THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút): *https://youtu.be/fKLgz0VTSFM*
- Link slides (dạng .pdf đặt trên Github của nhóm): *https://github.com/SoulOfWindTGN/CS2205.CH183/blob/main/Slides.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: Trần Gia Nghĩa
- MSSV: 240101019

- Lớp: CS2205.CH183
- Tự đánh giá (điểm tổng kết môn): 9.5/10
- Số buổi vắng: 0
- Số câu hỏi QT cá nhân: 6
- Link Github: https://github.com/mynameuit/CS2205.CH183/

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

CÁ NHÂN HÓA MÔ HÌNH TẠO SINH ẢNH TỪ VĂN BẢN VỚI ÍT MẪU DỮ LIỆU

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

FEW-SHOT PERSONALIZATION OF TEXT-TO-IMAGE GENERATION MODELS

TÓM TẮT *(Tối đa 400 từ)*

Trong những năm gần đây, *mô hình tạo sinh ảnh từ văn bản* [1] đã đạt được nhiều bước tiến đáng kể, cho phép tạo ra hình ảnh có độ chân thực và sáng tạo cao từ các mô tả ngôn ngữ. Tuy nhiên, phần lớn mô hình này được huấn luyện trên lượng dữ liệu rất lớn và có tính tổng quát cao, dẫn đến hiện tượng thiếu cá nhân hóa trong ảnh tạo sinh. Khi người dùng chỉ có một tập ảnh hoặc phong cách riêng hạn chế, mô hình thường không thể tái hiện đầy đủ chi tiết cũng như đặc điểm phong cách cá nhân, đòi hỏi giải pháp nâng cao khả năng cá nhân hóa dù dữ liệu huấn luyện không nhiều.

Một số tiếp cận đã được đề xuất nhằm cá nhân hóa mô hình, chẳng hạn như *tinh chỉnh trên tập dữ liệu nhỏ (fine-tuning)* [2] hoặc sử dụng kỹ thuật *mã hóa đặc trưng tùy biến (embedding)* [3] nhưng vẫn tồn tại nhiều khó khăn. Cụ thể, các mô hình tạo sinh ảnh thường có số lượng tham số lớn, dễ dẫn đến quá khớp (overfitting) khi dữ liệu hạn chế. Ngược lại, những phương pháp hạn chế việc can thiệp vào tham số gốc lại khó duy trì chi tiết của chủ thể. Đồng thời, việc bảo toàn chất lượng ảnh và khả năng tạo sinh đa dạng, đặc biệt trong các kịch bản phức tạp (như tạo sinh đồng thời nhiều chủ thể) vẫn là một thách thức.

Nhằm giải quyết những hạn chế này, nghiên cứu đề xuất một phương pháp mới để cải thiện khả năng cá nhân hóa mô hình tạo sinh ảnh từ văn bản với ít mẫu dữ liệu của người dùng. Trọng tâm của chúng tôi là nghiên cứu áp dụng **kỹ thuật tăng cường dữ liệu** nhằm mở rộng tính đa dạng cho tập huấn luyện, giúp mô hình học được nhiều đặc trưng phong phú hơn từ số lượng mẫu nhỏ. Bên cạnh đó, chúng tôi nghiên cứu **cải thiện quy trình huấn luyện** nhằm hạn chế tối đa tình trạng mất mát kiến thức (catastrophic forgetting) của mô hình gốc, trong khi vẫn đảm bảo khả năng tái tạo chi tiết của đối tượng cần cá nhân hóa.

GIỚI THIỆU *(Tối đa 1 trang A4)*

Những năm gần đây, mô hình khuếch tán (diffusion models) [1] đã cho thấy ưu thế nổi bật trong tác vụ tạo sinh ảnh có độ chi tiết và tính nghệ thuật cao. Nhiều mô hình khuếch tán khác [4] thậm chí còn cho phép người dùng kiểm soát bố cục (layout guidance) nhằm chỉ định cách sắp xếp không gian, vị trí đối tượng, rất hữu ích trong thiết kế, kể chuyện hay sáng tác nghệ thuật. Dù vậy, hầu hết các mô hình này đều được huấn luyện trên bộ dữ liệu lớn và mang tính tổng quát, làm hạn chế khả năng

đưa những khái niệm (concept) cá nhân hóa của người dùng vào ảnh sinh ra. Để khắc phục, nhiều phương pháp cá nhân hóa đã được đề xuất. Ví dụ, Textual Inversion [3] chỉ tối ưu một số embedding, gán từ khóa mới cho các khái niệm thị giác mong muốn (chẳng hạn phong cách nghệ thuật hoặc một đối tượng cụ thể). Tương tự, LoRA (Low-Rank Adaptation) thêm các ma trận hạng thấp vào một số lớp của mô hình [5] giảm thiểu việc chỉnh sửa trọng số gốc và tiêu tốn nhiều tài nguyên tính toán. Tuy nhiên, chúng phụ thuộc nhiều vào kiến thức đã có trong mô hình gốc, nên khó tái hiện những chi tiết phức tạp. Ngược lại các kỹ thuật tinh chỉnh trực tiếp mô hình gốc [2] có thể học hiệu quả dữ liệu mới, nhưng dễ xảy ra quá khớp (overfitting) hoặc làm giảm tính đa dạng đầu ra. Bên cạnh đó, việc sinh đồng thời nhiều khái niệm (multiple concepts) lại làm gia tăng độ phức tạp, do các mô hình khuếch tán thường gặp khó khăn trong việc xử lý quan hệ giữa nhiều đối tượng, nhất là khi chúng có tính tương đồng cao.

