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

1. 23520326 – Đỗ Minh Dũng
2. 23521555 – Huỳnh Diên Thục
3. 23521688 – Đinh Trần Duy Trường
4. TS. Nguyễn Vinh Tiệp

Multimodal Prompting

1. Giới thiệu
2. Dataset
3. Phương pháp
4. Kết quả
5. Demo

Multimodal Prompting

1. Giới thiệu
2. Dataset
3. Phương pháp
4. Kết quả
5. Demo

GIỚI THIỆU

Bối cảnh

- Hiện nay, Các mô hình khuếch tán sinh ảnh từ văn bản đang rất mạnh, cho phép sinh ảnh chất lượng cao từ text prompt.
- Tuy nhiên, chỉ dùng text prompt thì khó diễn đạt chính xác ý người dùng, phải "prompt engineering" phức tạp.

Mục tiêu

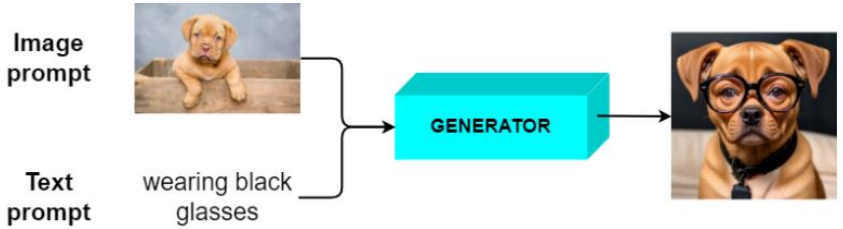
- Xây dựng một bộ điều hợp (adapter) nhẹ giúp các mô hình text-to-image đã huấn luyện nhận thêm image prompt mà không cần sửa hoặc fine-tune lại mô hình gốc.

Ý nghĩa

- Giúp tận dụng lại các mô hình text-to-image phổ biến, bổ sung image prompt mà vẫn giữ được text prompt và tương thích với các công cụ điều khiển khác (như ControlNet).
- Mở rộng mạnh khả năng ứng dụng trong sáng tạo nội dung, thiết kế, giải trí,…

Mô tả

- Input:
  + Reference image: ảnh dùng để điều khiển phong cách, bố cục, màu sắc, …
  + Text prompt: Chuỗi mô tả bằng ngôn ngữ tự nhiên.
- Output: Ảnh mới được sinh ra, bám nội dung/phong cách của ảnh tham chiếu và mô tả văn bản đi kèm.



Multimodal Prompting

1. Giới thiệu
2. Dataset
3. Phương pháp
4. Kết quả
5. Demo

DATASET

Dataset

- Conceptual Captions (CC3M) là tập dữ liệu ~3,3M ảnh kèm caption, được lấy từ alt-text HTML trên web và xử lý bằng pipeline tự động (trích xuất, lọc, biến đổi) để tạo caption vừa sạch, đủ thông tin, trôi chảy và phù hợp cho huấn luyện.

