IMPROVE IMAGE EDITING BY INSTRUCTING

Lê Ngọc Thành

[1] Trường ĐH Công Nghệ Thông Tin

[2] University of Science
HCMC, Vietnam

[3] National Institute of Informatics

What ?

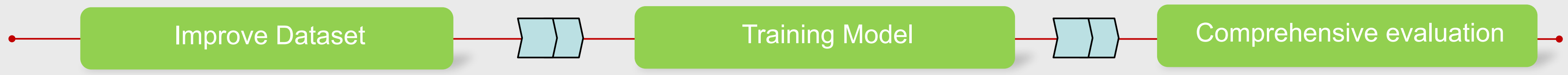
The main goal is to improve image editing performance based on text instructions by:

- Increasing spatial reasoning accuracy: Improving the ability to handle location-specific instructions.
- Reducing bias in the training data.
- Adding metrics such as FID and Human Evaluation Score for more comprehensive evaluation.

Why ?

- This study aims to overcome the limitations of **InstructPix2Pix**, especially in handling spatial positioning and bias reduction, and improving image quality.
- These improvements not only improve performance but also expand applications in design, advertising, and education, meeting the growing demand for intelligent and accurate image editing tools.

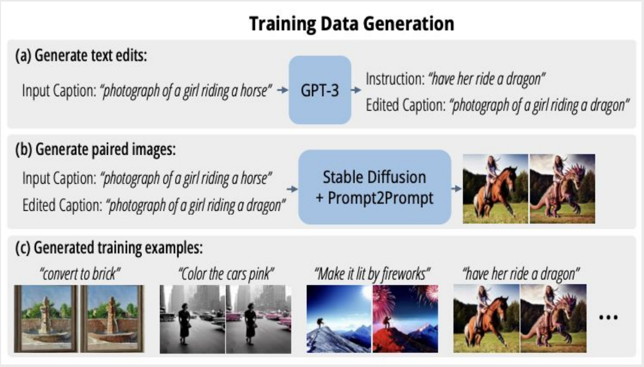
Overview



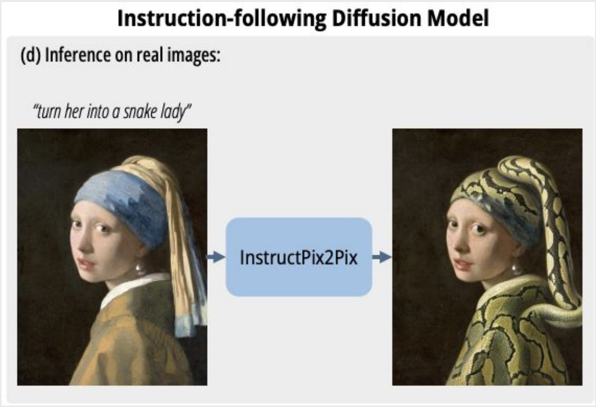
Improve Dataset

Training Model

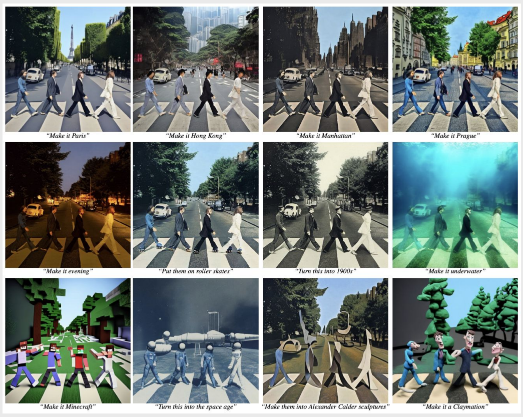
Comprehensive evaluation



Training Data Generation
(a) Generate text edits:
Input Caption: "photograph of a girl riding a horse"
GPT-3
Instruction: "have her ride a dragon"
Edited Caption: "photograph of a girl riding a dragon"
(b) Generate paired images:
Input Caption: "photograph of a girl riding a horse"
Edited Caption: "photograph of a girl riding a dragon"
Stable Diffusion + Prompt2Prompt
(c) Generated training examples:
"convert to brick"
"Color the cars pink"
"Make it lit by fireworks"
"have her ride a dragon"
...



Instruction-following Diffusion Model
(d) Inference on real images:
"turn her into a snake lady"
InstructPix2Pix



Description

1. Improve Dataset

- Combine simulation data from 3D tools like Blender (e.g. "put the ball in the lower left corner") with real-world data from COCO, Open Images, and manual annotations to improve spatial localization and reduce bias.

2. Training Model

- Model Architecture Improvement for Spatial Guidance Understanding.
- Integrating a spatial attention layer and flexible guidance processing mechanism enables the model to perform complex commands like "rotate the view 45 degrees" or "move the object to the left".

3. Comprehensive Evaluation

- Apply FID to measure image quality, Human Evaluation Score to evaluate the level of compliance with requirements, and build specific tests like "add two cats" or "move the object to the right 10%".

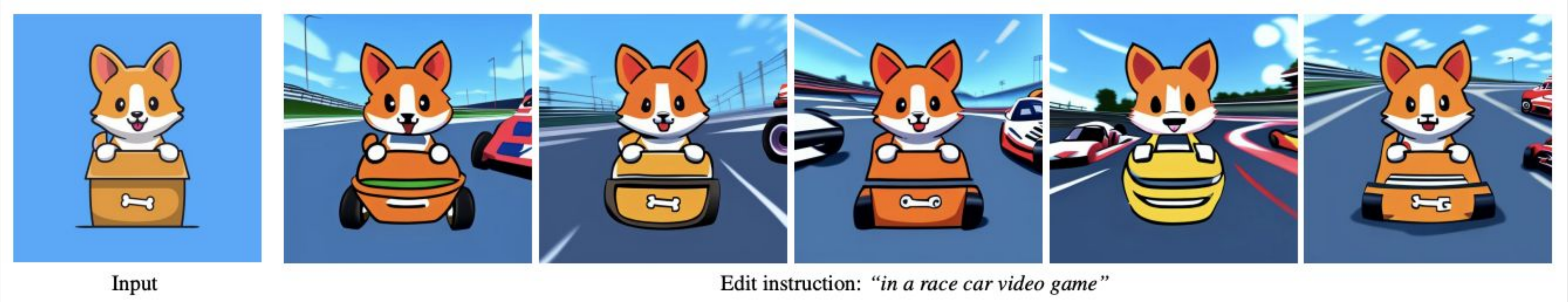


Figure 1 . By varying the latent noise, our model can produce many possible image edits for the same input image and instruction.

Lê Ngọc Thành– Trường Đại học Công Nghệ Thông Tin (UIT)
TEL : 0978839799 Email : thanhln.19@grad.uit.edu.vn