



IMAGE COMPLETION DDPM vs. Stable Diffusion

Thành viên:

Nguyễn Thị Thanh Huyền - 23020381
Nguyễn Thị Minh Ly - 23020399
Đặng Minh Nguyệt - 23020407

Giảng viên hướng dẫn:

PGS.TS. Nguyễn Việt Hà
ThS. Nguyễn Thị Thùy Linh

NỘI DUNG TRÌNH BÀY

01

Đặt vấn đề

02

Cơ sở lý thuyết

03

Thực nghiệm

04

Kết luận

ĐẶT VẤN ĐỀ

01

Bài toán Image Completion

Định nghĩa: Khôi phục các vùng bị khuyết trong ảnh đầu vào

- Nhất quán cấu trúc
 - Tự nhiên về thị giác
 - Phù hợp ngữ nghĩa toàn cục
- Khác với khử nhiễu / siêu phân giải:
 - Không chỉ dựa vào thông tin cục bộ
 - Cần suy luận ngữ cảnh toàn ảnh

02

Ứng dụng

- Phục hồi ảnh cũ, ảnh bị hư hỏng
- Chỉnh sửa & hậu kỳ ảnh
- Xóa vật thể không mong muốn
- Hỗ trợ sáng tạo nội dung số (design, nghệ thuật)
- Diffusion models → nâng cao chất lượng & tính thẩm mỹ ảnh sinh

ĐẶT VÂN ĐỀ PHẠM VI DỰ ÁN

1

DDPM

Nền tảng cho các mô hình diffusion hiện đại, là mô hình baseline.

2

Stable Diffusion v1.5 + LoRA

- Latent Diffusion Model (LDM) - giới thiệu năm 2022
- Mô hình phổ biến, đại diện cao trong cộng đồng
- Chất lượng sinh ảnh & khả năng nắm bắt ngữ nghĩa tốt
- Cho phép fine-tune theo tác vụ

CÔNG TRÌNH NGHIÊN CỨU LIÊN QUAN

Trước Diffusion

Phương pháp chính:

- CNN, GAN (Context Encoder, Conditional GAN)
- Attention, multi-scale, structural guidance

Hạn chế:

- Khó xử lý vùng khuyết lớn
- Thiếu nhất quán ngữ nghĩa
- Kết quả ít đa dạng, phụ thuộc discriminator

Với Diffusion

Sinh ảnh qua quá trình khử nhiễu dần

Inpainting có điều kiện: Vùng không khuyết làm ràng buộc

Ưu điểm:

- Chi tiết sắc nét
- Nhất quán ngữ nghĩa toàn cục
- Sinh được nhiều nghiệm hợp lý

→ *Stable Diffusion mở ra hướng tiếp cận hiệu quả cho chỉnh sửa ảnh*

CƠ SỞ LÝ THUYẾT

DIFFUSION MODELS

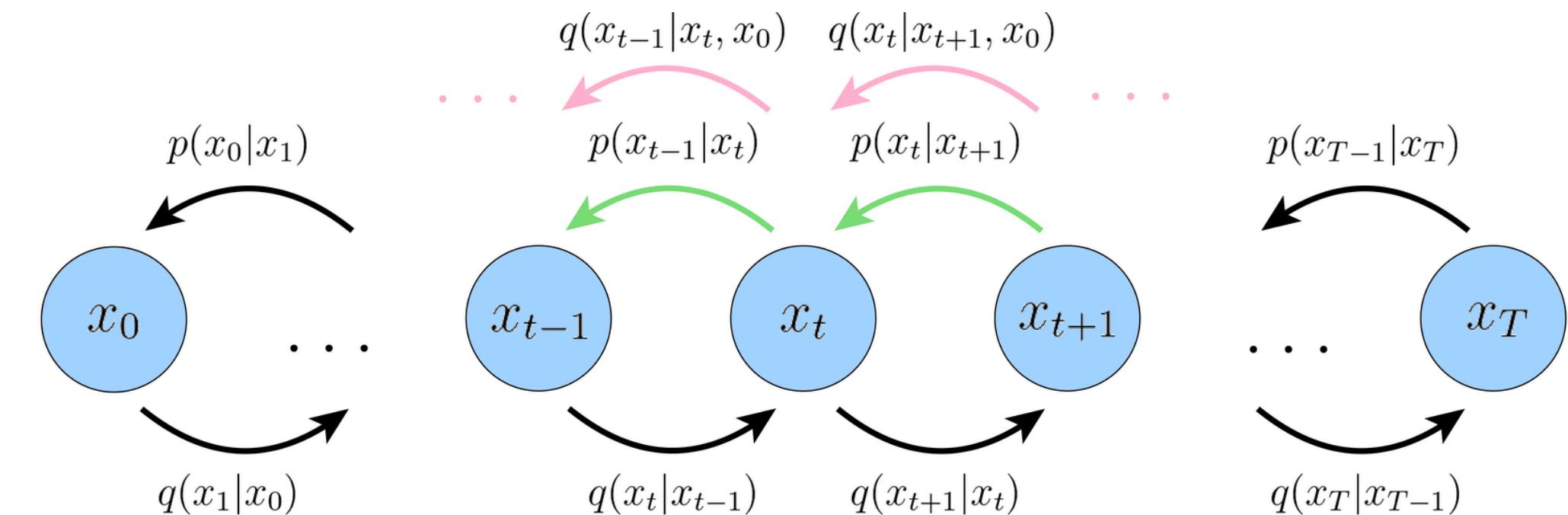
Mô hình sinh xác suất dựa trên chuỗi Markov, gồm hai quá trình:

- **Forward process:** thêm nhiễu Gaussian dần vào dữ liệu
- **Reverse process:** học cách khử nhiễu để tái tạo dữ liệu

Mục tiêu: Học phân phối dữ liệu thông qua việc đảo ngược quá trình nhiễu hóa

Huấn luyện:

- Tối ưu cận dưới biến phân (VLB) của log-likelihood
- Mô hình hóa phân phối khử nhiễu bằng mạng nơ-ron sâu



Khuếch tán tiến

$$q(x_t | x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t \mathbf{I})$$

Khuếch tán ngược

$$p_\theta(x_{t-1} | x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

Variational loss

$$\mathcal{L} = \mathbb{E}_q \left[-\log p(x_T) - \sum_{t=1}^T \log \frac{p_\theta(x_{t-1} | x_t)}{q(x_t | x_{t-1})} \right].$$

DENOISING DIFFUSION PROBABILISTIC MODELS (DDPM)

Đặc điểm chính:

- Forward process có dạng đóng
- Có thể lấy mẫu trực tiếp x_t từ x_0

$$q(x_t | x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)\mathbf{I}).$$

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}).$$

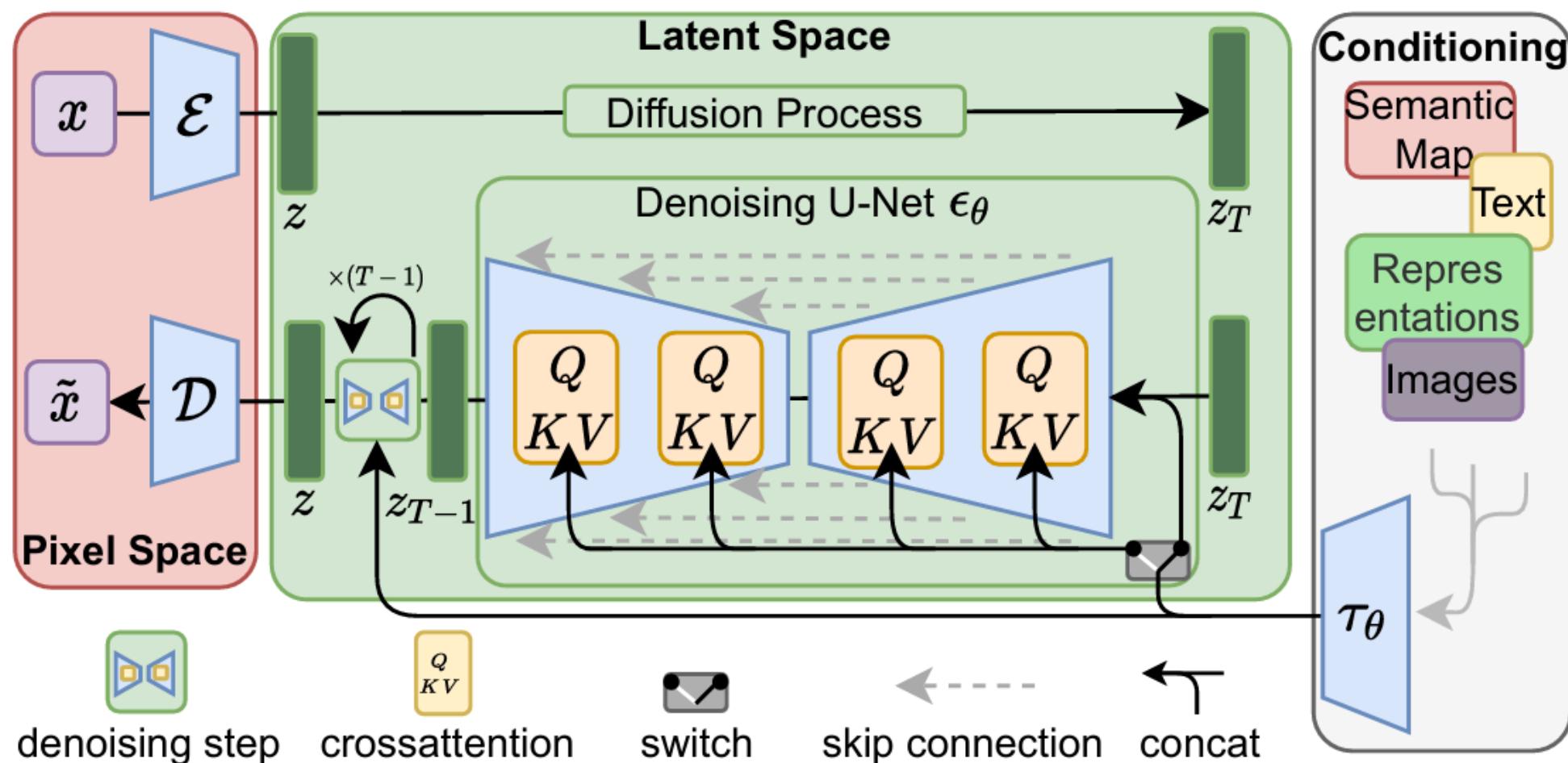
Giả định Reverse process là Gaussian với phương sai cố định

