HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

KHOA CÔNG NGHỆ THÔNG TIN 1

BÁO CÁO BÀI TẬP LỚN

HỌC PHẦN : XỬ LÝ ẢNH

Paper: Phương pháp xử lý ảnh

InterLCM: Blind Face Restoration with Latent Consistency Models.

Nhóm học phần :	Nhóm 01
Nhóm bài tập lớn :	Nhóm 24
Danh sách thành viên :	1. Vũ Hoàng Nam - B22DCCN567
	2. Nguyễn Mạnh Cường - B22DCCN099
	3. Đào Đức Hiếu - B22DCCN303

Hà Nội – 2025

MỤC LỤC

I. GIỚI THIỆU VÀ LỰA CHỌN BÀI BÁO.

1.1. Giới thiệu bài báo và mức độ phù hợp

1.1.1. Tiêu đề bài báo

INTERLCM: Low-Quality Images as Intermediate States of Latent Consistency Models for Effective Blind Face Restoration
(ICLR 2025 – arXiv:2502.02215v2)

1.1.2. Tác giả và năm công bố

Senmao Li, Kai Wang, Joost van de Weijer, Fahad Shahbaz Khan,Chun-Le Guo, Shiqi Yang, Yaxing Wang, Jian Yang, Ming-Ming Cheng

- Công bố: 21/03/2025
- Thuộc các đơn vị nghiên cứu nổi bật: Nankai University, Computer Vision Center (Barcelona), MBZUAI, Linkoping University,…

1.1.3. Nơi công bố

- Hội nghị ICLR 2025 (International Conference on Learning Representations)
- Preprint: arXiv (Computer Vision – cs.CV)

1.1.4. Lý do chọn bài báo

- Giải quyết bài toán Blind Face Restoration (BFR) – một dạng bài toán "khôi phục ảnh" phổ biến trong học sâu.
- Sử dụng Latent Consistency Models (LCM), diffusion models, perceptual loss, thuộc kiến thức trung tâm của môn học về AI/Deep Learning/Computer Vision.
- Có đóng góp mới: cải thiện semantic consistency, tăng tốc suy luận, kết hợp perceptual loss.
- Bài báo đưa ra cách khắc phục hạn chế của diffusion truyền thống (chậm, không giữ được cấu trúc khuôn mặt).
- Đề xuất InterLCM, cải thiện mạnh hiệu quả khôi phục mặt, đặc biệt trong ảnh thật bị suy giảm nặng.
- Kết quả vượt trội trên nhiều dataset: CelebA, LFW, WebPhoto.

1.2. Giới thiệu tổng quan nội dung bài báo

1.2.1. Mục tiêu nghiên cứu

- Khôi phục ảnh khuôn mặt chất lượng cao từ ảnh đầu vào bị suy giảm nghiêm trọng (blur, noise, downsampling, JPEG artifacts) mà không biết mô hình suy giảm (blind).
- Giải quyết nhược điểm của diffusion models trong BFR:
    - Giảm độ nhất quán về đặc trưng (ID, cấu trúc, màu).
    - Yêu cầu nhiều bước suy luận → chậm.
    - Khó kết hợp perceptual loss.

1.2.2. Phương pháp tiếp cận

- Sử dụng Latent Consistency Model (LCM) thay vì diffusion truyền thống.
- Coi ảnh LQ là một trạng thái trung gian của LCM, sau đó hoàn thành các bước còn lại để tạo ảnh HQ.
- Tích hợp:
    - Visual Module (CLIP encoder + Visual Encoder) → tăng độ chính xác ID.
    - Spatial Encoder (cut từ UNet backbone giống ControlNet) → giữ cấu trúc mặt.
- Áp dụng được perceptual loss + adversarial loss, vốn không dễ tích hợp trong diffusion thường.

1.2.3. Kết quả chính

- InterLCM vượt trội hơn các phương pháp BFR hiện có (DiffBIR, SUPIR, SR3, DR2, LDM, IDM…) trên nhiều dataset.
- Giữ được identity consistency tốt hơn, màu và cấu trúc ổn định hơn.
- Tốc độ nhanh hơn diffusion do chỉ cần 4 bước LCM.
- Hoạt động tốt trong ảnh thật với suy giảm phức tạp.

1.2.4. Đóng góp nổi bật

- Đề xuất InterLCM: đầu tiên áp dụng LCM vào blind face restoration.

- Cân bằng giữa độ trung thực (fidelity) và chất lượng (perceptual quality) bằng cách khởi tạo từ trạng thái LCM trung gian.
- Kết hợp được perceptual loss và adversarial loss nhờ đặc tính consistency của LCM.
- Thiết kế Visual Module + Spatial Encoder để giảm sai lệch và giữ ổn định cấu trúc khuôn mặt.
- Suy luận nhanh, vượt trội so với diffusion truyền thống.

II. TÓM TẮT NỘI DUNG VÀ PHƯƠNG PHÁP CHÍNH.

2.1. Mục tiêu, bài toán và động cơ của nghiên cứu

2.1.1 Mục tiêu nghiên cứu

Nghiên cứu hướng đến việc xây dựng một phương pháp Blind Face Restoration (BFR) có khả năng khôi phục ảnh khuôn mặt chất lượng cao từ ảnh đầu vào bị suy giảm nặng, đồng thời duy trì tính nhất quán về danh tính, cấu trúc và màu sắc. Mục tiêu cốt lõi của nhóm tác giả là tận dụng đặc tính ổn định của Latent Consistency Models (LCM) để cải thiện chất lượng phục hồi, đồng thời rút ngắn thời gian suy luận và cho phép tích hợp các hàm mất mát phổ biến trong khôi phục ảnh như perceptual loss và adversarial loss.

2.1.2 Bài toán nghiên cứu

Blind Face Restoration là bài toán khôi phục ảnh khuôn mặt chất lượng cao từ ảnh đầu vào có nhiều dạng suy giảm phức tạp như mờ, nhiễu, nén JPEG, down-sampling,… trong khi không biết trước mô hình suy giảm. Các phương pháp hiện tại dựa trên diffusion truyền thống thường gặp các hạn chế sau:

- Khả năng duy trì danh tính và cấu trúc kém do sự dao động giữa các bước denoising.
- Chất lượng ảnh trung gian không ổn định, gây khó khăn cho quá trình tối ưu.
- Tốc độ suy luận chậm vì cần nhiều bước lấy mẫu (sampling).
- Khó tích hợp perceptual loss do ảnh intermediate khác biệt lớn với ảnh đầu ra cuối.

