TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN, ĐHQG-HCM

KHOA KHOA HỌC MÁY TÍNH

BÁO CÁO ĐỒ ÁN MÔN HỌC

ĐỀ TÀI:

Adapter cho Multimodal Prompting nhằm mở rộng khả năng điều khiển của Stable Diffusion

Môn học: CS431.Q12 - Kỹ thuật học sâu và ứng dụng

Giảng viên hướng dẫn: TS. Nguyễn Vinh Tiệp

Thực hiện bởi nhóm các sinh viên, bao gồm:

1. Đỗ Minh Dũng 23520326
2. Huỳnh Diên Thục 23521555
3. Đinh Trần Duy Trường 23521688

Mục Lục

Chương 1 Tóm tắt 4

Chương 2 Giới thiệu đề tài 4

2.1 Giới thiệu đề tài. 4
2.1.1 Bối cảnh 4
2.1.2 Mục tiêu 5
2.1.3 Ý nghĩa 5
2.2 Mô tả. 6
2.2.1 Input 6
2.2.2 Output 7

Chương 3 Phương pháp thực hiện 7

3.1 UniControlNet(Global) 7
3.2 IP-Adapter 9
3.3 IP-Adapter MLP 11
3.4 IP-Adapter LoRA 11
3.5 IP-Adapter MLP_LoRA 12

Chương 4 Dữ liệu và độ đo 13

4.1 Dữ liệu 13
4.1.1 Tóm tắt bộ dữ liệu 13
4.1.2 Phân chia dữ liệu 14

4.2 Độ đo 14
4.2.1 CLIP–T 14
4.2.2 CLIP–I 15

Chương 5 Kết quả đánh giá 15

5.1 Thiết lập thực nghiệm 15

5.2 Kết quả 16
5.2.1 Kết quả định lượng 16
5.2.2 Kết quả định tính 17

Chương 6 Những cải tiến mới so với bài trình bày trước 20

Chương 7 Kết luận 21

Chương 8 Reference 21

CHƯƠNG 1 TÓM TẮT

Trong đề tài này, nhóm tập trung nghiên cứu và triển khai các kỹ thuật Adapter nhằm bổ sung khả năng multimodal prompting cho các mô hình sinh ảnh dựa trên khuếch tán, cụ thể là Stable Diffusion v1.5. Mục tiêu chính là cho phép mô hình kết hợp thông tin từ image prompt và text prompt, giúp tạo ra ảnh mới vừa bám sát đặc trưng thị giác của ảnh tham chiếu, vừa tuân thủ nội dung văn bản mô tả.

Nhóm tiến hành khảo sát hai hướng tiếp cận phổ biến gồm UniControlNet (Global) và IP-Adapter, đồng thời triển khai ba biến thể cải tiến của IP-Adapter là IP-Adapter MLP, IP-Adapter LoRA và IP-Adapter MLP_LoRA. Các thiết kế cải tiến nhằm giải quyết những hạn chế của phiên bản gốc. Kết quả thực nghiệm cho thấy IP-Adapter gốc đạt hiệu năng cao nhất, trong khi IP-Adapter LoRA mang lại sự cân bằng tốt giữa chất lượng và số tham số huấn luyện. UniControlNet (Global) tuy nhẹ nhưng vẫn duy trì được chất lượng cạnh tranh. Ngược lại, các biến thể dựa trên MLP không mang lại cải thiện mà còn làm suy giảm hiệu năng do ảnh hưởng đến không gian embedding CLIP.

Đề tài khẳng định rằng các Adapter là một hướng tiếp cận nhẹ, hiệu quả và dễ tích hợp, phù hợp để mở rộng khả năng điều khiển của mô hình diffusion bằng ảnh tham chiếu mà không cần can thiệp vào mô hình gốc.

CHƯƠNG 2 GIỚI THIỆU ĐỀ TÀI

2.1 Giới thiệu đề tài.

2.1.1 Bối cảnh

- Gần đây, các mô hình diffusion sinh ảnh từ văn bản như GLIDE, DALL·E 2, Imagen, Stable Diffusion, eDiff-I hay RAPHAEL đã đạt chất lượng tạo sinh ấn tượng. Tuy vậy, việc điều khiển mô hình chỉ thông qua text prompt thường không đủ chính xác: người dùng phải dựa vào các kỹ thuật prompt engineering phức tạp, và bản thân văn

bản khó mô tả đầy đủ những cảnh phức tạp hoặc các khái niệm giàu chi tiết.

- Một hướng tiếp cận tự nhiên là sử dụng image prompt, vì hình ảnh mang thông tin trực quan phong phú hơn nhiều so với câu chữ. Một số mô hình như SD Image Variations hoặc SD unCLIP cải thiện đáng kể độ bám sát hình tham chiếu bằng cách fine-tune trực tiếp diffusion model trên embedding từ ảnh. Tuy nhiên, các phương pháp này thường đòi hỏi tài nguyên huấn luyện lớn, thiếu tính linh hoạt, khó áp dụng trên nhiều base model khác nhau, và kém tương thích với text prompt hoặc các mô-đun điều khiển cấu trúc như ControlNet.
- Các hướng dựa trên adapter hiện có ví dụ như T2I-Adapter hoặc Uni-ControlNet tận dụng CLIP image encoder kết hợp với một mạng chiếu nhỏ để hòa trộn đặc trưng ảnh vào không gian đặc trưng văn bản. Dù nhẹ và linh hoạt hơn so với fine-tuning toàn mô hình, các phương pháp này thường chỉ cung cấp mức kiểm soát cơ bản, vẫn còn hạn chế đáng kể so với các mô hình fine-tune đầy đủ.