$$p_\theta(x_{t-1} | x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \beta_t\mathbf{I}).$$

Chiến lược huấn luyện: Dự đoán nhiễu ϵ thay vì dự đoán trực tiếp ảnh. Khi đó, hàm mất mát:

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{x_0, \epsilon, t} [\|\epsilon - \epsilon_\theta(x_t, t)\|^2]$$

STABLE DIFFUSION V1.5



Thuộc nhóm Latent Diffusion Models (LDM)

Ý tưởng chính: Thực hiện diffusion trong không gian tiềm ẩn thay vì pixel

Thành phần:

- VAE: mã hóa ảnh → latent, giải mã latent → ảnh
- U-Net: học quá trình khử nhiễu trong latent space

Phiên bản 1.5:

- Sử dụng CLIP Text Encoder
- Kết hợp vào U-Net qua cross-attention

Kết quả: Cân bằng tốt giữa chất lượng ảnh, điều khiển ngữ nghĩa và hiệu năng

THỰC NGHIỆM

METRICS ĐÁNH GIÁ

01

PSNR: Đo mức độ khác biệt pixel-wise giữa ảnh sinh và ảnh gốc; giá trị càng cao → ảnh được khôi phục càng gần với ảnh thật về mặt pixel

02

SSIM: Đánh giá mức độ tương đồng về cấu trúc, độ tương phản và độ sáng; phản ánh khả năng bảo toàn cấu trúc không gian của ảnh sau khi completion

03

LPIPS: Đo độ khác biệt cảm nhận thị giác dựa trên đặc trưng học được từ mạng sâu; giá trị càng thấp thì ảnh sinh càng giống ảnh thật theo cảm nhận của con người

04

FID: Đánh giá mức độ tương đồng phân phối giữa tập ảnh sinh và ảnh thật trong không gian đặc trưng; giá trị thấp → chất lượng và tính đa dạng của ảnh sinh tốt hơn

DDPM

- **Tiền xử lý dữ liệu đầu vào:** các ảnh đầu vào được thay đổi kích thước độ phân giải về 256×256 . Dữ liệu ảnh (x_{gt} và x_{masked}) được chuẩn hóa về khoảng giá trị $[-1, 1]$
- **Chi tiết thực thi và Kiến trúc mô hình**
 - Dựa trên kiến trúc U-Net của DDPM
 - **Điều chỉnh đầu vào mô hình:** mở rộng lớp tích chập đầu tiên để chấp nhận đầu vào 7 kênh, bao gồm: $\text{input} = \text{Concat}(x_{masked}, \text{mask}, x_t)$
 - Trong đó x_{masked} là ảnh đầu vào đã bị che một phần (3 kênh), mask là mặt nạ (1 kênh), và x_t là trạng thái nhiễu tại bước thời gian t (3 kênh).
 - **Hyperparameters:**
 - Resolution: 256×256 .
 - Batch Size: 8.
 - Learning Rate: 1×10^{-5} , sử dụng bộ tối ưu hóa AdamW.
 - Epochs: 10.
 - Gradient Clipping: Norm tối đa là 1.0 để tránh bùng nổ gradient.
 - Mixed Precision: Sử dụng Automatic Mixed Precision (AMP) với `torch.amp` để giảm thiểu bộ nhớ VRAM và tăng tốc độ huấn luyện.