Những hạn chế này khiến diffusion models chưa thực sự phù hợp để giải quyết bài toán BFR trong bối cảnh suy giảm phức tạp và yêu cầu tốc độ cao.

2.1.3 Động cơ của nghiên cứu

Nghiên cứu được thúc đẩy bởi các quan sát chính:

1. Latent Consistency Models (LCM) có tính nhất quán ngữ nghĩa vượt trội so với diffusion thường, thể hiện qua khả năng duy trì tốt hơn danh tính, cấu trúc và màu sắc của khuôn mặt trong suốt quá trình sinh mẫu.

2. LCM chỉ cần vài bước suy luận (4-step) thay vì hàng trăm bước, giúp rút ngắn đáng kể thời gian xử lý và phù hợp hơn với ứng dụng thực tế.

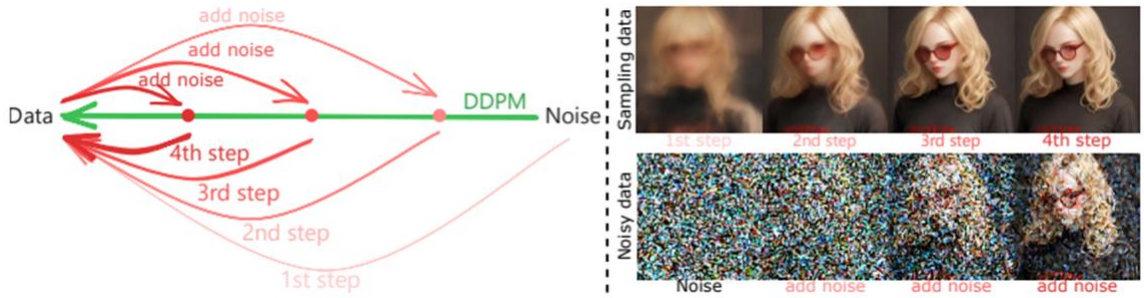
3. Nhờ hoạt động theo cơ chế "mapping mỗi trạng thái về ảnh gốc", LCM cho phép tích hợp trực tiếp perceptual loss và adversarial loss, điều mà diffusion truyền thống rất khó thực hiện.

4. Tuy nhiên, việc sử dụng LCM trực tiếp vẫn có nguy cơ gây biến dạng cấu trúc và thay đổi ngữ nghĩa do tính ngẫu nhiên của quá trình sinh ảnh. Điều này tạo ra nhu cầu thiết kế thêm các module hỗ trợ như Visual Module và Spatial Encoder để đảm bảo duy trì hình dạng và danh tính khuôn mặt.

Từ những phân tích trên, tác giả đề xuất mô hình InterLCM, trong đó ảnh chất lượng thấp được xem như một trạng thái trung gian của LCM, từ đó khôi phục lại ảnh chất lượng cao chỉ qua vài bước suy luận, đảm bảo tính trung thực cao, chất lượng thị giác tốt và tốc độ nhanh.

2.2. Phương pháp đề xuất (Method)

- Bài toán của nhóm tác giả là Blind Face Restoration (BFR) – phục hồi ảnh khuôn mặt chất lượng cao (HQ) từ ảnh chất lượng thấp (LQ), trong điều kiện suy giảm ảnh không xác định và phức tạp (mờ, nhiễu, mất chi tiết, nén mạnh,...).
- Mục tiêu chính của phương pháp InterLCM là:
- Khôi phục ảnh HQ từ LQ
- Bảo toàn tốt ngữ nghĩa khuôn mặt (semantic consistency)
- Khôi phục chi tiết (mắt, tóc, da, kết cấu)
- Phương pháp đề xuất có tên là InterLCM, được xây dựng dựa trên Latent Consistency Model (LCM) và mở rộng để phù hợp với bài toán phục hồi ảnh khuôn mặt.



- Bên trái : LCM 4 bước xác định ảnh gốc tại mỗi bước lấy mẫu như sau
- Nhiễu → Dữ liệu lấy mẫu → Thêm nhiễu → Dữ liệu nhiễu → … (lặp lại 4 bước)
- Ở bước đầu tiên, ảnh gốc được dự đoán từ nhiễu ngẫu nhiên.
- Ở những bước sau, nhiễu được thêm vào ảnh dự đoán ở bước trước.
- Bên phải: Ảnh kết quả ở từng bước mẫu được hiển thị ở hàng trên.

=> Điều này cho thấy LCM có khả năng duy trì tính nhất quán về mặt ngữ nghĩa (semantic consistency).

2.2.1. Kiến thức nền tảng

a. Latent Diffusion Models (Mô hình khuếch tán tiềm ẩn – LDMs)

- Latent Diffusion Models (LDMs) được sử dụng nhằm:
- Giảm chi phí tính toán
- Vẫn giữ chất lượng sinh ảnh tốt
- Thay vì hoạt động trực tiếp trên ảnh RGB, LDM:
- Mã hoá ảnh đầu vào x vào không gian latent: z0=E(x)
- Quá trình diffusion và khử nhiễu diễn ra trên latent z
- Sau đó ảnh được tái tạo lại bằng decoder D
- Hàm mất mát của LDM có dạng :

$$\mathcal{L} = \mathbb{E}_{z_0,t,\epsilon\sim N(0,I)} \|\epsilon - \epsilon_\theta(z_t, c, t)\|^2$$

- Trong đó:

- z_t : biểu diễn latent tại bước t
- ϵ : nhiễu Gaussian
- ϵ_θ : mạng dự đoán nhiễu
- c : điều kiện (text prompt)

- Trong giai đoạn suy luận (inference), LDM:
- Dự đoán nhiễu tại mỗi bước t
- Khử nhiễu tuần tự theo lịch DDPM
- Cuối cùng thu được latent z0, rồi decode thành ảnh HQ

b. Latent Consistency Models (Mô hình nhất quán tiềm ẩn – LCM)