2.1.2 Mục tiêu

- Mục tiêu của đề tài là phát triển một bộ Adapter nhằm bổ sung khả năng sinh ảnh theo *image prompt* cho các mô hình diffusion *text-to-image* đã được huấn luyện sẵn (ví dụ Stable Diffusion v1.5), mà không cần chỉnh sửa hay fine-tune UNet gốc.
- Đề tài hướng đến thiết kế một mô-đun gọn nhẹ, mang tính plug-and-play, có thể dễ dàng tích hợp vào nhiều mô hình Stable Diffusion khác nhau mà không làm thay đổi cấu trúc hay trọng số của mô hình nền.
- Bộ Adapter cần đảm nhiệm vai trò trích xuất, nén và ánh xạ đặc trưng hình ảnh vào không gian điều kiện của mô hình diffusion một cách hiệu quả, từ đó giúp mô hình tạo ra ảnh mới vừa tuân theo nội dung hình tham chiếu, vừa giữ được tính sáng tạo và khả năng điều khiển bằng văn bản.

2.1.3 Ý nghĩa

2.1.3.1 Về mặt khoa học / kỹ thuật

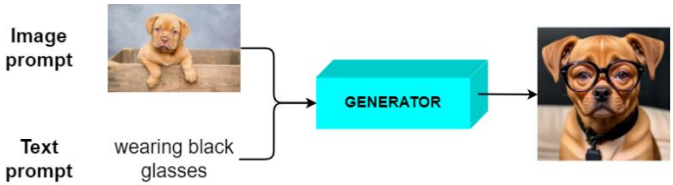
- Tái hiện và kiểm chứng các phương pháp Adapter cho image prompt trong bối cảnh các mô hình Stable Diffusion hiện đại, góp phần củng cố tính đúng đắn và khả năng tổng quát của phương pháp trong những môi trường và cấu hình khác nhau.

- Phân tích và đánh giá định lượng/định tính cách mà đặc trưng ảnh được hòa trộn với đặc trưng văn bản trong quá trình sinh ảnh, từ đó cung cấp hiểu biết sâu hơn về cơ chế multimodal prompting trong mô hình diffusion.
- Thử nghiệm một số điều chỉnh kỹ thuật nhỏ (ví dụ: thay thế các lớp Linear ở phương pháp IP-Adapter), giúp làm rõ ảnh hưởng của từng thành phần đến chất lượng đầu ra.
- Góp phần hoàn thiện quy trình đánh giá tính hiệu quả, nhẹ và khả năng tích hợp của các Adapter, một hướng tiếp cận có giá trị trong tối ưu chi phí và khả năng mở rộng của mô hình diffusion.

2.1.3.2 Về mặt ứng dụng / thực tiễn

- Hỗ trợ các ứng dụng sáng tạo như thiết kế đồ họa, minh họa, concept art, thời trang, media…, nơi cần tạo các biến thể hình ảnh dựa trên một ảnh nguồn theo cách nhanh chóng và tiết kiệm chi phí.
- Tối ưu chi phí triển khai hệ thống generative AI vì mô hình chỉ cần bổ sung thêm Adapter nhẹ, không cần huấn luyện lại hoặc hiệu chỉnh UNet, dễ vận hành trên phần cứng hạn chế.
- Tăng tính linh hoạt khi sử dụng Stable Diffusion, bởi Adapter có thể plug-and-play với nhiều phiên bản mô hình và tương thích với các mô-đun hỗ trợ như LoRA.
- Tạo nền tảng thực tiễn cho việc mở rộng thêm chức năng trong tương lai, như điều kiện hóa bằng phong cách, bố cục, mặt người hoặc các đặc trưng đa phương thức khác.

2.2 Mô tả.



Hình 1: *Trực quan hóa Input & Output*

2.2.1 Input

- **Image prompt**: Một ảnh RGB do người dùng cung cấp, đóng vai trò là nguồn thông tin thị giác tham chiếu. Ảnh này chứa các đặc trưng quan trọng như bố cục tổng thể, phong cách hình ảnh, chủ thể chính hoặc màu sắc chủ đạo.

- **Text prompt**: Một mô tả bằng ngôn ngữ tự nhiên, giúp bổ sung hoặc làm rõ các yếu tố mà người dùng mong muốn xuất hiện trong ảnh sinh ra.

2.2.2 Output

- Ảnh mới được tạo sinh bởi mô hình diffusion, trong đó kết quả đầu ra được kỳ vọng bám sát nội dung hoặc phong cách từ ảnh tham chiếu, đồng thời phản ánh đúng các ràng buộc hoặc hướng dẫn được nêu trong text prompt.

CHƯƠNG 3 PHƯƠNG PHÁP THỰC HIỆN

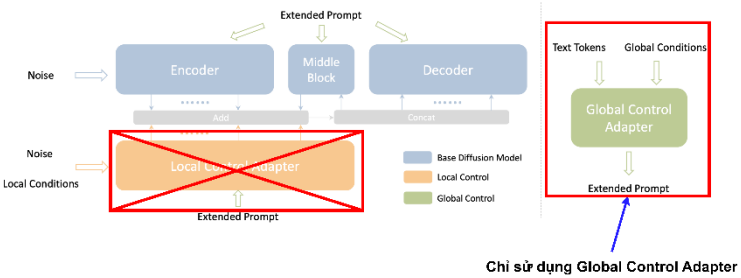
Trong nghiên cứu này, nhóm lựa chọn và áp dụng các phương pháp khác nhau để giải quyết bài toán, bao gồm **Uni-ControlNet (Global), IP-Adapter**. Bên cạnh việc áp dụng IP-Adapter, nhóm còn triển khai ba phương pháp cải tiến của IP-Adapter bao gồm **IP-Adapter MLP_LoRA, IP-Adapter LoRA và IP-Adapter MLP**. Các cải tiến này được thiết kế nhằm nâng cao khả năng điều khiển đặc trưng ảnh, cải thiện khả năng sinh ảnh từ image prompt và text prompt, đồng thời tối ưu hóa hiệu quả huấn luyện và triển khai so với phiên bản gốc.