Trên cơ sở đó, đề tài này nghiên cứu một chiến lược cá nhân hóa mô hình tạo sinh ảnh, hướng đến việc sinh một hoặc nhiều chủ thể cùng lúc. Mô hình nhận đầu vào là (i) các ảnh tham chiếu của người dùng, trong đó chứa chủ thể cần đưa vào ảnh sinh ra; và (ii) câu mô tả văn bản (prompt) mô tả nội dung hay bối cảnh mong muốn. Đầu ra là hình ảnh mới, có sự hiện diện của khái niệm (concept) mới tương ứng với nội dung mô tả. Để đạt được điều này, chúng tôi nghiên cứu cải thiện quy trình tinh chỉnh tham số mô hình (fine-tuning), kết hợp với kỹ thuật tăng cường dữ liệu nhằm giảm nguy cơ quá khớp và hạn chế việc trộn lẫn đặc trưng giữa các khái niệm. Đặc biệt, phương pháp hướng đến việc không cần thêm điều kiện phụ trợ khi suy luận, nhưng vẫn duy trì khả năng tái hiện chi tiết chủ thể.

Hình 1: Minh họa đầu vào và đầu ra của bài toán

MỤC TIÊU *(Viết trong vòng 3 mục tiêu)*
Nhóm nghiên cứu đặt ra các mục tiêu sau đây:

- Nghiên cứu quy trình tinh chỉnh tham số cho mô hình khuếch tán (ví dụ Stable Diffusion) nhằm nhúng các khái niệm (concept) cá nhân hóa một cách hiệu quả từ một tập ảnh tham chiếu nhỏ vào quá trình sinh ảnh.
- Xây dựng, tích hợp và đánh giá hiệu quả của kỹ thuật tăng cường dữ liệu để hạn chế quá khớp (overfitting) trong điều kiện ít dữ liệu.
- Triển khai và thử nghiệm và đánh giá phương pháp đề xuất cho cả tác vụ tạo sinh đơn hoặc đa khái niệm.

NỘI DUNG VÀ PHƯƠNG PHÁP
Nội dung 1: Tìm hiểu kiến trúc của một số mô hình khuếch tán (Diffusion Models).

- Mục tiêu:
    - Hiểu rõ về kiến trúc và nguyên lý hoạt động của các mô hình khuếch tán.
    - Phân tích ý nghĩa của từng thành phần trong mô hình, nhằm nắm bắt cách các khối (block) tương tác với nhau trong việc xử lý dữ liệu đa thể thức (hình ảnh - văn bản).
- Phương pháp:
    - **Nghiên cứu lý thuyết**: Đọc các bài báo khoa học, tài liệu và giáo trình liên quan để nắm rõ nền tảng lý thuyết về quá trình khuếch tán (forward diffusion) và khôi phục ngược (reverse diffusion).
    - **Phân tích triển khai thực tế**: Tham khảo mã nguồn chính thức (hoặc mã nguồn mở) của một số mô hình khuếch tán tiêu biểu (VD: DDPM, Latent Diffusion) nhằm hiểu sâu hơn về cách hiện thực các thành phần lý thuyết (UNet, scheduler, noise prediction network, v.v.).
    - **So sánh và đối chiếu**: Từ việc nghiên cứu lý thuyết và mã nguồn, rút ra điểm chung, khác biệt, ưu/nhược của từng mô hình khi áp dụng cho bài toán sinh ảnh.

Nội dung 2: Tìm hiểu các phương pháp cá nhân hóa mô hình khuếch tán cho việc tạo sinh ảnh từ văn bản

- Mục tiêu:
    - Nắm bắt được các hướng tiếp cận phổ biến cho bài toán cá nhân hóa việc tạo sinh hình ảnh từ văn bản.
    - Xây dựng một quy trình tinh chỉnh mô hình hiệu quả dựa trên những tiến bộ của các công trình nghiên cứu trước đó, từ đó đề xuất cải tiến hoặc lựa chọn phương án phù hợp với bối cảnh dữ liệu ít (few-shot).
- Phương pháp:
    - **Khảo sát tài liệu**: Nghiên cứu các bài báo khoa học liên quan đến cá nhân hóa mô hình tạo sinh ảnh từ văn bản. Tổng hợp, phân loại các phương pháp chính (VD: tinh chỉnh embedding, thêm ma trận hạng thấp, tinh chỉnh trực tiếp mô hình gốc).
    - **Phân tích ưu/nhược điểm**: Đánh giá chi tiết từng hướng (tài nguyên huấn luyện, chất lượng ảnh, khả năng tổng quát, mức độ phức tạp triển khai, v.v.).