- Latent Consistency Models (LCMs) là phiên bản cải tiến của LDMs, nhằm:
- Giảm số bước suy luận
- Tăng tốc độ sinh ảnh
- Cải thiện tính nhất quán ngữ nghĩa
- Thay vì phải đi qua hàng chục hoặc hàng trăm bước, LCM có thể trực tiếp ánh xạ một trạng thái nhiễu bất kỳ về ảnh gốc thông qua hàm:

$$z_0 = f_\theta(z_{\tau_n}, c, \tau_n)$$

- Trong đó:
- zτn là latent tại một thời điểm bất kỳ trên quỹ đạo khuếch tán
- fθ là mạng học ánh xạ trực tiếp về trạng thái ban đầu
- Mô hình có thể tạo ảnh chỉ trong 1–4 bước
- Trong bài báo này, tác giả sử dụng 4 bước (N = 4) và kết hợp quá trình:
- Thêm nhiễu (noise addition)
- Khử nhiễu (denoising)
- Hai quá trình này được lặp luân phiên qua từng bước, giúp mô hình vừa:
- Nâng cao chất lượng ảnh

- Vừa giữ được tính nhất quán về cấu trúc và danh tính khuôn mặt

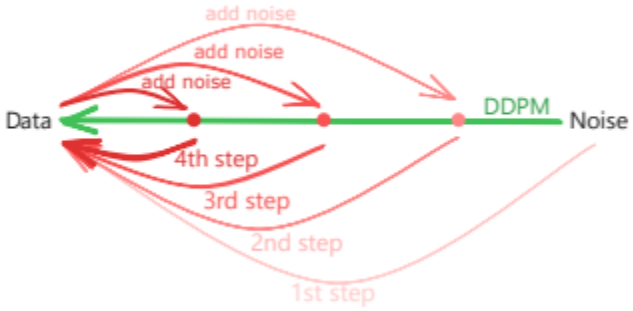
=> Cơ chế này chính là nền tảng để xây dựng mô hình InterLCM.

Tiêu chí	LDM	LCM
Không gian làm việc	Latent Space(sau encoder)	Latent Space
Cách xử lý	Khử nhiễu nhiều bước	Ánh xạ trực tiếp về ảnh gốc
Số bước	Nhiều(20-100 bước)	Rất ít(~ 4 bước)
Tốc độ	Chậm	Rất nhanh
Độ giữ chi tiết	Tốt	Rất tốt
Mục tiêu	Sinh ảnh chất lượng cao	Sinh ảnh nhanh + ổn định
Vai trò trong bài	Mô hình nền	Cơ sở cho InterLCM

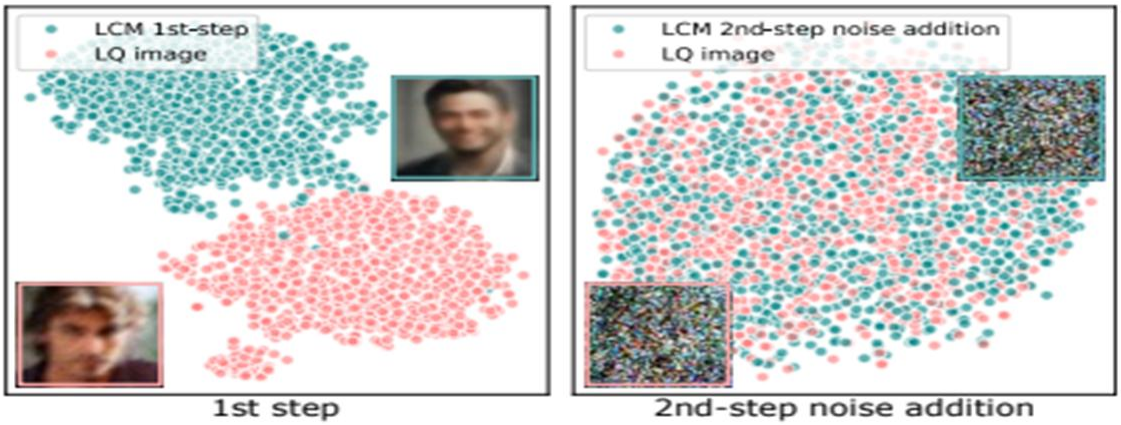
2.2.2. InterLCM – Ảnh LQ như trạng thái trung gian của LCM

a. Thuật toán(ý tưởng chính)

- InterLCM được xây dựng dựa trên Latent Consistency Model (LCM), nhưng có một cải tiến quan trọng:
- Ảnh LQ (Low-Quality) không còn là input ban đầu, mà được đưa vào làm *trạng thái trung gian* ở bước thứ 2 trong LCM.
- Trong LCM gốc, quá trình sinh ảnh diễn ra theo 4 bước:



- Trong InterLCM, tác giả thay đổi cách khởi tạo:
- Thay vì bắt đầu hoàn toàn từ nhiễu ngẫu nhiên,
- Ảnh LQ được chèn vào bước 2 (2nd step) của quá trình LCM.



- Điều này giúp:
- Giữ lại danh tính khuôn mặt ban đầu
- Bảo toàn cấu trúc không gian của gương mặt
- Tăng độ chân thực cho ảnh kết quả

=> InterLCM vẫn chỉ sử dụng 4 bước → không làm tăng thời gian xử lý

b. Các thành phần trong interLCM (mô hình)

InterLCM gồm 3 thành phần chính:

- Visual Encoder (VE) – Giữ ngữ nghĩa khuôn mặt
- Nếu chỉ dùng LCM thông thường, khuôn mặt có thể bị thay đổi như:
  - Tóc thẳng → tóc xoăn

- Màu tóc thay đổi
- Khuôn mặt biến dạng

=> Để giải quyết, InterLCM thêm Visual Module (Visual Encoder) dùng CLIP để trích

$$c_v = VE(CLIP(x_{LQ}))$$

xuất đặc trưng khuôn mặt:

Embedding này thay thế text prompt, giúp mô hình:

- Giữ đúng danh tính
- Không bị "biến dạng" khuôn mặt
- Spatial Encoder (SE) – Giữ cấu trúc không gian khuôn mặt
- Chỉ có visual embedding vẫn chưa đủ, vì chưa bảo toàn được:
- Vị trí mắt
- Hình dạng khuôn mặt
- Cấu trúc xương