3.1 UniControlNet(Global)

- **UniControlNet** là một kiến trúc mở rộng dựa trên ControlNet, cho phép mô hình sinh ảnh được điều khiển bởi nhiều dạng thông tin khác nhau.
- UniControlNet cho phép tích hợp nhiều loại ảnh điều kiện (image condition) từ tổng quát đến chi tiết dựa vào hai module Global Control Adapter và Local Control Adapter.
    + Global Control Adapter: Global Control Adapter sử dụng tín hiệu toàn cục (global control), cụ thể là embedding của ảnh hoặc các thông tin tổng quan, để ảnh hưởng đến quá trình sinh ảnh.
    + Local Control Adapter:

- Local Control Adapter tập trung vào điều khiển chi tiết tại từng vùng ảnh, sử dụng các đặc trưng cục bộ như mask, keypoints hoặc texture.
- Nó cho phép mô hình điều chỉnh chính xác từng vị trí trong ảnh mà không ảnh hưởng đến cấu trúc tổng thể.

– Trong đề tài, mục tiêu là multimodal prompt, tức là kết hợp thông tin văn bản với embedding tổng quan của ảnh để điều khiển sinh ảnh. Vì vậy, Global Control Adapter đã đủ khả năng tích hợp tín hiệu toàn cục với prompt văn bản, đáp ứng yêu cầu multimodal mà không cần Local Control Adapter.



***Hình 2**: Tổng quan về UniControlNet*

– Global Control Adapter:

+ Tín hiệu toàn cục c_g (image embedding) được đưa qua condition encoder h_g gồm nhiều lớp feedforward.

$$y^g = h_g(c_g)$$

+ Embedding được chia thành K global tokens

$$y_i{}^g = h_g(c_g)[(i-1)d \sim id],\ \ i \in [1, K]$$

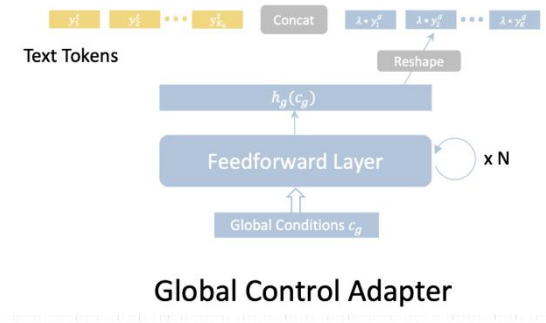
+ Ghép với K_0 token văn bản để tạo prompt mở rộng y_{ext}

$$y_{ext} = [y_1{}^t, \dots, y_{K_0}{}^t, \lambda y_1{}^g, \dots, \lambda y_K{}^t]$$

+ Toàn bộ các lớp cross-attention dùng prompt mở rộng thay vì chỉ token văn bản.

$$Q = ZW_Q,\ \ K = y_{ext}W_K,\ \ V = y_{ext}W_V$$

Trong đó: λ dùng để cân bằng mức độ ảnh hưởng của hình ảnh đầu vào khi đưa vào cross-attention.



***Hình 3**: Chi tiết cơ chế Global Control Adapter*

3.2 IP-Adapter

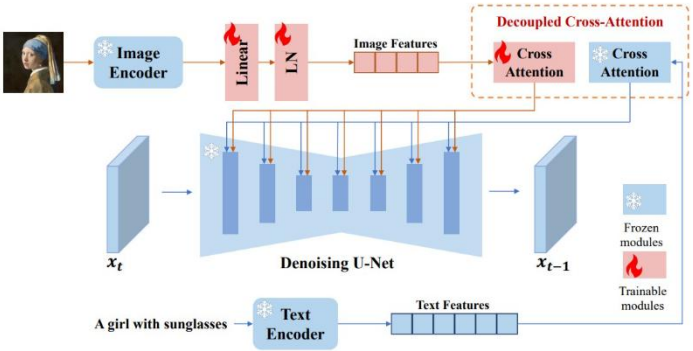
- **IP–Adapter** là một phương pháp mở rộng cho các mô hình sinh ảnh dựa trên Stable Diffusion hoặc các mô hình LLM kết hợp visual embedding.
- Mục tiêu chính của IP–Adapter là tích hợp thông tin hình ảnh (image prompt) trực tiếp vào quá trình sinh ảnh mà không cần huấn luyện lại toàn bộ mô hình. Phương pháp này tận dụng các adapter nhỏ gọn để điều chỉnh mô hình theo tín hiệu đầu vào, mang lại tính linh hoạt và hiệu quả.
- IP–Adapter sử dụng một lớp **Linear Projection** và **Normalization** để chiếu embedding hình ảnh toàn cục thành một chuỗi đặc trưng có chiều tương đương với đặc trưng đầu vào của mô hình khuếch tán.
- IP–Adapter sử dụng một Linear Projection để ánh xạ embedding hình ảnh vào không gian đặc trưng của mô hình khuếch tán. Lựa chọn này mang lại hiệu quả tính toán, nhưng cũng đặt ra một câu hỏi quan trọng: *(1)Tại sao mô hình IP-Adapter lại sử dụng lớp Linear, trong khi UniControlNet sử dụng một mạng neural phi tuyến là các lớp Feedforward để ánh xạ? Các mạng neural phi tuyến (ví dụ MLP) có tiềm năng học ánh xạ biểu diễn phức tạp hơn tại sao IP-Adapter lại không sử dụng?*
- IP–Adapter sử dụng cơ chế **Decoupled Cross-Attention**, tách biệt vai trò của prompt văn bản và embedding hình ảnh trong quá trình attention.

