

THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút):
<https://youtu.be/sn8AfeHwQtA>
- Link slides (dạng .pdf đặt trên Github của nhóm):
<https://github.com/thanhlengoc/CS2205.FEB2025/blob/main/Tha%CC%80nh%20Le%CC%82%20Ngo%CC%A3c%20-%20CS2205.FEB2025.DeCuong.FinalReport.Template.Slide.pdf>
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*
- *Lớp Cao học, mỗi nhóm một thành viên*

- Họ và Tên: Lê Ngọc Thành
- MSSV: 240101074



- Lớp: CS2205.FEB2025
- Tự đánh giá (điểm tổng kết môn): 8.5/10
- Số buổi vắng: 0
- Số câu hỏi QT cá nhân: 2
- Số câu hỏi QT của cả nhóm: 5
- Link Github:
<https://github.com/thanhlengoc/CS2205.FEB2025>

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

CẢI TIẾN CHỈNH SỬA HÌNH ẢNH THEO HƯỚNG DẪN VĂN BẢN SỬ DỤNG
MÔ HÌNH INSTRUCTPIX2PIX

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

IMPROVED IMAGE EDITING WITH TEXT GUIDE USING INSTRUCTPIX2PIX
MODEL

TÓM TẮT *(Tối đa 400 từ)*

Đề tài "Cải tiến chỉnh sửa hình ảnh theo hướng dẫn văn bản sử dụng mô hình InstructPix2Pix" tập trung khắc phục các hạn chế của InstructPix2Pix [1] trong việc lý luận không gian, giảm thiên kiến dữ liệu và nâng cao đánh giá hiệu suất. Phương pháp đề xuất bao gồm cải thiện dữ liệu huấn luyện bằng cách tích hợp các nguồn dữ liệu thực tế và dữ liệu không gian được tạo ra, nâng cấp kiến trúc mô hình với nền tảng Stable Diffusion mới hơn cùng các cơ chế chú ý không gian và hướng dẫn động, đồng thời áp dụng các phương pháp đánh giá toàn diện để đo lường chất lượng và sự phù hợp của kết quả. Nghiên cứu này hướng đến tăng cường độ chính xác và khả năng ứng dụng thực tiễn trong các lĩnh vực sáng tạo nội dung.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Chỉnh sửa hình ảnh theo hướng dẫn văn bản là một bài toán quan trọng trong lĩnh vực trí tuệ nhân tạo, nhằm mục đích cho phép người dùng thay đổi hình ảnh một cách tự nhiên và chính xác dựa trên các chỉ dẫn bằng ngôn ngữ tự nhiên, ví dụ: “thêm một cái cây vào góc phải” hoặc “chuyển phong cách thành tranh vẽ tay”. Bài toán mà đề tài muốn giải quyết xuất phát từ những hạn chế của InstructPix2Pix, một mô hình tiên phong trong lĩnh vực này, bao gồm: cải thiện khả năng lý luận không gian (spatial reasoning) khi xử lý các yêu cầu như “di chuyển vật sang trái”, thiên kiến dữ liệu từ các nguồn huấn luyện không đồng đều, và thiếu các thang đo hiệu quả để đánh giá

chất lượng chỉnh sửa.

- Input: Đầu vào là một hình ảnh gốc và một chuỗi văn bản mô tả thay đổi mong muốn (ví dụ: “thêm mặt trời ở góc trên”).
- Output: Kết quả là một hình ảnh đã được chỉnh sửa, phản ánh chính xác hướng dẫn văn bản với chất lượng cao và chi tiết rõ ràng.

Lý do chọn đề tài nằm ở tiềm năng lớn của công nghệ này trong việc đơn giản hóa các công cụ sáng tạo, đáp ứng nhu cầu ngày càng tăng về chỉnh sửa hình ảnh tự động và thông minh. Khả năng ứng dụng thực tế rất rộng, từ thiết kế đồ họa (tạo mẫu nhanh), quảng cáo (chỉnh sửa hình ảnh sản phẩm), đến giáo dục (hỗ trợ học sinh minh họa ý tưởng). Tính thời sự của nghiên cứu được thể hiện qua sự phát triển mạnh mẽ của các mô hình Diffusion và nhu cầu thực tiễn về các giải pháp AI sáng tạo, khi mà các công cụ như DALL·E 2 hay Stable Diffusion đang định hình xu hướng công nghệ hiện nay.