    - **Đề xuất quy trình tinh chỉnh**: Dựa trên kết quả nghiên cứu, xây dựng quy trình tinh chỉnh phù hợp với mục tiêu giữ lại tri thức gốc đồng thời nhúng hiệu quả khái niệm mới.

Nội dung 3: Tìm hiểu các phương pháp tăng cường dữ liệu cho bài toán few shot learning.

- Mục tiêu:
    - Phân tích ưu điểm và hạn chế của những kỹ thuật tăng cường dữ liệu (data augmentation) thường dùng cho kịch bản few-shot, nơi dữ liệu tham chiếu rất hạn chế.
    - Lựa chọn và triển khai một hoặc nhiều phương pháp tăng cường dữ liệu giúp bổ sung tính đa dạng cho tập ảnh tham chiếu.
- Phương pháp:
    - **Nghiên cứu học thuật**: Tìm hiểu các bài báo, tài liệu về kỹ thuật tăng cường dữ liệu cho hình ảnh (VD: xoay, cắt, lật, thay đổi độ sáng, ghép nền, v.v.) và các phương pháp nâng cao (mixup, cutmix, style transfer, v.v.).
    - **Lựa chọn chiến lược tăng cường**: Dựa trên đặc điểm của tập ảnh tham chiếu (ít dữ liệu, cùng/khác bối cảnh, cùng/khác kiểu), đề xuất danh sách các kỹ thuật augmentation tương thích.
    - **Thử nghiệm & đánh giá**: Thực hiện augmentation thực tế trên tập ảnh nhỏ, đánh giá hiệu quả dựa trên mức độ gia tăng tính đa dạng, tránh trùng lặp, và xem xét tác động đến chất lượng mô hình khi tinh chỉnh.

Nội dung 4: Thực nghiệm và đánh giá kết quả

- Mục tiêu:
    - Áp dụng quy trình tinh chỉnh cá nhân hóa (từ Nội dung 2) cùng chiến lược tăng cường dữ liệu (từ Nội dung 3) lên một mô hình khuếch tán cụ thể.
    - Đánh giá toàn diện chất lượng mô hình sau tinh chỉnh, bao gồm khả năng tái hiện khái niệm mới, giữ lại tri thức gốc, sinh ảnh đa dạng và tránh quá khớp.
    - So sánh hiệu suất với các phương pháp hiện có.
- Phương pháp:
    - **Chuẩn bị dữ liệu và cài đặt môi trường**: Thu thập tập ảnh tham chiếu, cài đặt mô hình khuếch tán gốc và lập kịch bản huấn luyện.
    - **Triển khai:** tích hợp phương pháp tăng cường dữ liệu kết hợp với quy trình tinh chỉnh đã đề xuất.
    - **Đánh giá định lượng và định tính**:
        - **Định lượng:** Sử dụng các chỉ số đánh giá độ giống với ảnh tham chiếu (DreamSim, DINO), mức độ bám sát văn bản (CLIP, ImageReward) và chất lượng hình ảnh (CLIP-IQA).
        - **Định tính:** Tiến hành khảo sát người dùng nhằm đánh giá trực quan về độ trung thực, tính đa dạng, khả năng bám sát mô tả văn bản.

- **Phân tích và so sánh:** Thực hiện các thử nghiệm khác nhau để đánh giá về hiệu năng, độ chính xác của phương pháp và so sánh mô hình với các phương pháp hiện có.

KẾT QUẢ MONG ĐỢI

- Tài liệu báo cáo chi tiết cơ sở lý thuyết về phương pháp đề xuất.
- Chi tiết kết quả thực nghiệm, đưa ra các đánh giá về hiệu năng, số liệu định lượng và định tính và tính hiệu quả của phương pháp.
- Kết quả so sánh với các phương pháp hiện nay trên cùng một tập dữ liệu.

TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

[1]. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer: High-Resolution Image Synthesis with Latent Diffusion Models. CVPR 2022: 10674-10685

[2]. Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, Kfir Aberman: DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. CVPR 2023: 22500-22510

[3]. Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, Daniel Cohen-Or: An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. ICLR 2023

[4]. Lvmin Zhang, Anyi Rao, Maneesh Agrawala: Adding Conditional Control to Text-to-Image Diffusion Models. ICCV 2023: 3813-3824

[5]. Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, Yixiao Ge, Ying Shan, Mike Zheng Shou: Mix-of-Show: Decentralized Low-Rank Adaptation for Multi-Concept Customization of Diffusion Models. NeurIPS 2023