DDPM

Training Objective & Spatially Weighted Loss

- Thay đổi hàm mục tiêu MSE tiêu chuẩn của DDPM bằng biến thể có trọng số theo không gian, giải quyết sự mất cân bằng thông tin giữa vùng cần khôi phục và vùng nền đã biết

$$L_{total} = \mathbb{E}_{t, x_0, \epsilon} [\text{mean} (\|\epsilon - \epsilon\theta(x_t, t, x_{masked}, mask)\|^2 \odot W)]$$

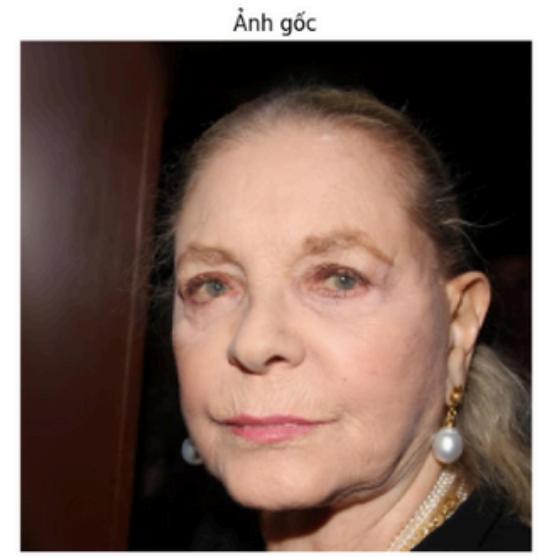
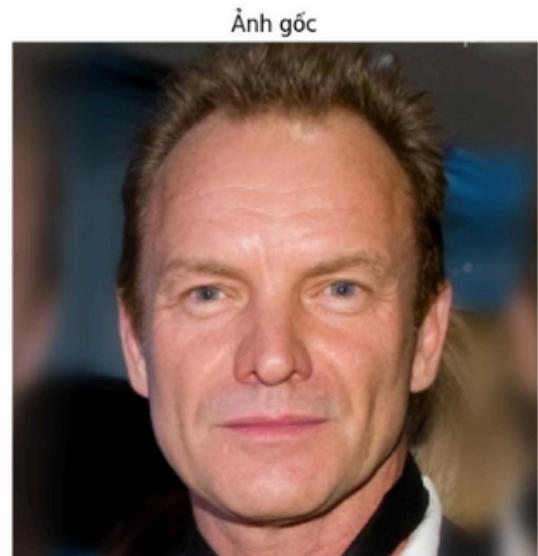
