Combining Large Language and Diffusion Models for Intuitive Image Editing.

What is the problem?

- **Challenge in Image Editing**: The primary problem addressed in this paper is the difficulty in teaching generative models to follow human-written instructions for image editing. Traditional methods either require full image descriptions, rely on manual fine-tuning, or struggle with generalizing edits to real-world images. Furthermore, existing approaches lack the ability to make diverse and accurate image edits based on simple text instructions.

What has been done earlier?

- Earlier models focused on specific editing tasks like style transfer or domain translation, often requiring complex setups such as latent space manipulation or inverting images for edits or full descriptions. Recent models combine pretrained models but lack consistency in edits and require complex processes.

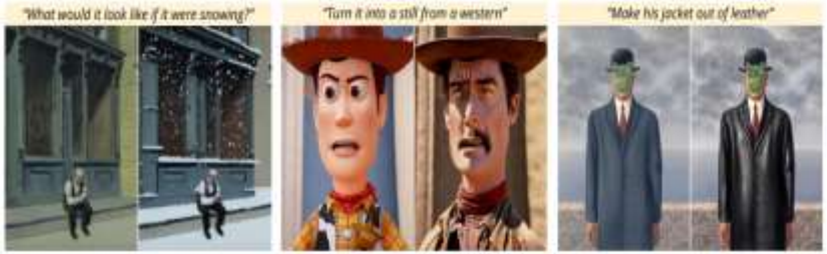
"What would it look like if it were snowing?"
"Turn it into a still from a western"
"Make his jacket out of leather"

Figure 1. Given **an image** and **an instruction** for how to edit that image, our model performs the appropriate edit. Our model does not require full descriptions for the input or output image, and edits images in the forward pass without per-example inversion or fine-tuning.

Mohammad Talha Quamar, B421027

What are the remaining challenges? What novel solution proposed by the authors to solve the problem?

- Limited visual quality, struggles with spatial edits, accumulation of artifacts, and inherited biases

.

Solution

- Data Generation: Combines GPT-3 and Stable Diffusion to generate a large, diverse dataset for training a model that follows editing instructions.
- Efficient Editing Model: Trains a diffusion model, InstructPix2Pix to perform quick and diverse image edits directly in the forward pass without requiring fine-tuning or additional images.

Mohammad Talha Quamar, B421027