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

Lê Ngọc Thành - 240101074

Tóm tắt

- Lớp: CS2205.FEB2025
- Link Github của nhóm: https://github.com/thanhlengoc/CS2205.FEB2025
- Link YouTube video: https://youtu.be/sn8AfeHwQtA

Lê Ngọc Thành - 240101074

Giới thiệu

- Chỉnh sửa hình ảnh là công việc khá phức tạp, đòi hỏi kỹ năng chuyên môn về ảnh. Đề tài hướng tới công cụ đơn giản, thân thiện cho người dùng không chuyên.
- **InstructPix2Pix** [1] là mô hình diffusion chỉnh sửa ảnh theo văn bản, nhưng còn hạn chế về lý luận không gian và bị ảnh hưởng bởi thiên kiến dữ liệu.
- Vậy làm sao cải thiện InstructPix2Pix để nâng cao khả năng lý luận không gian và giảm thiên kiến trong chỉnh sửa ảnh?
- Đề tài nhằm cải tiến bài toán chỉnh sửa hình ảnh dựa trên mô hình **InstructPix2Pix**:
  - **Input:** Hình ảnh gốc và hướng dẫn văn bản (ví dụ: "di chuyển vật sang trái").
  - **Output:** Hình ảnh đã chỉnh sửa theo yêu cầu.

Mục tiêu

Nâng cấp mô hình **InstructPix2Pix** nhằm khắc phục hạn chế về lý luận không gian (Spatial Reasoning):

- **Cải thiện lý luận 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 FID**: 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.

Nội dung và Phương pháp

Phương pháp:

- Tạo tập dữ liệu huấn luyện chỉnh sửa hình ảnh dựa trên LAION và GPT-3.

	Input LAION caption	Edit instruction	Edited caption
Human-written (700 edits)	*Yefim Volkov, Misty Morning*	*make it afternoon*	*Yefim Volkov, Misty Afternoon*
	girl with horse at sunset	*change the background to a city*	*girl with horse at sunset in front of city*
	painting-of-forest-and-pond	*Without the water.*	*painting-of-forest*
	...	...	...
GPT-3 generated (>450,000 edits)	*Alex Hill, Original oil painting on canvas, Moonlight Bay*	*in the style of a coloring book*	*Alex Hill, Original coloring book illustration, Moonlight Bay*
	The great elf city of Rivendell, sitting atop a waterfall as cascades of water spill around it	*Add a giant red dragon*	*The great elf city of Rivendell, sitting atop a waterfall as cascades of water spill around it with a giant red dragon flying overhead*
	Kate Hudson arriving at the Golden Globes 2015	*make her look like a zombie*	*Zombie Kate Hudson arriving at the Golden Globes 2015*
	...	...	...

Hình 2: Minh hoạ cách tạo dữ liệu huấn luyện từ InstructPix2Pix

Nội dung và Phương pháp

Phương pháp:

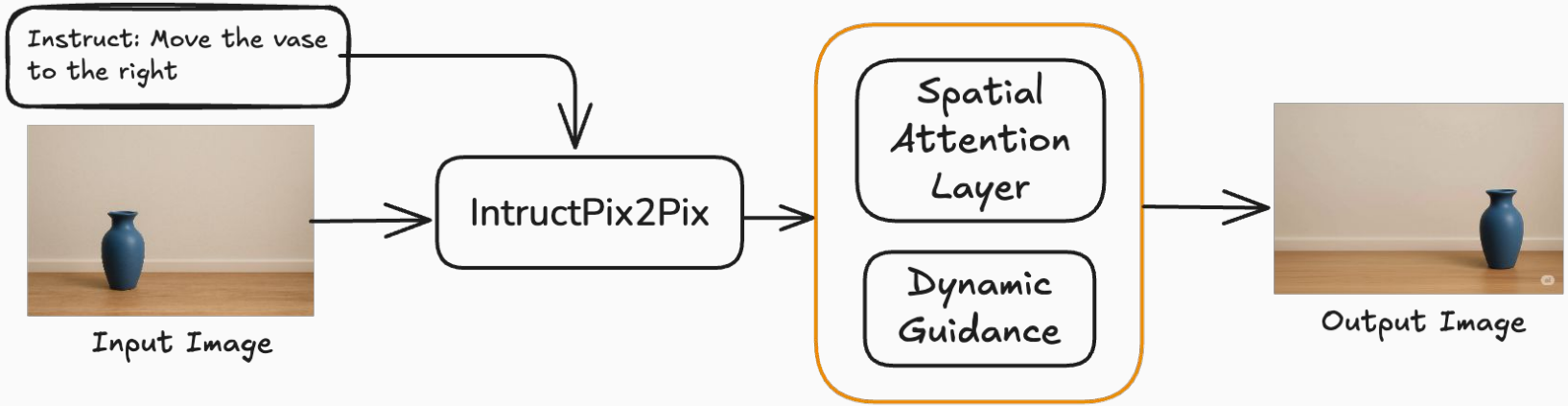
- **Bổ sung dữ liệu huấn luyện đa dạng**
  Kết hợp dữ liệu mô phỏng từ công cụ 3D như Blender (ví dụ: *"đặt quả bóng ở góc dưới bên trái"*) với dữ liệu thực tế từ COCO, Open Images và chú thích thủ công để cải thiện khả năng định vị không gian và giảm thiên kiến.

- **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%"*.

Nội dung và Phương pháp

Kiến trúc mô hình để hiểu hướng dẫn không gian:



Hình 3: Kiến trúc mô hình đề xuất

Kết quả dự kiến

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

- [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