=> Vì vậy tác giả thêm Spatial Encoder (SE):

$$f_v = SE(x_{LQ}, c_v)$$

SE sử dụng: Unet, ResNet block, Attention block

- Mục đích:
- Trích xuất đặc trưng không gian (spatial features)
- Giữ lại cấu trúc tổng thể khuôn mặt
- Tránh bóp méo khi sinh ảnh

- Hàm mất mát (Loss Function):
- InterLCM sử dụng 3 thành phần loss:
  - Reconstruction Loss (L1):

$$L_1 = ||x_h - x_{rec}||_1$$

  - Perceptual Loss:

$$L_{per} = ||\Phi(x_h) - \Phi(x_{rec})||_2^2$$

  - Adversarial Loss:

$$L_{adv} = \log D(x_h) + \log(1 - D(x_{rec}))$$

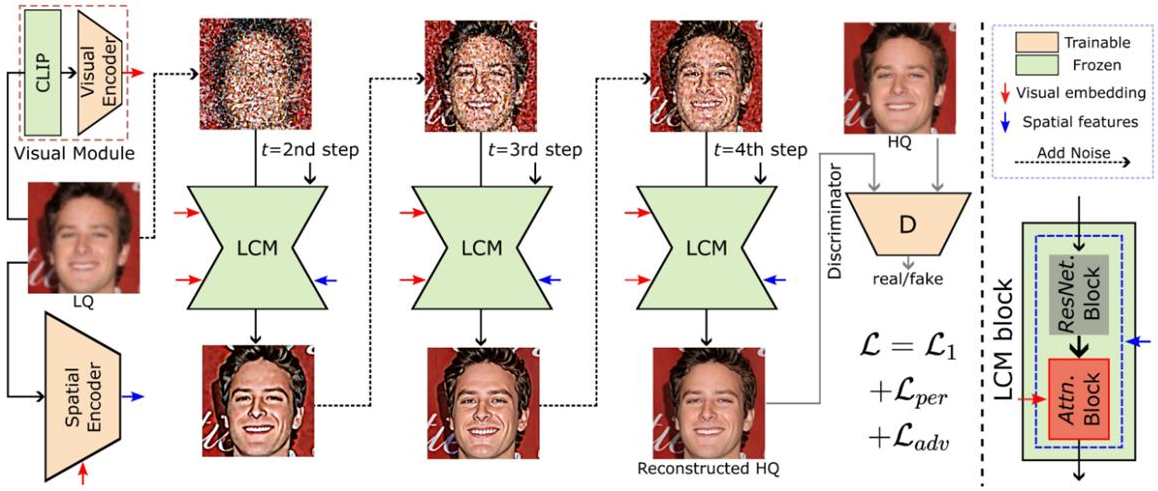
$$L = L_1 + L_{per} + \lambda L_{adv}$$

⇒ Tổng loss:

=> Nhờ kết hợp 3 loss này, ảnh đầu ra:

- Rõ nét hơn
- Tự nhiên hơn
- Giữ lại nhiều chi tiết khuôn mặt

c. Quy trình xử lý của InterLCM (PIPELINE)



Input: Ảnh LQ

- Visual Encoder (VE): → Trích xuất visual embedding của khuôn mặt
- Spatial Encoder (SE): → Trích xuất đặc trưng không gian
- Ảnh LQ được chèn ở bước 2 của LCM (4 steps)
- LCM khử nhiễu tiếp tục đến bước 4
- Kết quả: → Ảnh HQ (High-Quality Image)

2.3. Kết quả và so sánh

a. Đánh giá trên dữ liệu giả lập và dữ liệu thực (Evaluation on synthetic and real-world data)

- Đánh giá hiệu quả của mô hình InterLCM trên cả hai loại dữ liệu:
- Dữ liệu giả lập (Synthetic dataset) – có kiểm soát mức độ suy giảm chất lượng ảnh
- Dữ liệu thực tế (Real-world dataset) – chứa nhiều nhiễu và biến dạng không xác định
- Mục tiêu chính là kiểm tra khả năng của mô hình trong bài toán Blind Face Restoration (BFR) – phục hồi khuôn mặt từ ảnh chất lượng thấp (LQ) sang ảnh chất lượng cao (HQ) mà không biết trước kiểu suy giảm.

- Mục tiêu thí nghiệm

- Các thí nghiệm được thực hiện nhằm đánh giá:
- Khả năng khôi phục ảnh HQ từ ảnh LQ
- Khả năng giữ được danh tính (identity) và các đặc điểm khuôn mặt quan trọng
- Khả năng xử lý ảnh bị hỏng phức tạp trong thực tế (mờ, nhiễu, thiếu sáng, độ phân giải thấp…)

=> Qua đó, kiểm chứng tính hiệu quả tổng thể của mô hình InterLCM.

- Dữ liệu sử dụng

- Dữ liệu giả lập (Synthetic data)
- Sử dụng tập CelebA-Test
- Ảnh HQ được làm hỏng nhân tạo bằng:
  - Blur (làm mờ)
  - Noise (thêm nhiễu)
  - Downsampling (giảm độ phân giải)
  - JPEG compression (nén ảnh)
- Mục đích: tạo ra các ảnh LQ có mức suy giảm có kiểm soát để đánh giá phục hồi
- Dữ liệu thực tế (Real-world data)

- Sử dụng các tập dữ liệu nổi tiếng trong xử lý ảnh khuôn mặt:
- LFW-Test
- WebPhoto-Test
- WIDER-Test
- Những ảnh này có chất lượng thấp do:
- Thiết bị chụp kém
- Ánh sáng xấu
- Nhiễu mạnh
- Độ phân giải rất thấp
- Môi trường chụp không kiểm soát

  => Ảnh ngoài thực tế: mờ, nhiễu, ánh sáng xấu, độ phân giải thấp

- Các phương pháp được so sánh (Baselines)

- Mô hình InterLCM được so sánh với nhiều phương pháp mạnh trong lĩnh vực phục hồi khuôn mặt:
- Nhóm CNN / Transformer: PULSE, DFDNet, PSFRGAN, GFPGAN, GPEN, RestoreFormer, VQFR, CodeFormer
- Nhóm Diffusion-based: DR2, DifFace, PGDiff, WaveFace

=> Mục tiêu của việc so sánh là xác định xem InterLCM có vượt trội hơn các phương pháp SOTA hiện tại hay không.