Dataset

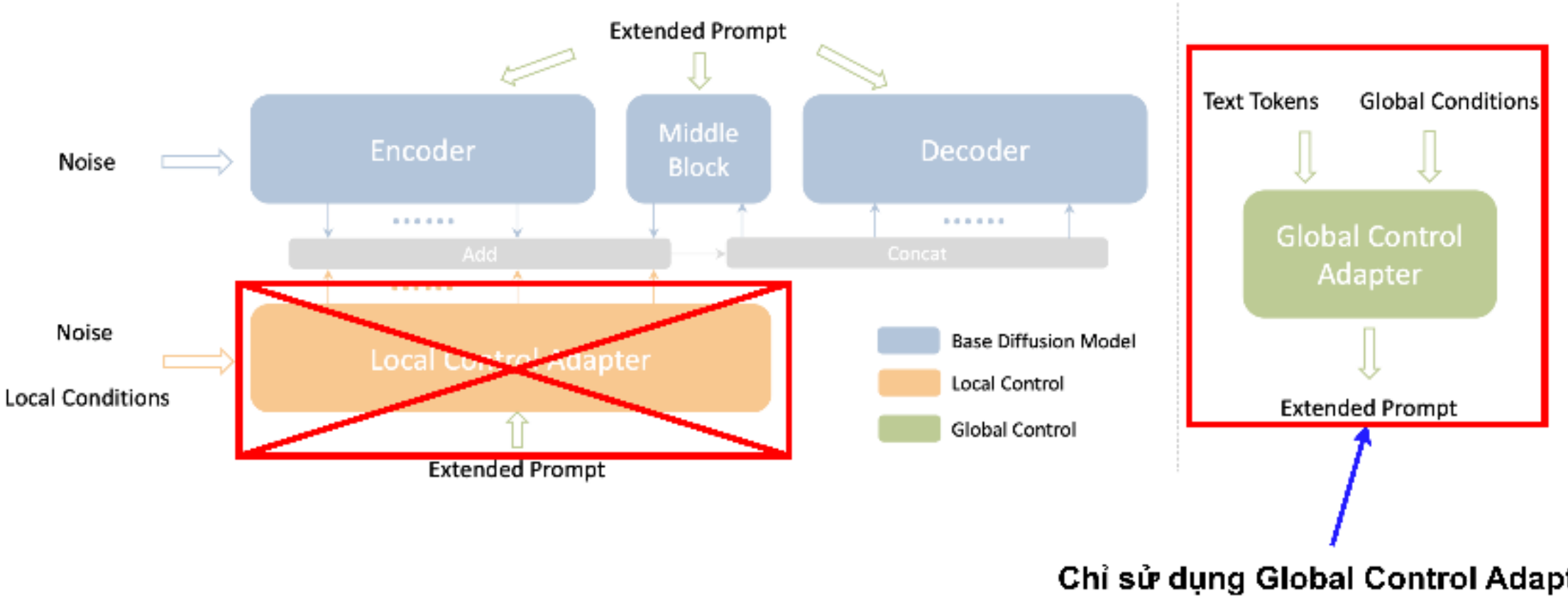
- Dữ liệu được đóng gói dạng webdataset .tar. Ảnh được tải bằng img2dataset, resize nếu cạnh ngắn nhất > 512 về 512.
  + Tập train: 576 shard, 2.905.954 / 3.318.333 mẫu.