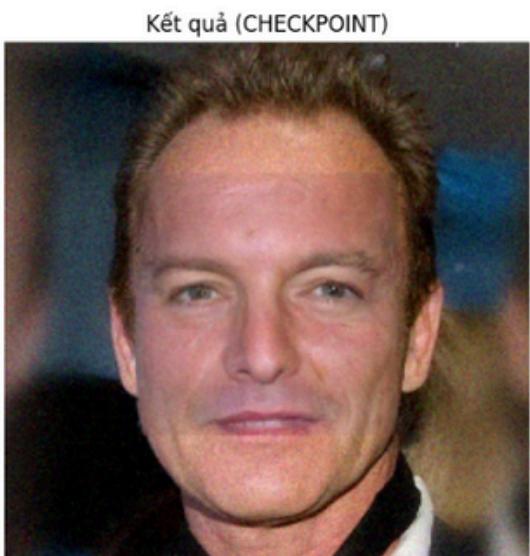
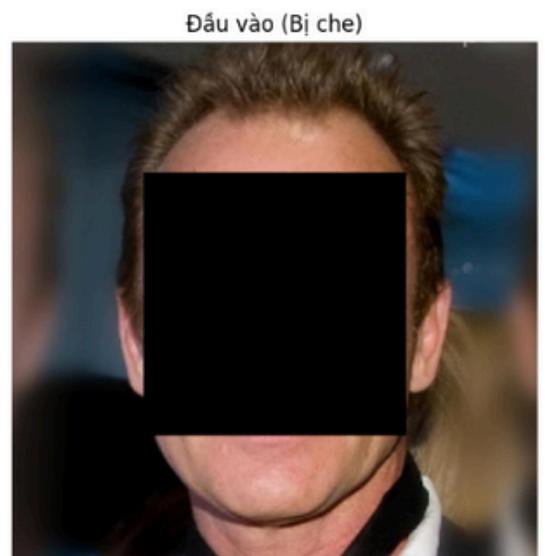
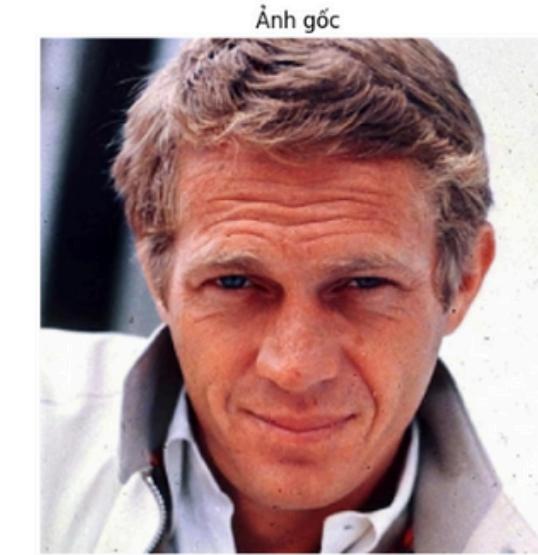
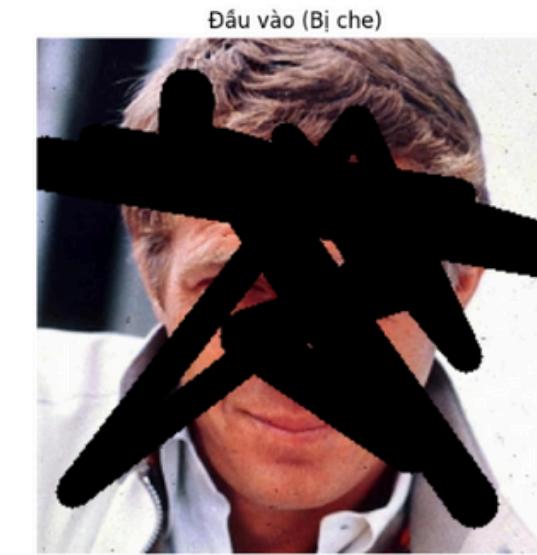
- Trong đó:
 - ϵ là nhiễu Gaussian thực tế được thêm vào ảnh.
 - $\epsilon\theta$ là nhiễu được dự đoán bởi mạng U-Net.
 - W: Ma trận trọng số không gian, có cùng kích thước với ảnh đầu vào.
- Ma trận W được xây dựng dựa trên mặt nạ nhị phân mask, điều hướng gradient tập trung vào khu vực bị mất thông tin.**