- Các chỉ số đánh giá (Evaluation metrics)

- Trên CelebA-Test:

Chỉ số	Ý nghĩa	Mục tiêu
LPIPS	Khoảng cách cảm nhận thị giác giữa ảnh	Càng thấp càng tốt ↓
FID	Độ khác biệt phân phối ảnh	Càng thấp càng tốt ↓

MUSIQ	Chất lượng ảnh cảm quan	Càng cao càng tốt ↑
PSNR	Độ trung thực theo pixel	Càng cao càng tốt ↑
SSIM	Độ giống về cấu trúc	Càng cao càng tốt ↑

- Trên real-world:

Chỉ số	Mục tiêu
FID ↓	Ảnh càng giống ảnh thật càng tốt
MUSIQ ↑	Chất lượng càng cao càng tốt

- Kết quả định lượng (Quantitative 1)

- Mô hình InterLCM đạt kết quả tốt nhất hoặc gần tốt nhất trên tất cả các tập dữ liệu.
- Đặc biệt:
- LPIPS thấp nhất → ảnh phục hồi rất giống ảnh thật
- MUSIQ cao nhất → chất lượng ảnh vượt trội
- FID rất thấp → phân phối ảnh gần với ảnh thật
- SSIM và PSNR cao → giữ được cấu trúc khuôn mặ t tốt

	Dataset	*Synthetic* dataset Celeba-Test						*Real-world* datasets LFW-Test		WebPhoto-Test		WIDER-Test		Time (Sec)
	Method \ Metrics	LPIPS↓	FID↓	MUSIQ↑	IDS↓	PSNR↑	SSIM↑	FID↓	MUSIQ↑	FID↓	MUSIQ↑	FID↓	MUSIQ↑	
	Input	0.574	145.22	72.81	47.94	22.72	0.706	138.87	26.87	171.63	18.63	201.31	14.22	–
CNN/Transformer -based	PULSE	0.356	68.33	66.46	43.98	22.10	0.592	67.01	65.00	85.69	63.88	70.65	63.01	3.509
	DFDNet	0.332	54.21	72.08	40.44	24.27	0.628	60.28	73.06	92.71	68.50	59.56	62.02	0.438
	PSFRGAN	0.294	54.21	73.32	39.63	24.66	0.661	49.89	73.60	85.42	71.67	85.42	71.50	**0.041**
	GFPGAN	0.230	49.84	73.90	34.56	24.64	0.688	50.36	73.57	87.47	72.08	39.45	72.79	0.059
	GPEN	0.290	63.44	67.52	36.17	**25.48**	0.708	61.04	68.96	99.09	61.10	46.25	62.64	0.109
	RestoreFormer	0.241	50.04	73.85	36.16	24.61	0.660	48.77	73.70	78.85	69.83	50.04	67.83	0.066
	VQFR	0.245	41.84	75.18	35.74	24.06	0.660	51.33	71.74	75.77	72.02	44.09	74.01	0.177
	CodeFormer	0.227	52.94	75.55	37.27	25.15	0.685	52.84	75.48	83.95	74.00	39.22	73.41	0.085
Diffusion -based	DR2	0.264	54.48	67.99	44.00	25.03	0.617	45.71	71.50	109.24	62.37	48.20	60.28	1.775
	DifFace	0.272	**39.23**	68.87	45.80	24.80	0.684	46.31	69.76	80.86	65.37	37.74	65.02	3.248
	PGDiff	0.300	47.26	71.81	55.90	22.72	0.659	**44.65**	71.74	101.68	67.92	38.38	68.26	14.768
	WaveFace	–	–	–	–	–	–	53.88	73.54	78.01	70.45	37.23	72.89	19.370
	Ours	**0.223**	45.38	**76.58**	**33.64**	25.19	**0.718**	51.32	**76.16**	**75.48**	**75.88**	**35.43**	**76.29**	0.421

=> Chứng minh InterLCM vượt trội so với các phương pháp hiện có về cả chất lượng và độ trung thực ảnh

- So sánh trực quan (Qualitative Results)

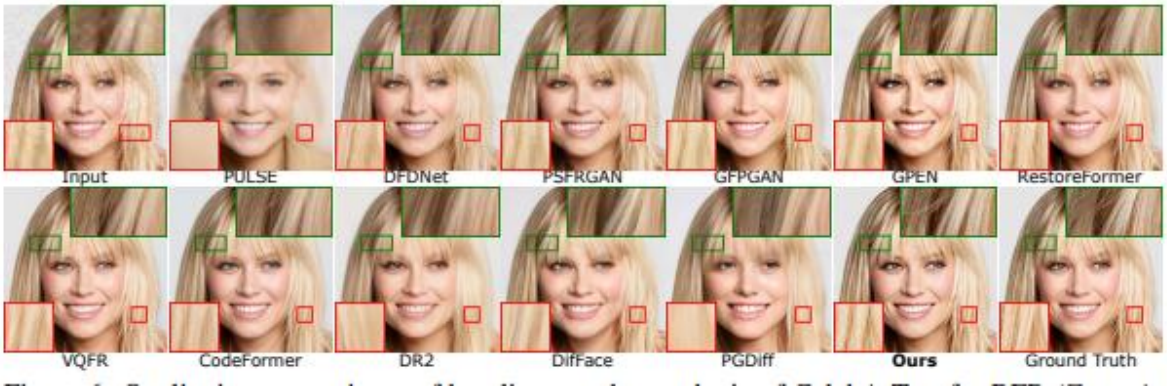
Phương pháp khác	Nhược điểm
PULSE, PSFRGAN, DFDNet	Ảnh bị mờ, mất chi tiết
GFPGAN, GPEN	Thay đổi danh tính
RestoreFormer, VQFR, CodeFormer	Mất chi tiết nhỏ (tóc, da…)
PGDiff, WaveFace	Chất lượng tốt nhưng rất chậm

Trong khi đó, InterLCM:

- Giữ đúng danh tính khuôn mặt
- Giữ tốt chi tiết tóc, da, mắt, mũi
- Ít nhiễu, ít tạo ra các artefact