Hình 1: Mô hình InstructPix2Pix gặp hạn chế trong việc thay đổi góc nhìn, dễ tạo ra thay đổi quá mức, khó cô lập đối tượng được chỉ định và gặp khó khăn khi hoán đổi hoặc sắp xếp lại các đối tượng.

MỤC TIÊU

- Cải thiện nhận thức không gian: Tăng khả năng xử lý các hướng dẫn chỉnh sửa ảnh liên quan đến vị trí, như “thêm cây ở góc dưới bên phải”.
- Giảm thiên kiến dữ liệu: Giảm thiểu thiên lệch từ dữ liệu huấn luyện, đặc biệt về giới tính và nghề nghiệp.
- Nâng cao đánh giá hiệu suất: Bổ sung FID (Fréchet Inception Distance) [7] và Human Evaluation Score, so sánh với DiffusionCLIP [5] và DALL·E 2 [6].

NỘI DUNG VÀ PHƯƠNG PHÁP

Nội dung:

- **Text generation:** Fine-tuning GPT-3 trên 700 mẫu dữ liệu do con người tạo, sinh ra hướng dẫn (instructions) và chú thích hình ảnh (image caption) từ LAION-Aesthetics V2 6.5+.
- **Image generation:** Dùng Stable Diffusion v2.1 [2] với Prompt-to-Prompt [4] để tạo cặp hình ảnh (trước và sau chỉnh sửa) từ input/edited image caption, đảm bảo tính nhất quán (chỉ cập nhật ảnh đúng chỗ yêu cầu).
- **Data filtering:** Áp dụng CLIP [3] (độ tương đồng hình ảnh: 0.75, image-caption: 0.2, hướng ảnh: 0.2) để giữ lại các mẫu data chất lượng.
- **Dynamic Guidance [8]:** Tự động điều chỉnh mức độ ảnh hưởng của văn bản dựa trên độ phức tạp của chỉ dẫn, chẳng hạn tăng trọng số cho các yêu cầu chi tiết như “thêm bóng đổ dưới cây”.
- **Thang đo FI:** để đo lường độ tương đồng giữa hình ảnh gốc và hình ảnh chỉnh sửa với các tiêu chuẩn thực tế.
- **Human Evaluation Score:** dựa trên khảo sát người dùng, để đánh giá mức độ khớp giữa kết quả và hướng dẫn văn bản.

Phương pháp:

1. Tạo tập dữ liệu chỉnh sửa hình ảnh.

	Input LAION caption	Edit instruction	Edited caption
Human-written (700 edits)	<i>Yefim Volkov, Misty Morning</i>	<i>make it afternoon</i>	<i>Yefim Volkov, Misty Afternoon</i>
	<i>girl with horse at sunset</i>	<i>change the background to a city</i>	<i>girl with horse at sunset in front of city</i>
	<i>painting-of-forest-and-pond</i>	<i>Without the water.</i>	<i>painting-of-forest</i>

GPT-3 generated (>450,000 edits)	<i>Alex Hill, Original oil painting on canvas, Moonlight Bay</i>	<i>in the style of a coloring book</i>	<i>Alex Hill, Original coloring book illustration, Moonlight Bay</i>
	<i>The great elf city of Rivendell, sitting atop a waterfall as cascades of water spill around it</i>	<i>Add a giant red dragon</i>	<i>The great elf city of Rivendell, sitting atop a waterfall as cascades of water spill around it with a giant red dragon flying overhead</i>
	<i>Kate Hudson arriving at the Golden Globes 2015</i>	<i>make her look like a zombie</i>	<i>Zombie Kate Hudson arriving at the Golden Globes 2015</i>

2. Huấn luyện mô hình Diffusion trên tập dữ liệu đã tạo.

- **Bổ sung dữ liệu huấn luyện đa dạng**
 - Kết hợp dữ liệu mô phỏng từ công cụ 3D như Blender (ví dụ: “đặt quả bóng ở góc dưới bên trái”) với dữ liệu thực tế từ COCO, Open Images và