  + Tập validation: 16 shard, 13.443 / 15.840 mẫu.

Multimodal Prompting

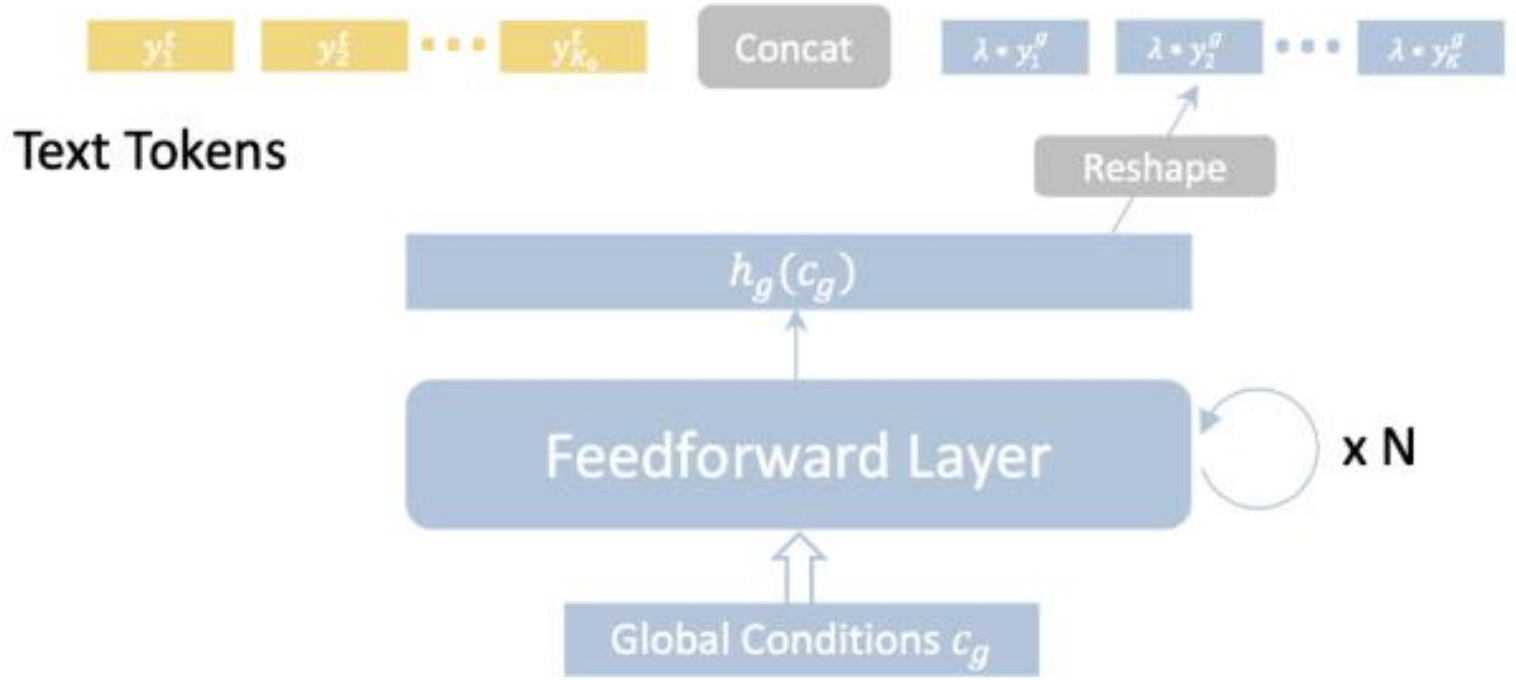
1. Giới thiệu
2. Dataset
3. Phương pháp
4. Kết quả
5. Demo