$$Z' = Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V$$

$$Z'' = Attention(Q, K', V') = Softmax\left(\frac{Q(K')^T}{\sqrt{d}}\right)V'$$

$$Z^{new} = Attention(Q, K, V) + \lambda \cdot Attention(Q, K', V')$$

Trong đó: $Q = ZW_Q, K = c_t W_K, V = c_t W_V, K' = c_i W'_K, V' = c_i W'_V$, Z là đặc trưng truy vấn, c_t và c_i lần lượt là đặc trưng văn bản và đặc trưng hình ảnh. λ dùng để cân bằng mức độ ảnh hưởng của hình ảnh đầu vào.



***Hình 4**: Tổng quan IP-Adapter*

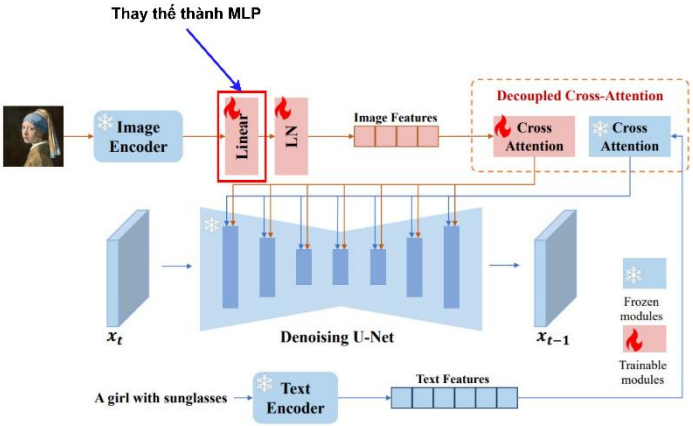
– IP–Adapter sử dụng Decoupled Cross-Attention để tách biệt vai trò giữa embedding văn bản và embedding hình ảnh. Việc sử dụng thêm các lớp cross-attention giúp tránh làm méo đặc trưng vốn rất nhạy của text encoder, nhưng đồng thời làm tăng số lượng tham số phải học, từ đó nảy sinh một câu hỏi quan trọng: *(2)Có cách nào để làm giảm tham số mà không ảnh hưởng đến khả năng tạo sinh ảnh không?*

3.3 IP-Adapter MLP

- Phiên bản **IP–Adapter MLP** được đề xuất nhằm trả lời câu hỏi số *(1)* là khắc phục hạn chế của lớp Linear Projection trong việc học ánh xạ biểu diễn phức tạp từ embedding hình ảnh sang không gian đặc trưng của mô hình khuếch tán.
- Thay vì sử dụng một lớp Linear duy nhất, biến thể này sử dụng một mạng Multilayer perceptron(MLP) nhỏ gọn.
- Với embedding hình ảnh đầu vào emb_i, MLP hai lớp ẩn được biểu diễn như sau

$$c_i = MLP(emb_i) = W_3\sigma_2\big(W_2\sigma_1(W_1 emb_i)\big)$$

Trong đó: σ_1, σ_2 là các hàm phi tuyến GELU



Hình 5: Tổng quan IP-Adapter MLP

3.4 IP-Adapter LoRA

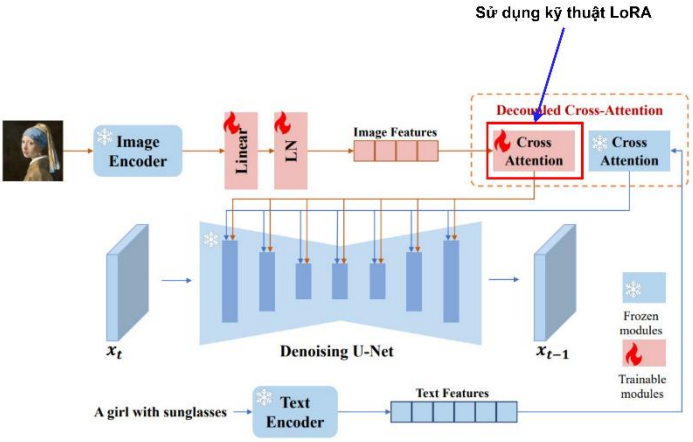
- Trong kiến trúc IP–Adapter gốc, cơ chế Decoupled Cross-Attention sử dụng hai ma trận chiếu đặc trưng:
  - $W'_K \in \mathbb{R}^{d \times d_k}$ để chiếu embedding hình ảnh thành không gian Key
  - $W'_V \in \mathbb{R}^{d \times d_v}$ để chiếu embedding hình ảnh thành không gian Value

- Hai ma trận này trong cơ chế Decoupled Cross-Attention có kích thước rất lớn khiến chúng làm tăng đáng kể số lượng tham số phải học. Trong khi mục tiêu của đề tài là tạo ra một module nhỏ gọn.
- Để khắc phục vấn đề này và trả lời câu hỏi số *(2)*, biến thể **IP-Adapter LoRA** áp dụng kỹ thuật Low-Rank Adaptation lên chính hai ma trận lớn nói trên, cho phép thay vì học toàn bộ trọng số đầy đủ, mô hình chỉ học các cập nhật hạng thấp.