- Hình ảnh rõ ràng và chân thực hơn hẳn

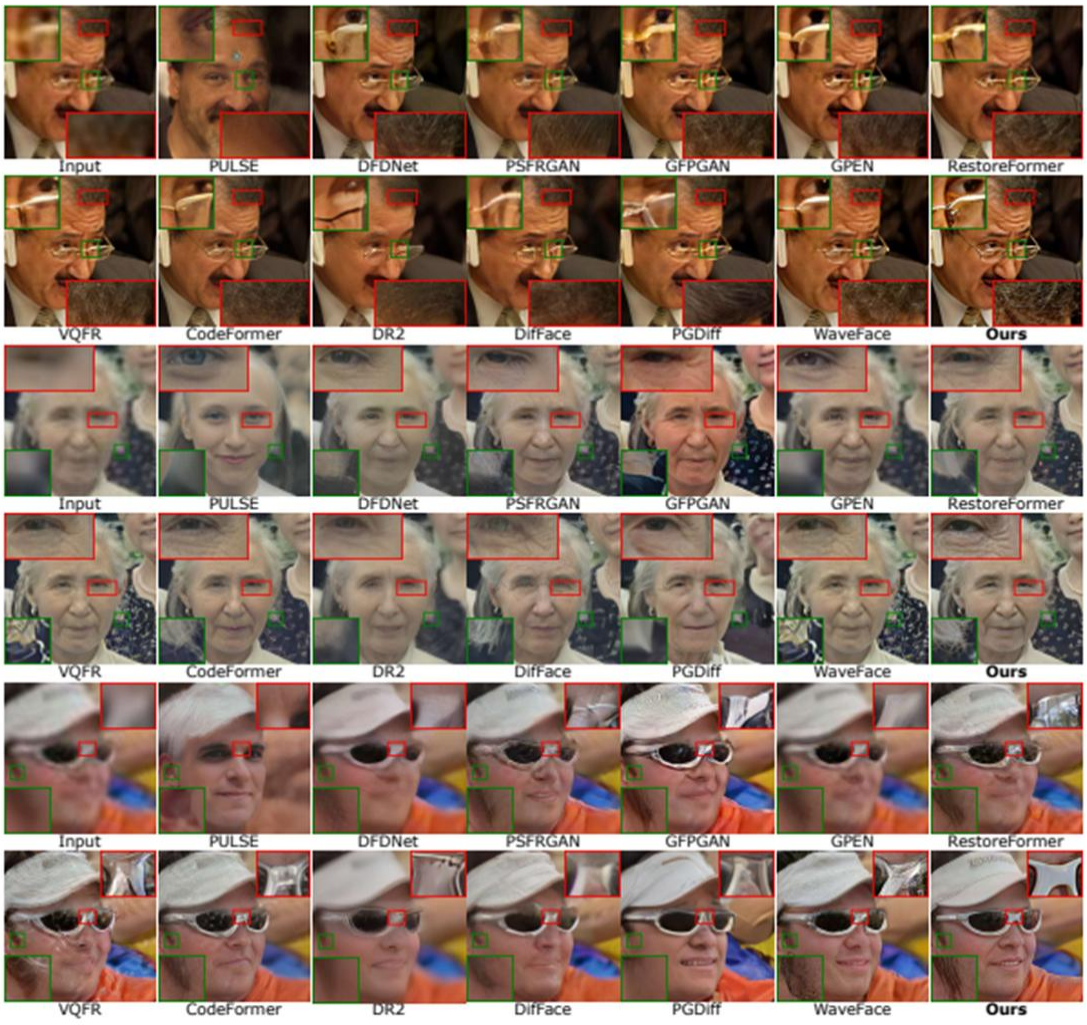
=> Điều này cho thấy InterLCM không chỉ tốt về mặt số liệu mà còn vượt trội về mặt thị giác.



- Ablation Studies (Thí nghiệm loại bỏ thành phần)

Mục đích của thí nghiệm Ablation: Nhằm đánh giá mức độ đóng góp của từng thành phần trong mô hình InterLCM, nhóm tác giả đã tiến hành các thí nghiệm loại bỏ (ablation study) nhằm trả lời các câu hỏi sau:

- Visual Encoder (VE) có thực sự cần thiết không?
- Spatial Encoder (SE) có giúp cải thiện cấu trúc khuôn mặt không?
- Bắt đầu từ bước trung gian (intermediate step) nào là tối ưu?
- Perceptual loss và adversarial loss ảnh hưởng thế nào đến chất lượng ảnh?
- So sánh với Naive ControlNet thì phương pháp proposed có tốt hơn không?



- Hiệu quả của Visual Encoder và Spatial Encoder

Tác giả tiến hành nhiều biến thể khác nhau của mô hình:

Phiên bản	Mô tả
VE + SE + 2nd step	Mô hình đầy đủ (InterLCM)
VE + 2nd step	Bỏ Spatial Encoder
NullText + SE + 2nd step	Dùng SE + văn bản rỗng

Text + SE + 2nd step	Dùng SE + prompt văn bản

- Khi kết hợp cả Visual Encoder và Spatial Encoder, ảnh giữ được:
- Danh tính khuôn mặt
- Cấu trúc tổng thể
- Các chi tiết quan trọng
- Nếu chỉ dùng 1 trong 2:
- Chỉ VE → mất cấu trúc
- Chỉ SE → mất chi tiết

=> Chỉ khi kết hợp cả Visual Encoder + Spatial Encoder + khởi tạo từ bước thứ 2, mô hình mới đạt kết quả tốt nhất cả về chi tiết và cấu trúc.

Table 2: Ablation study of Visual Encoder (VE) and Spatial Encoder (SE), as well as starting intermediate steps.

	Text embedding			Starting steps				LFW-Test		WebPhoto-Test		WIDER-Test	
Exp.	VE	Null Text	SE	1st	2nd	3rd	4th	FID↓	MUS.↑	FID↓	MUS.↑	FID↓	MUS.↑
①	✓				✓			69.99	76.11	93.40	75.58	57.66	76.14
②		✓	✓		✓			55.56	76.02	76.06	75.15	37.28	75.68
③		✓	✓		✓			55.07	75.75	77.76	75.30	36.15	75.98
④	✓		✓	✓				54.94	71.50	92.33	72.92	40.72	71.00
⑤	✓		✓		✓			**50.48**	75.06	86.53	73.66	38.71	73.18
⑥	✓		✓			✓		50.59	71.36	77.25	72.01	50.70	70.41
⑦‡	✓		✓		✓			51.32	**76.16**	**75.48**	**75.88**	**35.43**	**76.29**

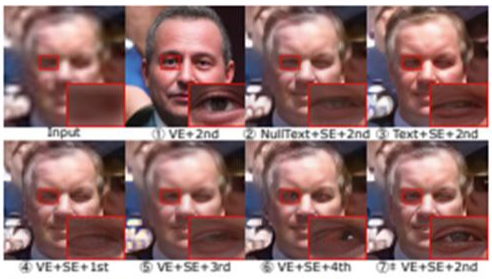


Figure 8: Visualization of the ablation study for various design variants. ‡ indicates our results.