PHƯƠNG PHÁP

Uni–ControlNet



Uni–ControlNet



Global Control Adapter

Uni–ControlNet

– Global Control Adapter:

+ Tín hiệu global control 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

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

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

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

Uni–ControlNet

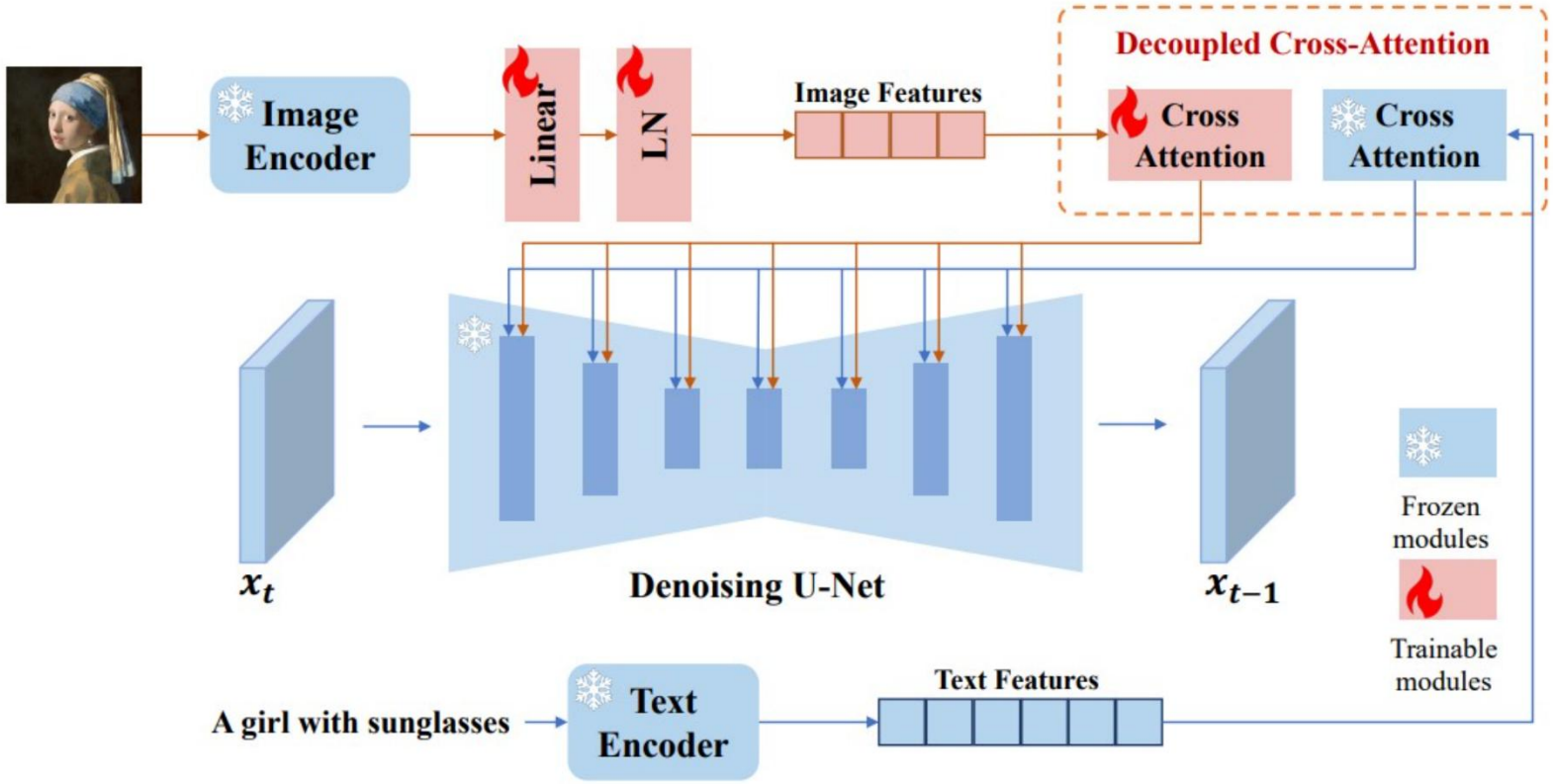
– Global Control Adapter:

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

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

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

IP–Adapter



Hình 3: Tổng quan Framework của IP-Adapter.

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

– Cơ chế Decoupled Cross-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.

IP–Adapter

- Có cách nào để làm giảm tham số mà không ảnh hưởng đến khả năng tạo sinh ảnh không?
- Tại sao lại sử dụng lớp Linear trong khi mạng Neural network cho phép học các ánh xạ phi tuyến tốt hơn?

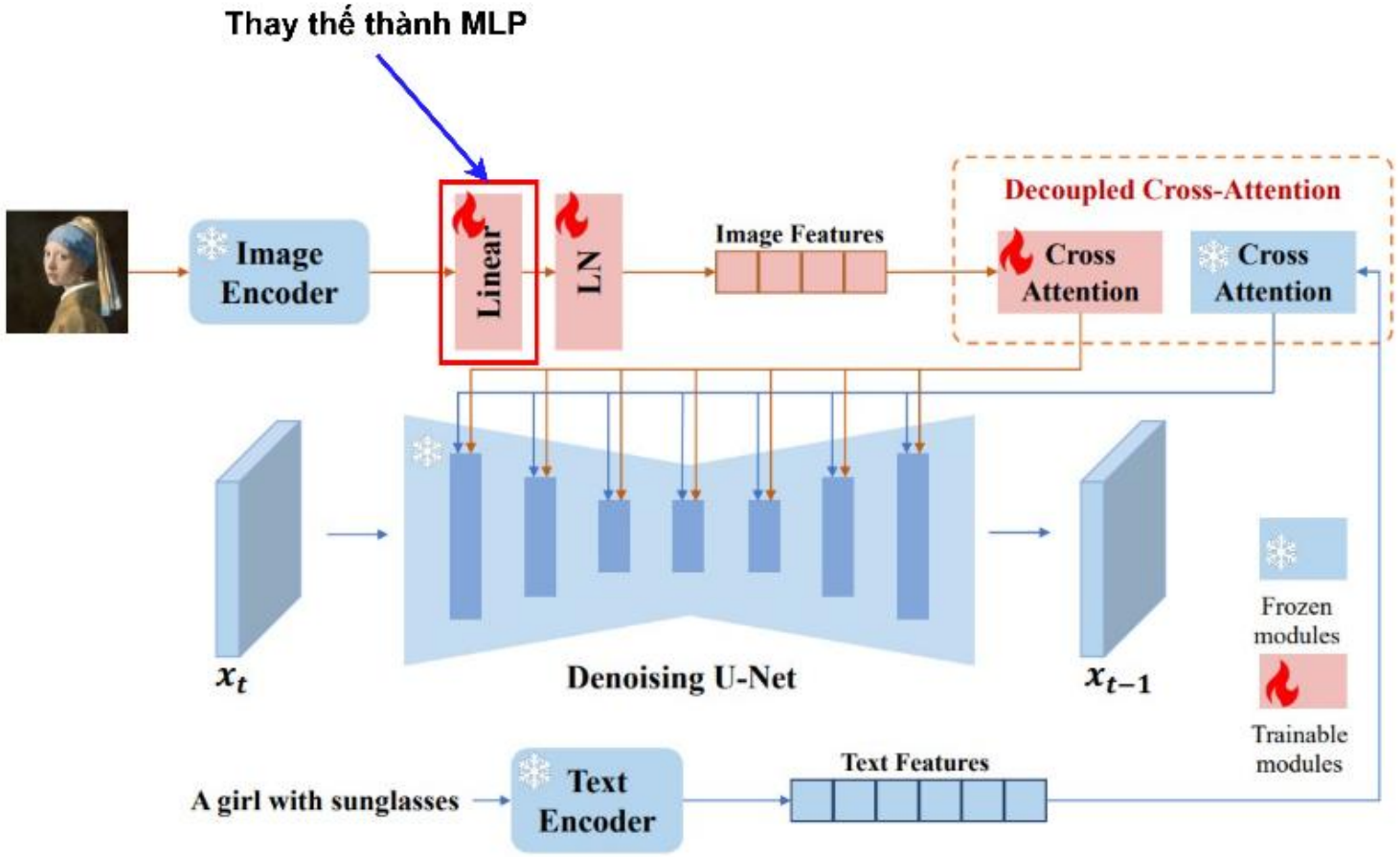
IP–Adapter

– Tiến hành thí nghiệm mở rộng trên mô hình IP–Adapter:

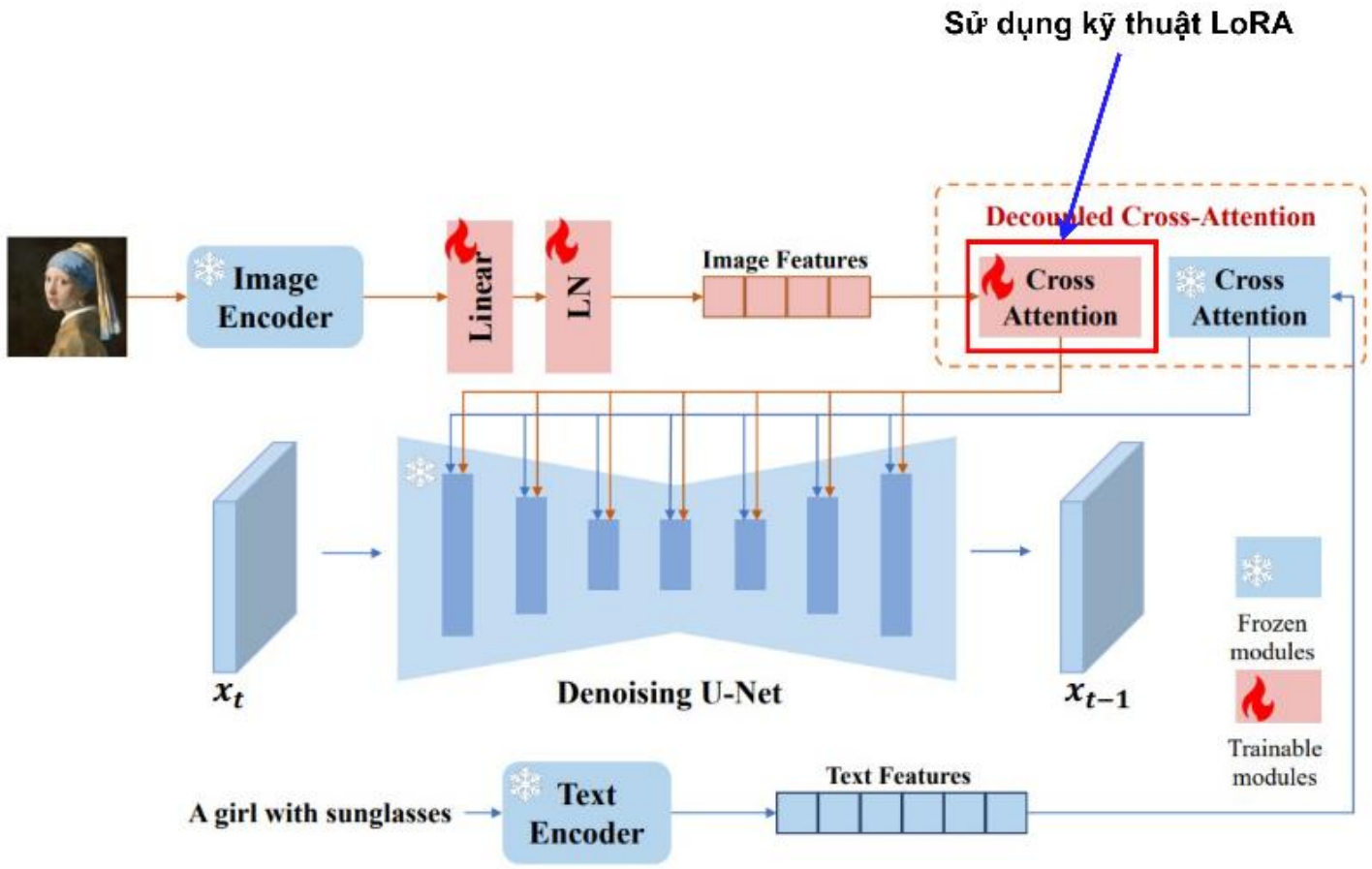
+ Thay thế lớp Linear thành Multi-Layer Perceptron 2 lớp ẩn.