$$W'_K = W_K + B_1 A_1$$

$$W'_V = W_V + B_2 A_2$$

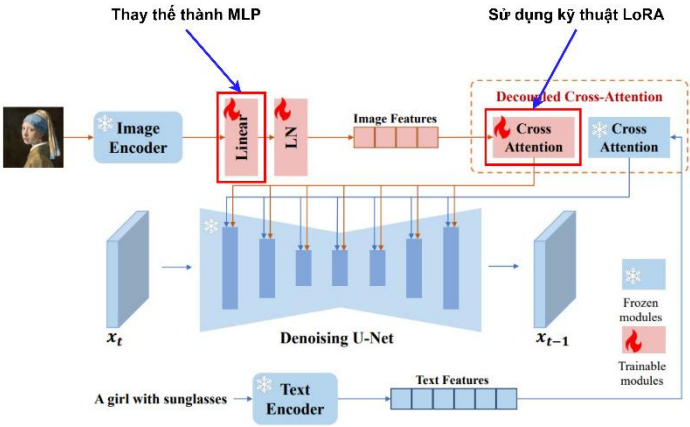
trong đó A, A' có kích thước rank nhỏ, còn B, B' là ma trận mở rộng.



***Hình 6**: Tổng quan IP-Adapter LoRA*

3.5 IP-Adapter MLP_LoRA

- Biến thể thứ ba, **IP–Adapter MLP_LoRA**, là sự kết hợp của hai hướng cải tiến trên:
  - + Thay Linear Projection bằng MLP
  - + Áp dụng LoRA cho các ma trận ánh xạ W'_K và W'_V trong Decoupled Cross-Attention.
- Mục tiêu là tạo ra phiên bản IP-Adapter vừa có khả năng biểu diễn mạnh hơn, vừa duy trì tính nhẹ và hiệu quả, phù hợp cho bài toán multimodal có yêu cầu cao về linh hoạt và tiết kiệm tài nguyên.



Hình 7: Tổng quan IP-Adapter MLP_LoRA

CHƯƠNG 4 DỮ LIỆU VÀ ĐỘ ĐO

4.1 Dữ liệu

4.1.1 Tóm tắt bộ dữ liệu

- **Conceptual Captions** là một bộ dữ liệu hình ảnh lớn, gồm khoảng 3,3 triệu ảnh kèm theo mô tả.
- Khác với nhiều bộ dữ liệu chú thích ảnh truyền thống, trong đó các mô tả thường được tuyển chọn và kiểm duyệt thủ công, Conceptual Captions sử dụng các chú thích thu thập trực tiếp từ web, nhờ đó thể hiện sự đa dạng về phong cách ngôn ngữ và cách diễn đạt.
- Nhờ vậy, Conceptual Captions trở thành một bộ dữ liệu phong phú, đa dạng, thích hợp cho việc huấn luyện các mô hình text-to-image và các nghiên cứu về multimodal learning.

4.1.2 Phân chia dữ liệu

- Bộ dữ liệu được chia thành hai tập chính:
  - Tập huấn luyện
    - 576 shard
    - 2.905.954 trên tổng số 3.318.333 mẫu
  - Tập kiểm thử
    - 16 shard
    - 13.443 trên tổng số 15.840 mẫu
- Trong đề tài này, tập dữ liệu huấn luyện được sử dụng để huấn luyện các bộ Adapter, trong khi quá trình đánh giá được thực hiện trên một tập con gồm 1000 ảnh của tập kiểm thử nhằm đánh giá khả năng sinh ảnh của các mô hình text-to-image khi kết hợp với các Adapter này.

4.2 Độ đo

Trong các mô hình sinh ảnh dựa trên multimodal prompt, việc đánh giá mức độ phù hợp giữa ảnh sinh ra và thông tin đầu vào (văn bản hoặc ảnh tham chiếu) thường được thực hiện trong không gian embedding của CLIP. Trong báo cáo này, hai độ đo được sử dụng là CLIP–T và CLIP–I.

4.2.1 CLIP–T

- CLIP–T dùng để đo mức độ tương đồng giữa ảnh được sinh ra và mô tả văn bản của ảnh gốc.
- Độ đo này được tính bằng điểm CLIPScore thông qua cosine similarity trong không gian embedding của CLIP.

$$CLIP - T = \frac{1}{N \times K} \sum_{n=1}^{N} \sum_{k=1}^{K} \cos\left(\hat{v}_{n,k}^{gen}, \hat{v}_{n}^{text}\right)$$

Trong đó:

- N: Số mẫu gốc.
- K: Số ảnh được sinh ra cho mỗi mẫu.
- $\hat{v}_{n,k}^{gen}$: vector embedding CLIP của ảnh sinh ra thứ k ứng với mẫu n.
- $\hat{v}_{n}^{text}$: vector embedding CLIP của text prompt của mẫu n.

4.2.2 CLIP–I

- CLIP–I dùng để đo mức độ tương đồng giữa ảnh sinh ra và ảnh tham chiếu (image prompt).
- Đây là độ đo phản ánh mức độ bảo toàn phong cách, màu sắc hoặc bố cục từ ảnh đầu vào.

$$CLIP - I = \frac{1}{N \times K} \sum_{n=1}^{N} \sum_{k=1}^{K} \cos\left(\hat{v}_{n,k}^{gen}, \hat{v}_{n}^{ref}\right)$$

Trong đó:

+ N: Số mẫu gốc.
+ K: Số ảnh được sinh ra cho mỗi mẫu.
+ $\hat{v}_{n,k}^{gen}$: vector embedding CLIP của ảnh sinh ra thứ k ứng với mẫu n.
+ $\hat{v}_{n}^{ref}$: vector embedding CLIP của image prompt của mẫu n.

