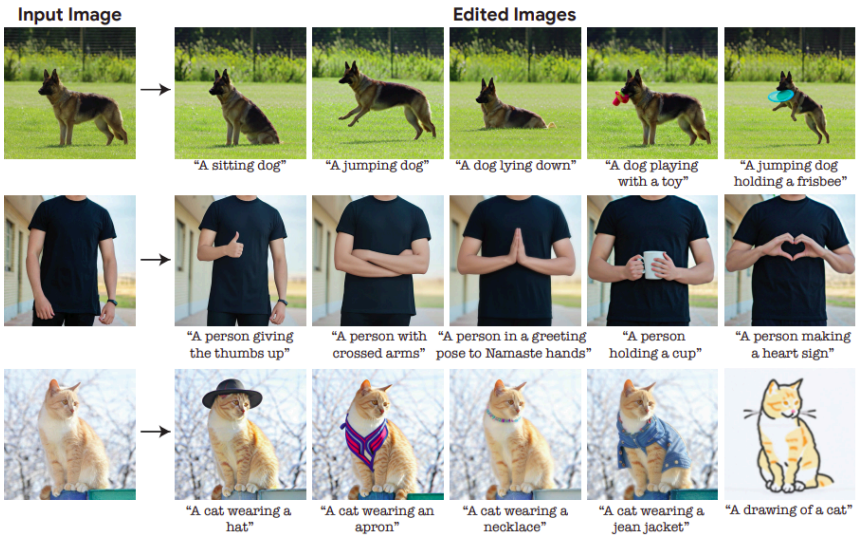
Proposal for the GenAI Project
IA-B 2025

Team name
Fourier No More

Team members
Robin Meneust, Maxime Cabrit, Antoine Charlet, Ethan Pinto

Subject
Robust Text-based Image Editing through Cross-Attention in Diffusion Models

Paper inspiration
[Imagic: Text-Based Real Image Editing with Diffusion Models](#) (Bahjat Kawar et al.)



Summary

"Robust Text-based Image Editing through Cross-Attention in Diffusion Models" is a subject that was given to Stanford students, and that was suggested for this Gen-AI project. We will use the paper "Imagic: Text-Based Real Image Editing with Diffusion Models" as a base since there doesn't seem to be a paper associated with this subject. This paper introduces a novel method for editing real images using text prompts. Unlike previous approaches that often require multiple images or are limited to specific editing types, Imagic enables complex, non-rigid semantic edits on a single high-resolution image. By leveraging a pre-trained text-to-image diffusion model, the method aligns text embeddings with both the input image and the desired textual edit, allowing for transformations such as altering object posture or composition while preserving original image characteristics. In our Gen-AI project, we aim to build upon this work by implementing the Cross-Attention mechanism and exploring further enhancements to improve the model's performance and versatility.

Limits

The paper’s authors identified different limits to their models, and suggested using cross-attention control, which is what this subject is about. Sometimes the edit is too weak or affects unwanted parts of the image (e.g. zoom…). They included those difficult examples in a benchmark (TEdBench). that is available, so that we can test our method.

Improvements suggestions

Our project aims to enhance text-based image generation by incorporating several key improvements. First, we plan to implement cross-attention control mechanisms to refine the alignment between textual prompts and generated images, ensuring more accurate and contextually relevant outputs. Typically it should help us solve the limits mentioned above. Additionally, we intend to add the possibility to generate consistent image sequences from a single text prompt, such as creating a series of images depicting the action of clapping hands, resulting in an animated GIF. Furthermore, we aim to enable the model to accept multiple input images, allowing for the generation of new images that seamlessly integrate elements from the provided references. For example, by inputting images of a cat and a dog along with the prompt "sleeping on a couch," the model would produce an image featuring both animals, resembling their respective references, resting together on a sofa.