$$W_{i,j} = mask_{i,j} \cdot \lambda_{mask} + (1 - mask_{i,j}) \cdot \lambda_{context}$$

- $\lambda_{mask} = 1$
- $\lambda_{context} = 0.05$

KẾT QUẢ

Thiết lập: Sử dụng DDIM scheduler với 50 bước khuếch tán ngược giảm thời gian sinh ảnh so với DDPM gốc



KẾT QUẢ

FINETUNE DDPM

PSNR	20.9651 dB	Chưa tái tạo chính xác màu sắc và chi tiết cục bộ ở vùng bị che
SSIM	0.8012	Giữ được cấu trúc tổng thể và hình dạng chính
LPIPS	0.1625	Anh sinh ra khá giống ảnh gốc về mặt cảm nhận
FID	40.4452	Phân phối ảnh sinh còn khác biệt so với ảnh thật, màu sắc chưa đồng đều

DIFFUSION 1.5

+ LORA

CẤU HÌNH

STABLE DIFFUSION 1.5 + LORA

Mô hình cơ sở: Stable Diffusion 1.5

- Thành phần chính: VAE, UNet, Text Encoder
(đóng băng để giảm chi phí huấn luyện)

LoRA trên UNet:

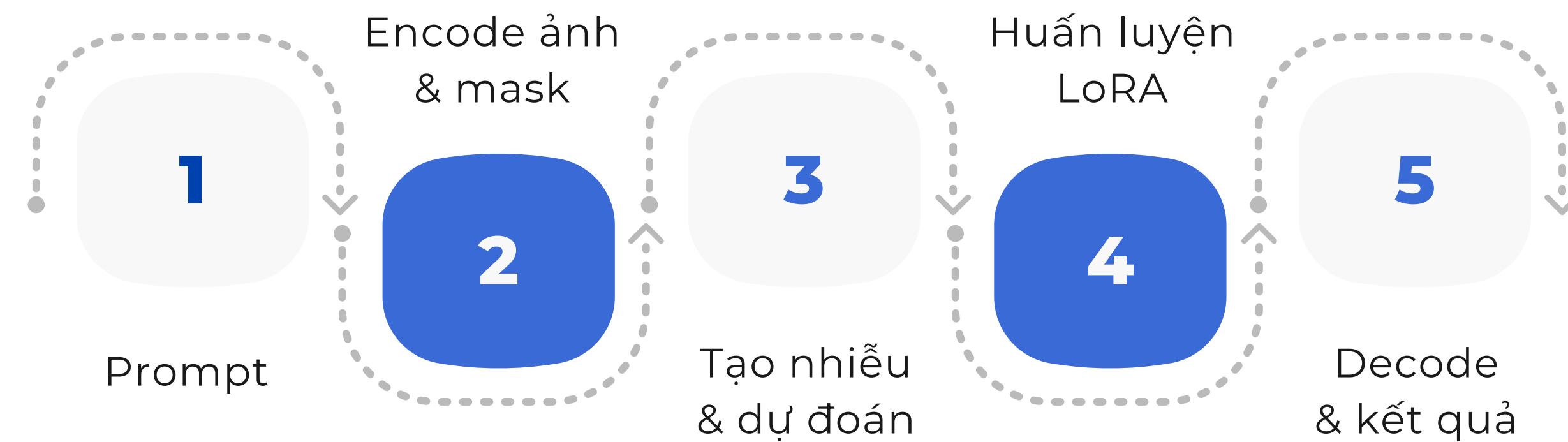
- Rank (r): 16 | Alpha: 16 | Dropout: 0.05
- Nhắm vào các module: to_q, to_k, to_v, to_out.0

Cấu hình huấn luyện:

- Batch size = 2 | Learning rate = 1×10^{-4} | Epoch = 3
- Mixed precision FP16 + GradScaler → giảm bộ nhớ GPU & tăng tốc

PIPELINE

STABLE DIFFUSION 1.5 + LORA



PIPELINE

1. Prompt

- Caption từ BLIP, ví dụ: “a woman with long brown hair”
- Tokenize & encode bằng Text Encoder (đóng băng để giảm chi phí huấn luyện)

2. Encode ảnh & mask

- Ảnh gốc & masked → encode VAE → latent 64×64
- Mask resize 64×64 , đảo màu (vùng trắng = vùng inpaint)

3. Tạo nhiễu & dự đoán

- Sinh Gaussian noise, thêm vào latent theo SD timestep scheduler
- Latent bị nhiễu + masked latent + mask → UNet + LoRA → dự đoán noise

4. Huấn luyện LoRA

- MSE loss giữa noise dự đoán & noise gốc
- Backpropagation chỉ cập nhật LoRA parameters
- CFG training: bỏ prompt 15% xác suất → cải thiện khả năng điều khiển

5. Decode & kết quả

- Latent dự đoán → decode VAE → ảnh RGB hoàn chỉnh

TIỀN XỬ LÝ DỮ LIỆU

STABLE DIFFUSION 1.5 + LORA

1. Chuẩn hóa ảnh & mask

- 512×512, RGB, [-1,1]
- Mask 64×64, đảo màu → vùng cần vẽ = trắng

2. Caption / Prompt

- Sinh tự động bằng BLIP
- Lưu offline vào JSON.

3. Xử lý dataset thiếu ảnh

- Kiểm tra tồn tại file GT/mask
- Nếu thiếu, dùng ảnh đầu tiên trong dataset để thay thế

4. CFG training

- Bỏ prompt 15% xác suất → cải thiện điều khiển khi inference

KẾT QUẢ

STABLE DIFFUSION 1.5 + LORA

Thiết lập:

- 50 steps, guidance scale = 7.5, seed cố định
- Mask đảo màu → vùng cần vẽ = trắng

Input (Masked) - ID 479



Mask Input (White=Draw)