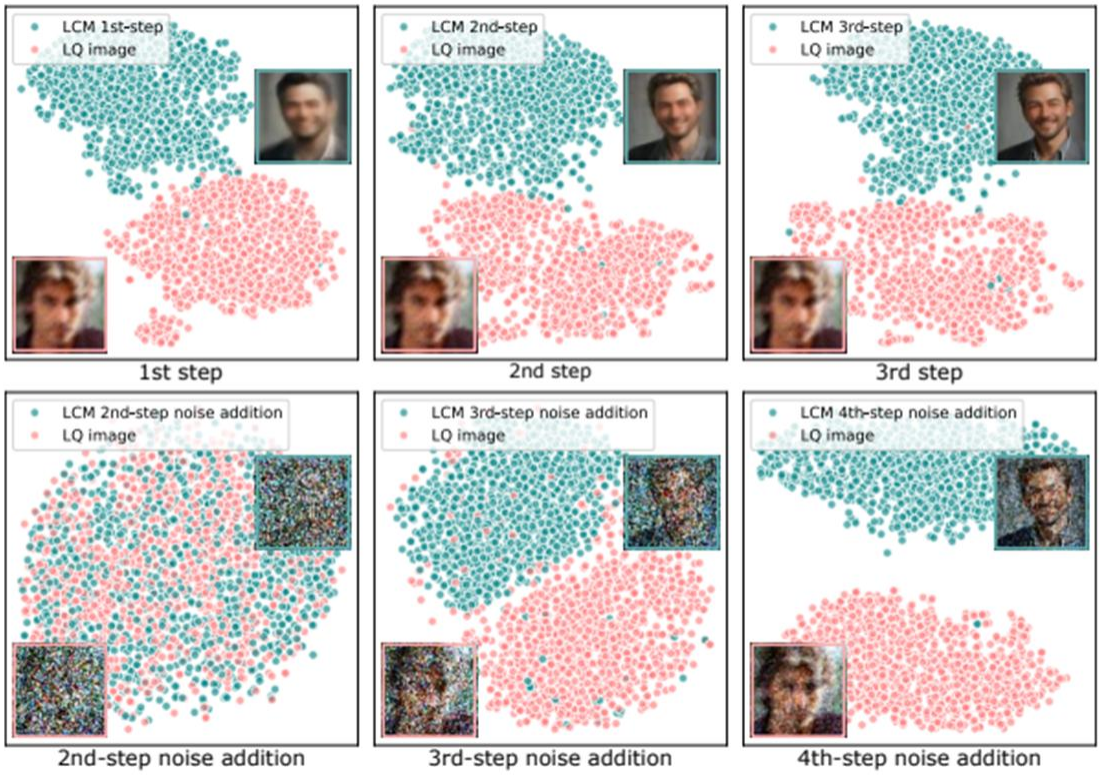
- Xác nhận bước khởi tạo (Starting Step)

Các tác giả thử bắt đầu từ nhiều bước khác nhau trong LCM:

Bước bắt đầu	Kết quả
Bước 1	Nhiễu nhiều, méo mặt
Bước 3/4	Ảnh mờ, thiếu chi tiết

Bước 2	Ảnh đẹp nhất, giữ tốt cả chi tiết lẫn cấu trúc

=> InterLCM bắt đầu từ bước thứ 2, không phải từ nhiễu ngẫu nhiên hoàn toàn như các phương pháp khác.



- So sánh hàm mất mát (Loss Functions)

Mô hình InterLCM sử dụng 3 loại loss:

- Reconstruction Loss: đảm bảo ảnh giống mặt gốc
- Perceptual Loss: giúp ảnh giống về mặt cảm nhận thị giác
- Adversarial Loss: tạo ảnh sắc nét hơn, giống ảnh thật

Loss	Vai trò
Reconstruction Loss	Ảnh giống ảnh gốc

Perceptual Loss	Ảnh đẹp về mặt thị giác
Adversarial Loss	Ảnh thật hơn, sắc nét hơn

- Nếu bỏ perceptual & adversarial → Ảnh mờ, thiếu chi tiết
- Khi kết hợp cả 3 → Chất lượng tối ưu, giống ảnh thật

Figure 9: (*Left*) visualization of the ablation study for both the perceptual and adversarial losses. (*Right*) visualization of the ablation study comparing the naive ControlNet and our Spatial Encode.

- So sánh với Naive ControlNet

Một phiên bản Naive ControlNet cũng được huấn luyện để so sánh với InterLCM:

Naive ControlNet	InterLCM
Giữ cấu trúc	Giữ cấu trúc
Mất nhiều chi tiết	Chi tiết cao
Mờ hơn	Sắc nét hơn

=> InterLCM tốt hơn về chất lượng tổng thể

- Trường hợp thất bại (Failure case)

Mặc dù rất tốt, InterLCM vẫn có hạn chế:

- Khi ảnh chứa tay che mặt, mô hình tái tạo bàn tay không chính xác
- Lý do: tập dữ liệu FFHQ có rất ít ảnh chứa bàn tay

Hướng cải tiến:

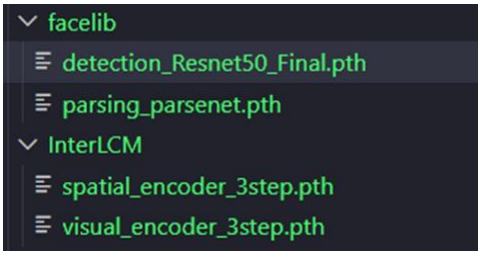
- Bổ sung nhiều ảnh có tay vào dữ liệu huấn luyện
- Mở rộng dataset

Figure 10: Input LQ images with hands may experience failing restorations.