CHƯƠNG 5 KẾT QUẢ ĐÁNH GIÁ

5.1 Thiết lập thực nghiệm

- Các thí nghiệm trong báo cáo được tiến hành chủ yếu trên mô hình Stable Diffusion v1.5 (SD v1.5) , một mô hình khuếch tán phổ biến và ổn định.
- Các đánh giá định lượng (CLIP–T và CLIP–I) được thực hiện hoàn toàn trên nền SD v1.5 nhằm đảm bảo tính nhất quán giữa các phương pháp.
- Bên cạnh đó, một số thí nghiệm định tính cũng được triển khai trên nhiều mô hình sinh ảnh khác nhau để kiểm tra khả năng tổng quát hóa và tính linh hoạt của các adapter, bao gồm: Stable Diffusion v1.4, ReV Animated, Realistic Vision v4.0, và Anything v4.0.
- Cách bố trí thí nghiệm nói trên cho phép vừa đảm bảo đánh giá nghiêm ngặt trên một mô hình tiêu chuẩn (SD v1.5), vừa kiểm tra khả

năng mở rộng của phương pháp sang những mô hình sinh ảnh khác trong thực tế.

5.2 Kết quả

5.2.1 Kết quả định lượng

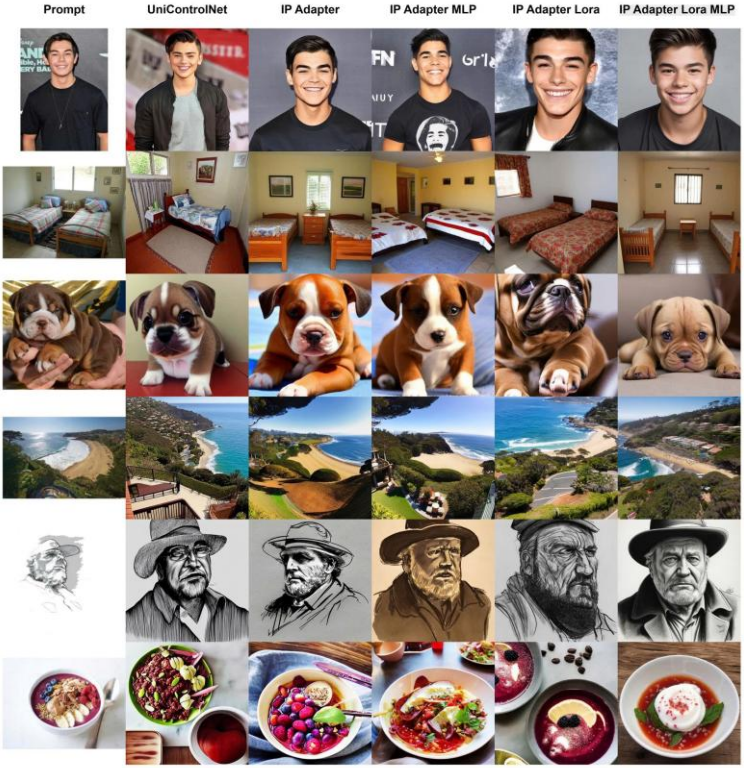
Method	Trainable parameters	CLIP – T	CLIP – I
UniControlNet (Global)	11.80 M	0.191	0.679
IP Adapter	20.75 M	0.198	0.709
IP Adapter MLP	30.71 M	0.184	0.665
IP Adapter LoRA	14.31 M	0.193	0.682
IP Adapter MLP_LoRA	24.27 M	0.172	0.625

***Bảng 1**: So sánh định lượng*

- Kết quả ở Bảng 1 cho thấy IP–Adapter gốc đạt hiệu năng cao nhất trên cả hai độ đo CLIP–T và CLIP–I, phản ánh khả năng duy trì đồng thời thông tin văn bản và đặc trưng hình ảnh tốt nhất. Tuy nhiên, chi phí tham số của phương pháp này vẫn ở mức tương đối cao.
- Ngược lại, UniControlNet (Global) mặc dù có số tham số huấn luyện thấp nhất trong tất cả các mô hình, nhưng vẫn duy trì được chất lượng sinh ảnh ở mức cạnh tranh, cho thấy cơ chế điều khiển toàn cục là đủ hiệu quả đối với các tác vụ dựa trên image prompt.
- Đáng chú ý, biến thể IP–Adapter LoRA cho thấy mức cân bằng tốt nhất giữa độ nhẹ và hiệu năng: số tham số giảm đáng kể so với bản gốc nhưng các độ đo CLIP chỉ suy giảm rất nhỏ. Điều này khẳng định hiệu quả của việc áp dụng LoRA lên các ma trận chiếu lớn trong cơ chế Decoupled Cross-Attention.
- Ngược lại, các biến thể tích hợp MLP (IP–Adapter MLP và MLP_LoRA) có xu hướng làm suy giảm hiệu năng, mặc dù số tham số tăng lên đáng kể. Điều này cho thấy việc thay thế linear projection bằng các tầng phi tuyến không mang lại lợi ích trong bối cảnh ánh xạ embedding CLIP sang không gian attention, thậm chí còn gây nhiễu và làm giảm khả năng bảo toàn thông tin hình ảnh.
- Tổng thể, các kết quả thực nghiệm khẳng định rằng: (i) IP–Adapter gốc vẫn là chuẩn mạnh nhất về chất lượng sinh ảnh; (ii) IP–Adapter LoRA là lựa chọn tối ưu khi cần giảm tham số mà vẫn giữ hiệu năng cao; (iii) UniControlNet(Global) là lựa chọn tốt khi cần mô hình nhẹ nhưng vẫn đảm bảo chất lượng sinh ảnh; và (iiii) các thiết kế dựa trên