+ Sử dụng kỹ thuật LoRA nhằm giảm số lượng tham số phải học.

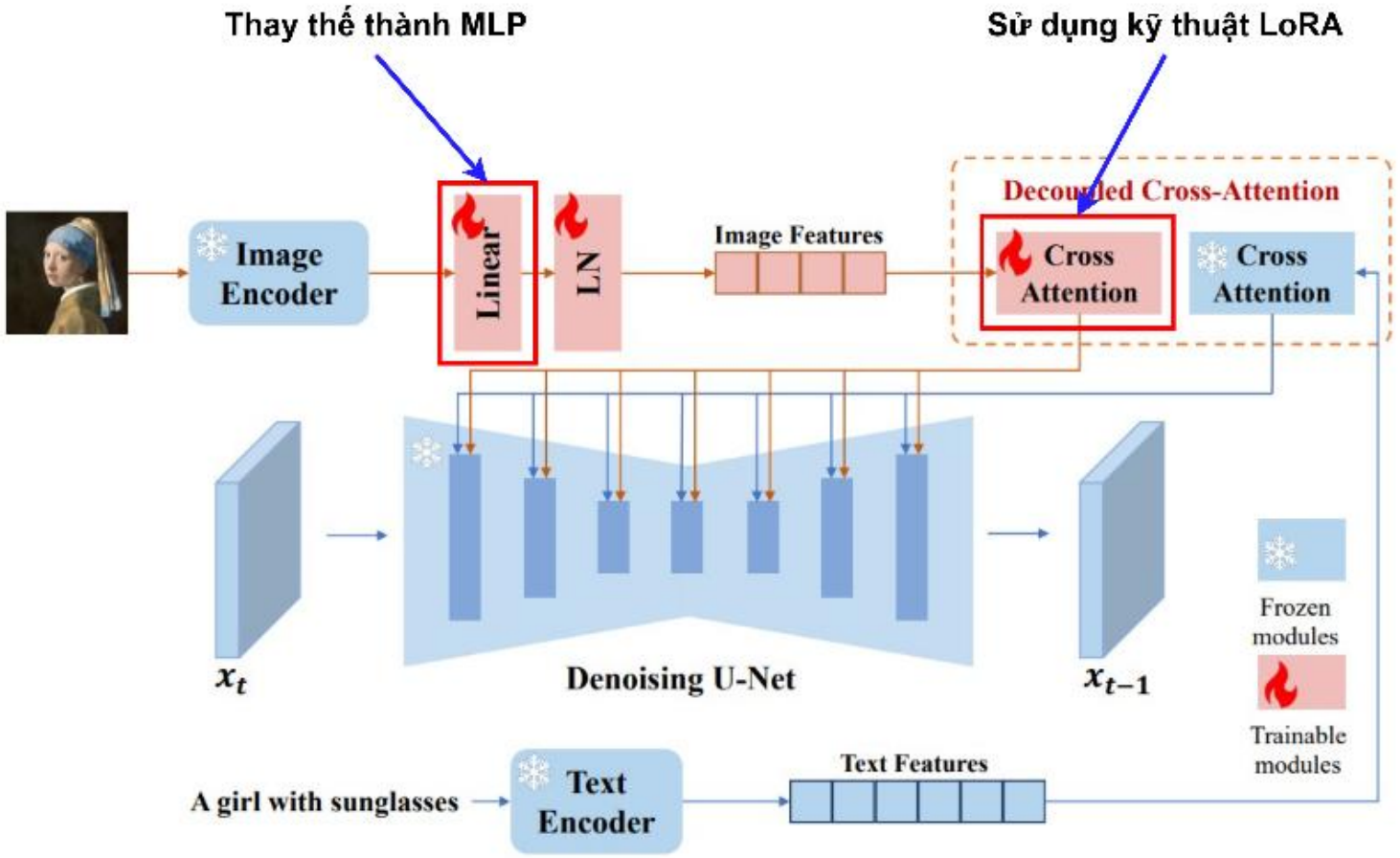
IP–Adapter


Thay thế thành MLP
Image Encoder
Linear
LN
Image Features
Decoupled Cross-Attention
Cross Attention
Cross Attention
x_t
Denoising U-Net
x_{t-1}
Frozen modules
Trainable modules
A girl with sunglasses
Text Encoder
Text Features

IP–Adapter


Sử dụng kỹ thuật LoRA
Decoupled Cross-Attention
Image Encoder
Linear
LN
Image Features
Cross Attention
Cross Attention
x_t
Denoising U-Net
x_{t-1}
Frozen modules
Trainable modules
A girl with sunglasses
Text Encoder
Text Features

IP–Adapter


Thay thế thành MLP
Sử dụng kỹ thuật LoRA
Image Encoder
Linear
LN
Image Features
Decoupled Cross-Attention
Cross Attention
Cross Attention
x_t
Denoising U-Net
x_{t-1}
Frozen modules
Trainable modules
A girl with sunglasses
Text Encoder
Text Features

Multimodal Prompting

1. Giới thiệu
2. Dataset
3. Phương pháp
4. Kết quả
5. Demo

KẾT QUẢ

Độ Đo

– CLIP–T: điểm CLIPScore giữa ảnh được sinh ra và phần mô tả của ảnh prompt.

$$CLIP - T_{mean} = \frac{1}{N * K} \sum_{n=1}^{N} \sum_{k=1}^{K} \cos(\hat{v}_{n,k}^{gen}, \hat{v}_{n}^{text})$$

Độ Đo

– CLIP–I: độ tương đồng trong không gian embedding hình ảnh của CLIP giữa ảnh được sinh ra và ảnh dùng làm prompt.

$$CLIP - I_{mean} = \frac{1}{N * K} \sum_{n=1}^{N} \sum_{k=1}^{K} \cos(\hat{v}_{n,k}^{gen}, \hat{v}_{n}^{ref})$$

Thiết lập thực nghiệm

- Thực nghiệm được tiến hành trên mô hình SDv1.5
- Ngoài ra một số thí nghiệm định tính thực hiện trên SDv1.4, ReV animeted, Realistic vision v4.0, Anything v4.0