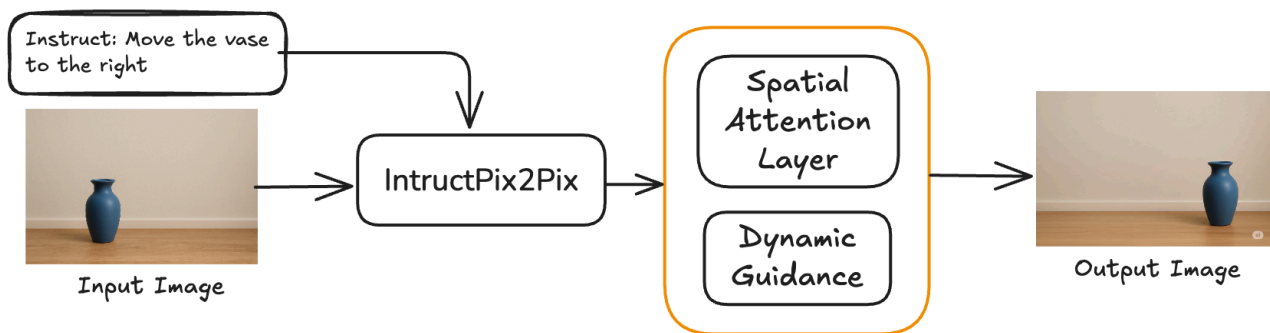
chú thích thủ công để cải thiện khả năng định vị không gian và giảm thiên kiến.

- **Nâng cấp kiến trúc mô hình để hiểu hướng dẫn không gian**

- Tích hợp lớp chú ý không gian (spatial attention layer) và cơ chế xử lý hướng dẫn linh hoạt giúp mô hình thực hiện các lệnh phức tạp như “xoay góc nhìn 45 độ” hoặc “di chuyển vật thể sang trái”.

- **Thiết kế tiêu chí đánh giá chuyên biệt cho chỉnh sửa không gian**

- Áp dụng FID để đo chất lượng ảnh, Human Evaluation Score để đánh giá mức độ phù hợp với yêu cầu, và xây dựng bài kiểm tra cụ thể như “thêm hai con mèo” hoặc “di chuyển vật sang phải 10%”.



Kiến trúc mô hình cải tiến

KẾT QUẢ MONG ĐỢI

Định lượng:

- Độ chính xác định vị không gian: Tăng từ ~50% lên ~75%, đo bằng tỷ lệ thành công trong các yêu cầu như “di chuyển vật sang trái”.
- Thiên kiến dữ liệu: Giảm ~20%, đánh giá qua sự đa dạng của đối tượng và ngữ cảnh.
- Chất lượng hình ảnh: FID giảm từ ~25 xuống ~22.5, tương đương cải thiện ~10%.
- Tỷ lệ thành công tổng thể: Đạt ~80-85% cho các chỉnh sửa đa dạng.

So sánh:

- So với InstructPix2Pix: Cải tiến dự kiến vượt trội về định vị không gian (~75%

so với ~50%) và giảm thiên kiến (~20% giảm so với cao).

- So với DiffusionCLIP: Nhanh hơn (~40 lần, ~10s/ảnh so với ~400s/ảnh), chính xác hơn (~15% về tỷ lệ thành công).
- So với DALL·E 2: Dự kiến chính xác hơn (~10% về định vị không gian) và linh hoạt hơn với hướng dẫn văn bản.

TÀI LIỆU THAM KHẢO (*Định dạng DBLP*)

- [1]. Tim Brooks, Aleksander Holynski, Alexei A. Efros: InstructPix2Pix: Learning to Follow Image Editing Instructions. CVPR 2023: 18392-18402
- [2]. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer: High-resolution image synthesis with latent diffusion models. CVPR 2022: 1-45
- [3]. Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, Daniel Cohen-Or: StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators. ACM Trans. Graph. 41(4): 149:1-149:15 (2022)
- [4]. Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, Daniel Cohen-Or: Prompt-to-Prompt Image Editing with Cross-Attention Control. CoRR abs/2208.01626 (2022)
- [5]. Gwanghyun Kim, Taesung Kwon, Jong Chul Ye: DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation. CVPR 2022: 2426-2435
- [6]. Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen: Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv preprint arXiv:2204.06125, 2022.
- [7]. Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. NeurIPS 2017
- [8]. Felix Koulischer, Johannes Deleu, Gabriel Raya, Thomas Demeester, Luca Ambrogioni: Dynamic Negative Guidance of Diffusion Models. ICLR 2025