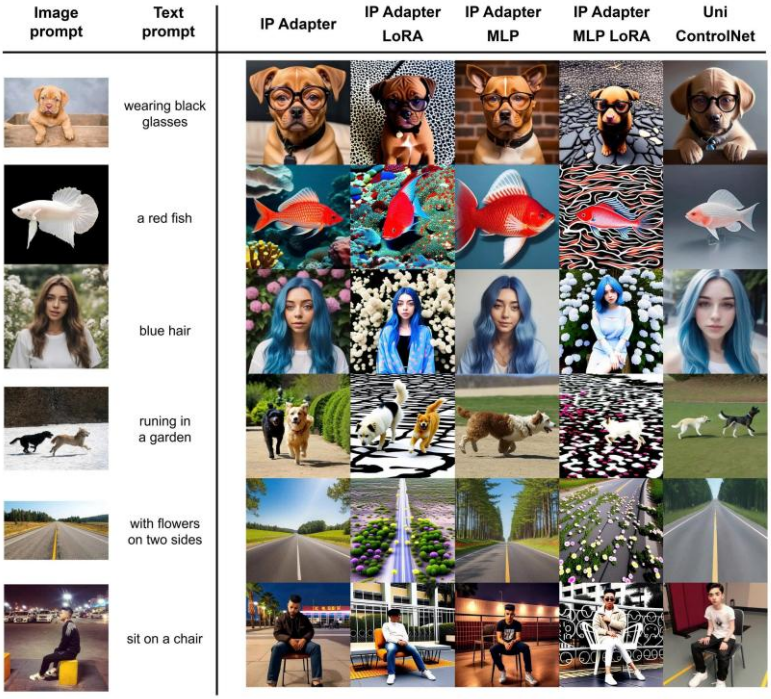
MLP không phù hợp với bài toán này do không cải thiện và thậm chí làm suy giảm chất lượng sinh ảnh.

5.2.2 Kết quả định tính



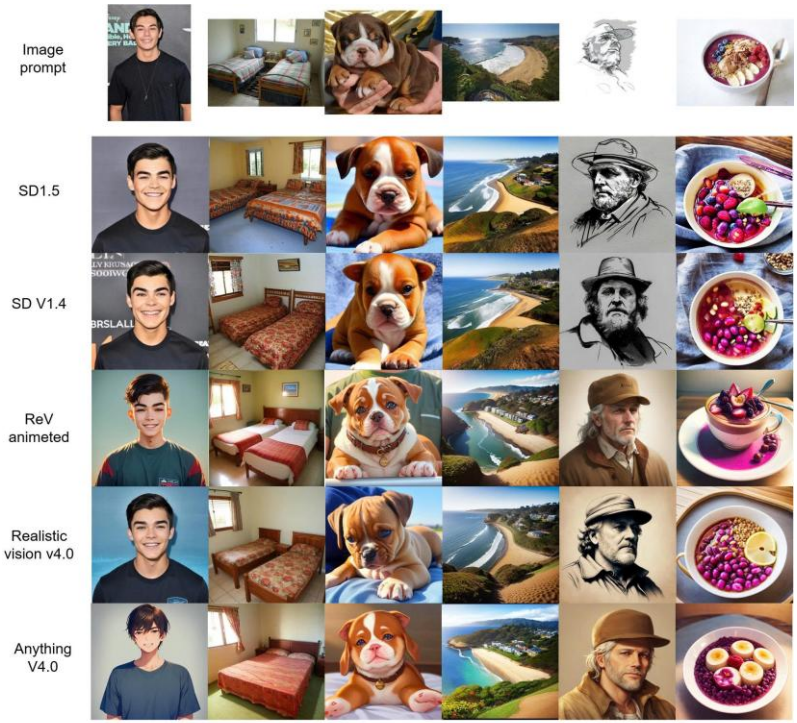
***Hình 8**: So sánh trực quan giữa các phương pháp khác dựa trên các loại và kiểu hình ảnh khác nhau.*

– Hình 8 cho thấy UniControlNet duy trì mức độ nhất quán cao hơn giữa ảnh điều kiện và ảnh sinh, đặc biệt ở các hàng chứa khuôn mặt, phòng và phong cảnh; trong khi các biến thể IP-Adapter có xu hướng tái tạo chi tiết sắc nét hơn, mức độ đa dạng hóa mạnh hơn, thể hiện qua sự thay đổi phông nền và tạo kiểu nghệ thuật.



***Hình 9**: So sánh trực quan giữa các cấu hình IP-Adapter với các phương pháp khác dựa trên các loại và kiểu hình ảnh khác nhau.*

- Hình 9 cho thấy UniControlNet duy trì sự cân bằng tốt nhất giữa việc bảo toàn thông tin hình ảnh gốc và tuân thủ ràng buộc văn bản.
- Các cấu hình IP-Adapter và IP-Adapter LoRA/MLP tuy thể hiện mức độ chỉnh sửa mạnh, nhưng thường xuyên làm biến dạng nội dung gốc hoặc tạo chi tiết thừa. Ngược lại, UniControlNet sinh ảnh ổn định hơn và giữ được cấu trúc đối tượng ban đầu.