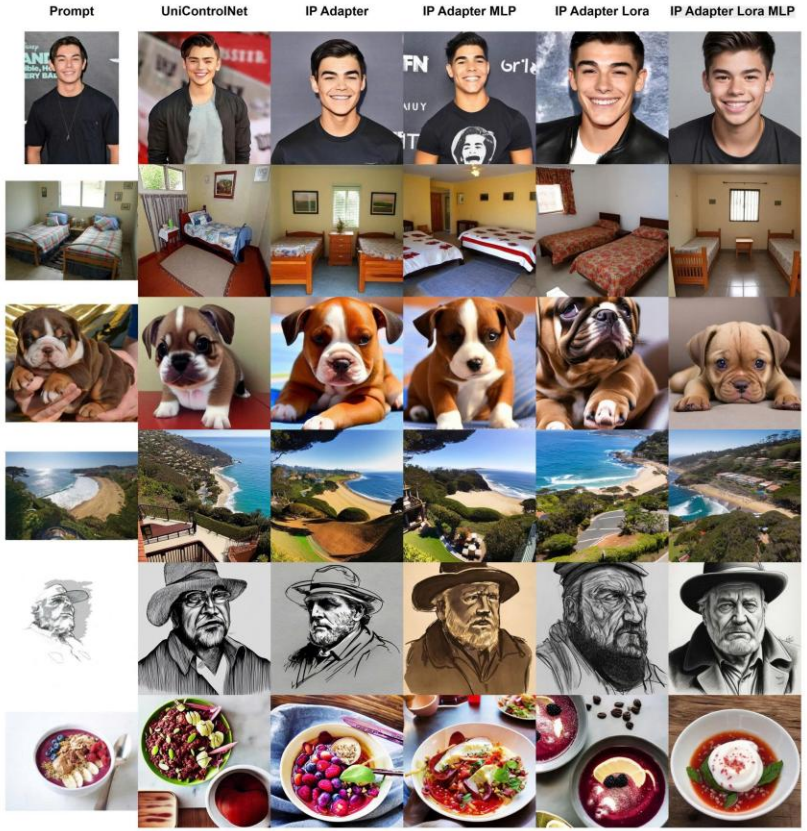
Kết quả định lượng

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

Kết quả định lượng

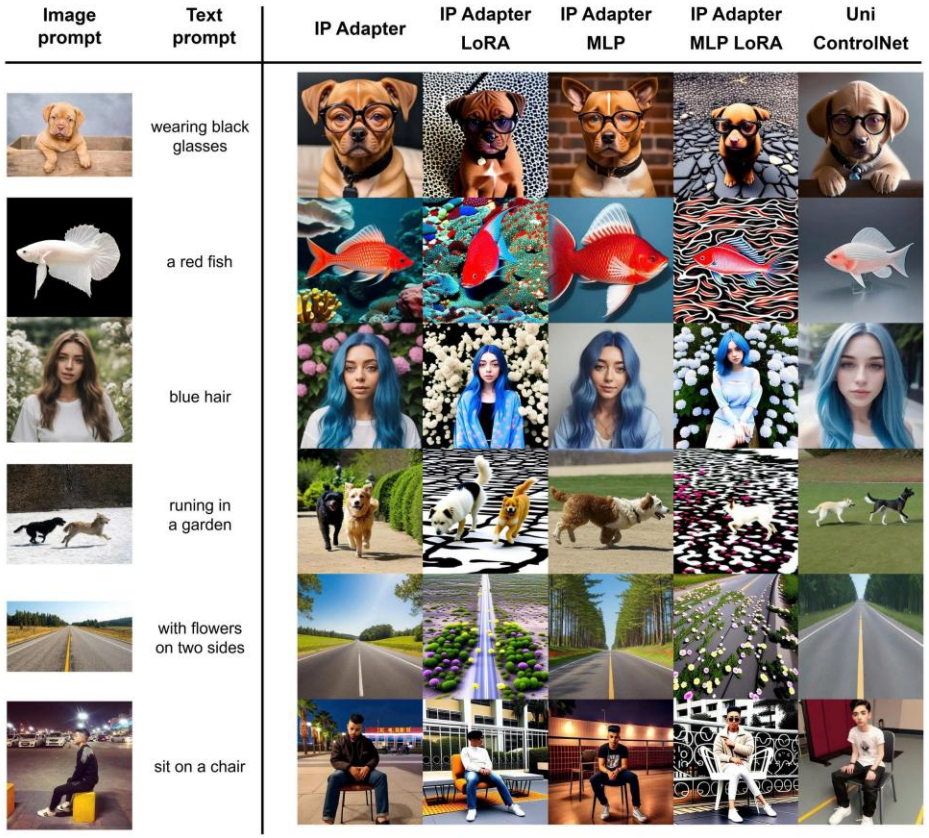
- Kết quả cho thấy IP–Adapter gốc đạt hiệu năng cao nhất trên cả hai độ đo CLIP–T và CLIP–I, phản ánh khả năng duy trì đồng thời thông tin văn bản và đặc trưng hình ảnh tốt nhất.
- Biến thể IP–Adapter LoRA cho thấy mức cân bằng tốt nhất giữa độ nhẹ và hiệu năng: số tham số giảm đáng kể so với bản gốc nhưng các độ đo CLIP chỉ suy giảm rất nhỏ.
- Ngược lại, các biến thể tích hợp MLP (IP–Adapter MLP và MLP_LoRA) có xu hướng làm suy giảm hiệu năng

Kết quả định tính



Hình 4: 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

Kết quả định tính



Hình 5: So sánh các multimodal prompts giữa các cấu hình IP-Adapter với các phương pháp khác.

Kết quả định tính

Hình 6: Hình ảnh được tạo ra từ các mô hình diffusion khác nhau với IP - Adapter. IP-Adapter chỉ huấn luyện một lần

Multimodal Prompting

1. Giới thiệu
2. Dataset
3. Phương pháp
4. Kết quả
5. Demo

DEMO

Demo

- Link project: here.

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).

Reference

1. 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.

Cảm ơn các bạn đã lắng nghe

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN TP.HCM