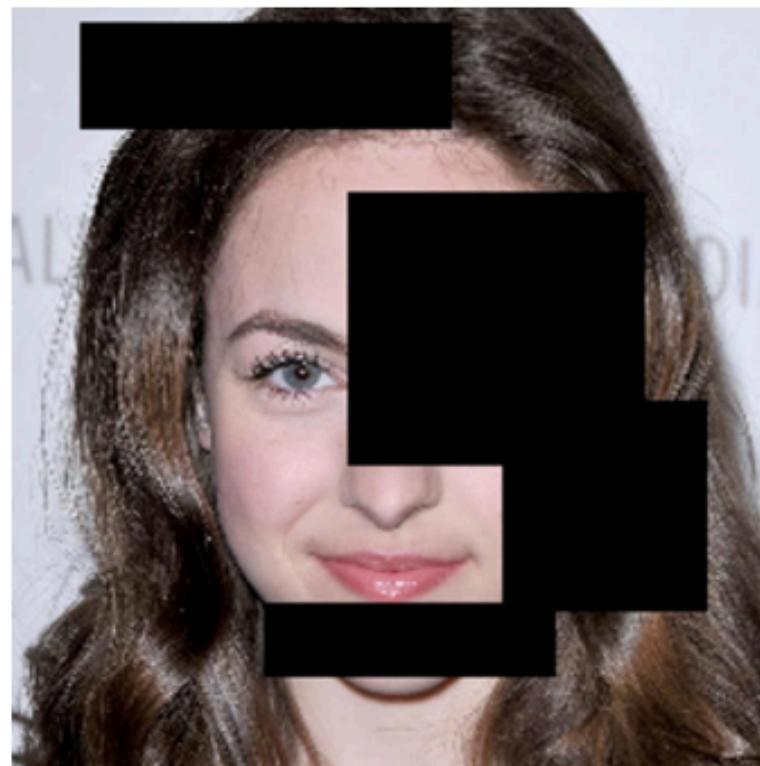
Prediction (AI)



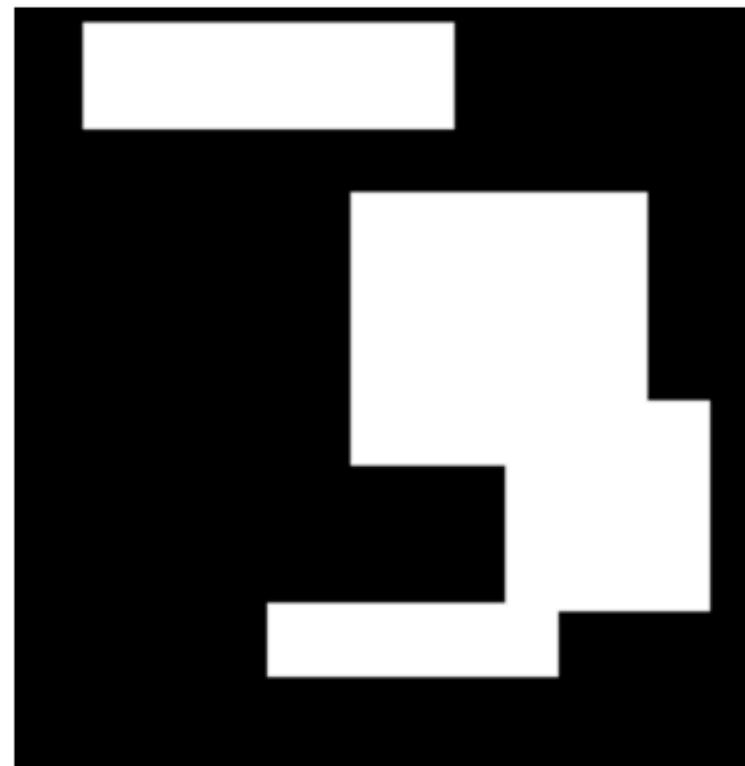
Ground Truth (Real)