Hình 10: *Hình ảnh được tạo ra từ các mô hình diffusion khác nhau với IP-Adapter (IP-Adapter chỉ huấn luyện một lần trên SDv1.5).*

- Hình 10 minh họa khả năng tổng quát hóa mạnh của IP-Adapter khi được áp dụng trực tiếp lên nhiều mô hình diffusion khác nhau, mặc dù chỉ được huấn luyện một lần trên SDv1.5. Các kết quả cho thấy IP-Adapter duy trì được đặc trưng cốt lõi của ảnh tham chiếu trên hầu hết các backbone, từ SDv1.4, ReV-animated, đến Realistic Vision và Anything V4.0.
- Đáng chú ý, khi chuyển sang các mô hình có phong cách khác biệt (ví dụ ReV-animated và Anything V4.0), IP-Adapter vẫn bảo toàn tốt danh tính đối tượng và bố cục chính, đồng thời thích nghi tự nhiên với phong cách riêng của từng mô hình. Điều này cho thấy IP-Adapter không phụ thuộc nặng vào kiến trúc diffusion cụ thể, mà có khả năng hoạt động ổn định và linh hoạt trên nhiều hệ sinh ảnh khác nhau.

CHƯƠNG 6 NHỮNG CẢI TIẾN MỚI SO VỚI BÀI TRÌNH BÀY TRƯỚC

Trong phiên bản báo cáo này, nhóm đã bổ sung và mở rộng nhiều nội dung quan trọng nhằm hoàn thiện phân tích, nâng cao tính trực quan và tăng giá trị thực tiễn so với bài trình bày trước. Các cải tiến bao gồm:

- Trình bày rõ ràng lý do hình thành các biến thể cải tiến của IP-Adapter, bao gồm IP-Adapter MLP, IP-Adapter LoRA và IP-Adapter MLP_LoRA, nhằm giải quyết các hạn chế về khả năng biểu diễn và số lượng tham số của phiên bản gốc.
- Nhóm đã thêm nhiều hình ảnh minh họa quan trọng:
    + Ảnh minh họa Input và Ouput.
    + Chỉ rõ các module sử dụng trong kiến trúc UniControlNet và các module Global/Local Adapter.
    + Sơ đồ IP-Adapter và các biến thể cải tiến như IP-Adapter MLP, IP-Adapter LoRA và IP-Adapter MLP_LoRA.
    + Làm nổi bật rõ ràng các module được sử dụng và các module được thay thế trong từng biến thể, giúp người đọc dễ dàng nhận biết sự khác biệt giữa các phương pháp.
- Bổ sung khuyến nghị chọn mô hình:
    + Cần chất lượng cao nhất: IP-Adapter
    + Cần mô hình nhẹ nhất: UniControlNet (Global)
    + Cần cân bằng tốt giữa chất lượng và số tham số: IP-Adapter LoRA.
    + Không khuyến nghị dùng: IP-Adapter MLP và IP-Adapter MLP_LoRA vì hiệu năng giảm.

Những bổ sung này giúp báo cáo trở nên trực quan hơn, dễ theo dõi hơn và phản ánh đầy đủ hơn quá trình nghiên cứu của nhóm. Việc đưa ra khuyến nghị lựa chọn mô hình cũng góp phần làm tăng tính ứng dụng thực tiễn của đề tài. Nhờ đó, phiên bản báo cáo hiện tại hoàn thiện hơn, rõ ràng hơn và hỗ trợ tốt hơn cho việc triển khai các Adapter trong sinh ảnh multimodal.

CHƯƠNG 7 KẾT LUẬN

Trong nghiên cứu này, nhóm tiến hành đánh giá nhiều biến thể của IP–Adapter và UniControlNet (Global) nhằm khảo sát khả năng tích hợp thông tin hình ảnh vào mô hình sinh ảnh dựa trên khuếch tán. Kết quả cho thấy IP–Adapter gốc vẫn là phương pháp hiệu quả nhất về chất lượng sinh ảnh, nhưng IP–Adapter LoRA nổi bật ở khả năng giảm tham số mạnh mẽ trong khi vẫn duy trì hiệu năng gần tương đương. UniControlNet (Global) cũng cho chất lượng cạnh tranh với chi phí tham số thấp.

Ngược lại, các biến thể dựa trên MLP không cho thấy hiệu quả và không phù hợp cho tác vụ ánh xạ embedding hình ảnh trong mô hình attention-based. Những quan sát này cho phép kết luận rằng việc giảm tham số thông qua LoRA là hướng tối ưu và bền vững cho các mô hình điều khiển bằng image prompt, trong khi các cải tiến dựa trên MLP cần được xem xét lại hoặc thiết kế lại để phù hợp với không gian embedding của CLIP.

Trong tương lai, việc mở rộng đánh giá lên các mô hình mới hơn (như SDXL) và thử nghiệm trên các domain đa dạng hơn sẽ giúp hoàn thiện hơn hiệu quả và khả năng tổng quát hóa của các adapter điều khiển hình ảnh.

CHƯƠNG 8 REFERENCE

1. Ye, H., Zhang, J., Liu, S., Han, X., & Yang, W. (2023). Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. ArXiv preprint arXiv:2308.06721 (IP-Adapter).
2. Zhao, S., Chen, D., Chen, Y.-C., Bao, J., Hao, S., Yuan, L., & Wong, K.-Y. K. (2023). Uni-ControlNet: All-in-one control to text-to-image diffusion models. Advances in Neural Information Processing Systems, 36, 11127–11150 (Uni-ControlNet).
3. Sharma, P., Ding, N., Goodman, S., & Soricut, R. (2018). Conceptual Captions: A cleaned, hypernymed, image alt-text

dataset for automatic image captioning. In Proceedings of ACL (CC3M).

4. Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718, 2021 (CLIPScore).