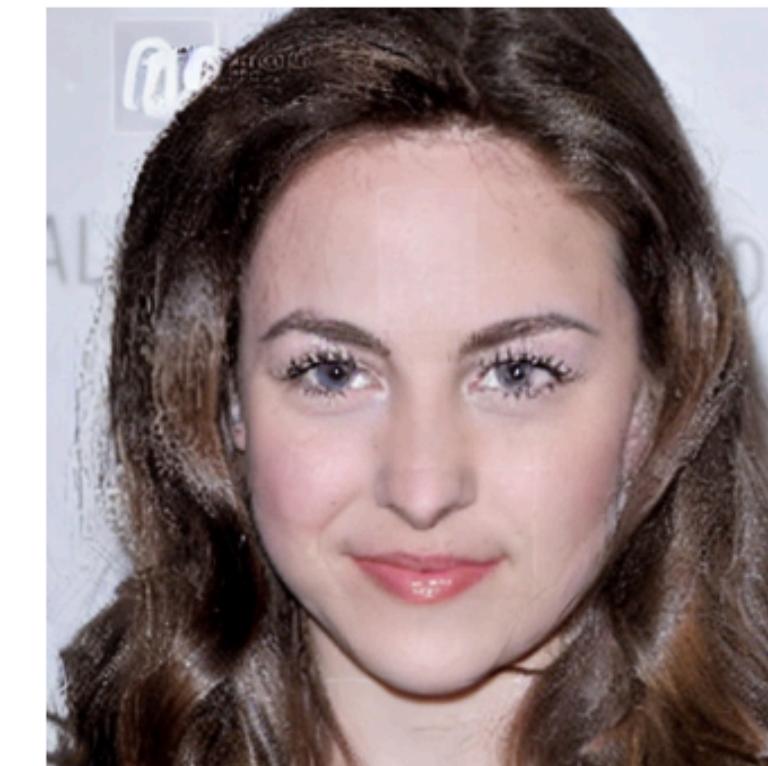
Input (Masked) - ID 311



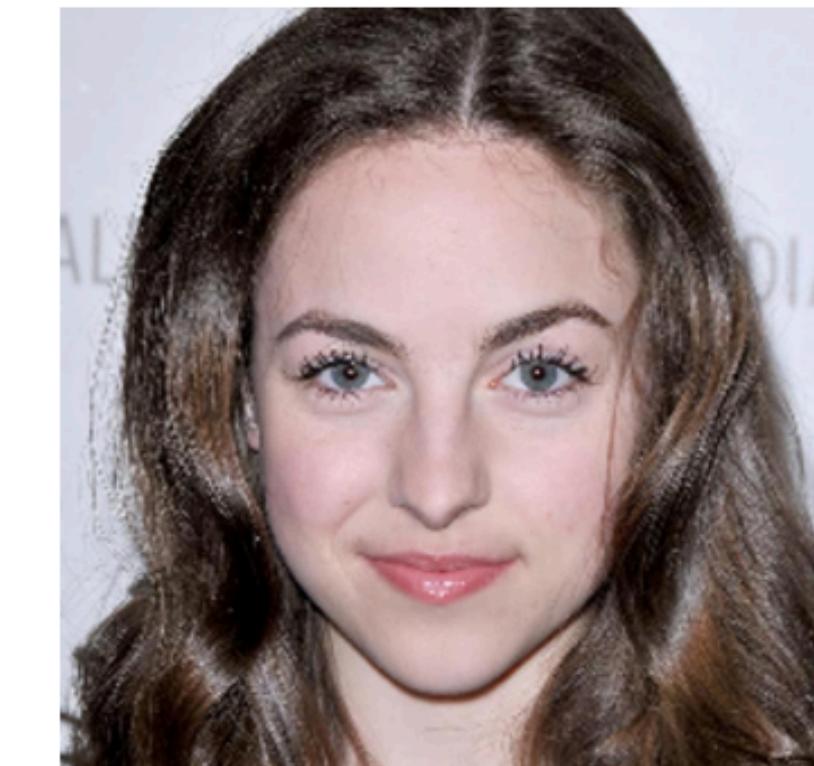
Mask Input (White=Draw)



Prediction (AI)



Ground Truth (Real)



PHÂN TÍCH KẾT QUẢ

STABLE DIFFUSION 1.5 + LORA

PSNR	23.45 dB	Khôi phục pixel chính xác
SSIM	0.845	Bảo toàn kết cấu tổng thể
LPIPS	0.112	Perceptual similarity tốt
FID	32.17	Phân phối ảnh sinh gần với ảnh thật

SO SÁNH KẾT QUẢ

STABLE DIFFUSION 1.5 + LORA VS. DDPM

Chỉ số	SD1.5	DDPM
PSNR	23.45 dB	20.9651 dB
SSIM	0.845	0.8012
LPIPS	0.112	0.1625
FID	32.17	40.4452

KẾT LUẬN

1

Kết quả chính

- Diffusion models hiệu quả cho Image Completion
- Stable Diffusion v1.5 vượt trội hơn DDPM

2

Nguyên nhân

- Khuếch tán trong không gian tiềm ẩn
- Khả năng nắm bắt cấu trúc và ngữ nghĩa tốt hơn

3

Vai trò của LoRA

- Fine-tuning hiệu quả tham số
- Phù hợp tài nguyên hạn chế
- Giữ được khả năng sinh ảnh tổng quát

HƯỚNG ĐI TƯƠNG LAI

MỞ RỘNG DỮ LIỆU

(cảnh tự nhiên, y tế,
vệ tinh, v.v.)

BỔ SUNG ĐIỀU KIỆN HÓA

(text prompt,
references, v.v.)

KỸ THUẬT FINETUNE KHÁC

(DoRA, QLoRA,
adapters, v.v.)

ĐÁNH GIÁ NÂNG CAO

(dựa trên nhận thức
người dùng)

DEMO