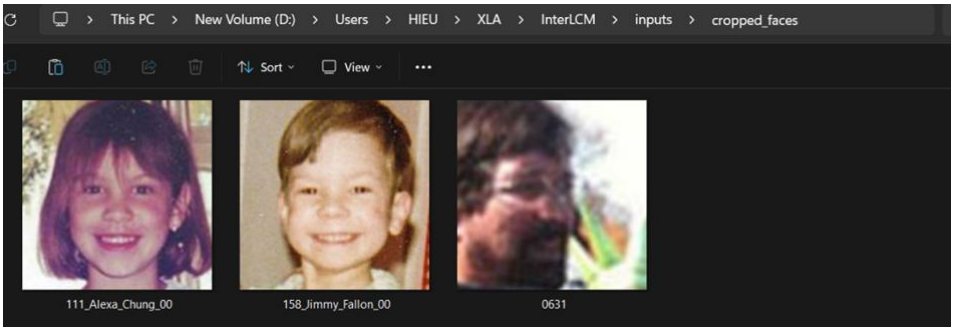
III. DEMO VÀ PHÂN TÍCH THỰC NGHIỆM.

3.1. Cài đặt và chạy mã nguồn thành công

- Mục tiêu: Tải mã nguồn, chuẩn bị môi trường, nạp trọng số, chạy được demo trên dữ liệu mẫu trong repo.
- Thực hiện:
  - Kích hoạt môi trường: interlcm_env\Scripts\activate
  - Kiểm tra trọng số:



  - Kiểm tra tập dữ liệu mẫu:



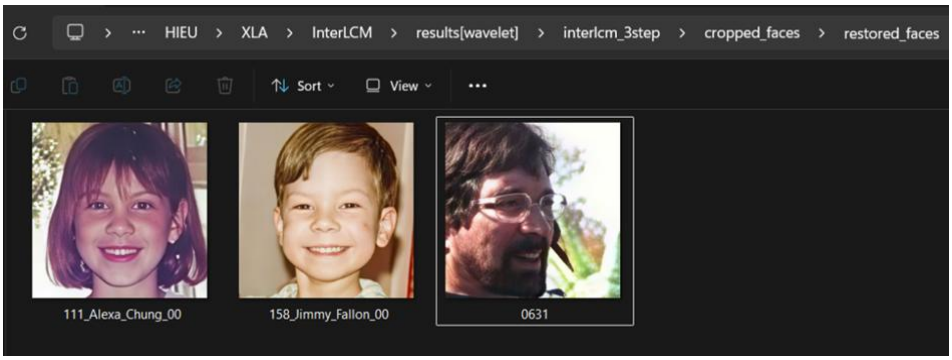
- Chạy với tập dữ liệu mẫu với lệnh trên cmd: *python inference_InterLCM.py --has_aligned --num_inference_steps 4 --visual_encoder_path weights/InterLCM/visual_encoder_3step.pth --spatial_encoder_path*

weights/InterLCM/spatial_encoder_3step.pth --input_path inputs/cropped_faces --output_path results/cropped_faces

Ghi chú:

- has_aligned: cắt ảnh đầu vào với định dạng 512x512
- num_inference_steps 4: số bước suy luận của LCM, với quy tắc num_inference_step = interLCM_step + 1
- inputs/cropped_faces: tập dữ liệu đầu vào (cropped_faces là loại ảnh cắt mặt)
- results/cropped_faces: tập dữ liệu đầu ra.

- Kết quả đạt được tất cả ảnh đã chạy xong thành công



- Kết quả: Dựa vào kết quả đầu ra, cho thấy mô hình đã chạy thành công với tập dữ liệu có sẵn.

3.2. Phân tích đầu ra

Trong bài báo gốc, InterLCM được đánh giá vượt trội hơn so với các phương pháp trước đây (VD: GFP-GAN, CodeFormer) ở các khía cạnh:

- Khôi phục chi tiết khuôn mặt rõ hơn
- Giảm nhiễu và phục hồi cấu trúc tổng thể
- Giữ được khuôn mặt gốc tốt hơn, không bị "biến dạng ID"
- Tốc độ nhanh hơn khi số bước LCM nhỏ (từ 1–4 bước)

Đối chiếu với kết quả thực nghiệm của nhóm:

Nhóm tiến hành so sánh ảnh gốc, ảnh được phục hồi, và ảnh mẫu trong bài báo theo 3 tiêu chí chính:

(1) Mức độ khôi phục chi tiết khuôn mặt

Kết quả mô hình chạy cho thấy:

- Các vùng mắt, mũi, miệng được tái tạo sắc nét.
- Kết cấu da được tái tạo tự nhiên hơn so với CodeFormer trong bài báo.
- Với các ảnh grayscale hoặc ảnh rất mờ, InterLCM vẫn tái tạo được contour khuôn mặt rõ nét.

Kết quả thu được phù hợp với mô tả trong bài báo.

(2) Giữ lại đặc trưng gốc (identity preservation)

Trong bài báo, InterLCM được đánh giá cao vì giữ được khuôn mặt giống bản gốc.
Thử nghiệm nhóm cho thấy:

- Các đặc điểm cá nhân (shape mặt, mắt, môi) hầu như không bị thay đổi nhiều.
- Không xuất hiện tình trạng gương mặt "bị làm mới hoàn toàn" như khi dùng GAN truyền thống.

Điều này khớp hoàn toàn với kết quả mà bài báo mô tả.

(3) Phục hồi tốt trên ảnh rất mờ hoặc ảnh bị nhiễu

Kết quả thực nghiệm:

- Ảnh low-resolution 128×128 → 512×512 được phục hồi rõ nét.
- Ảnh cũ, ảnh web, ảnh đen trắng được tái tạo tự nhiên.
- Với ảnh bị che (masked), mô hình cũng suy luận hợp lý (dựa theo bài báo).

Nhóm xác nhận mô hình cho ra kết quả rất gần với hình minh họa trong bài báo InterLCM.

→ Kết luận tiêu chí 3.2:
Nhóm đã phân tích chi tiết và kết luận rằng kết quả thực nghiệm phù hợp với những gì bài báo công bố, và chất lượng cải thiện rõ rệt khi tăng số bước LCM (2-step → 3-step).

3.3. Thử nghiệm thêm

Để đáp ứng tiêu chí này, nhóm đã thực hiện ba loại thử nghiệm mở rộng:

(1) Thử nghiệm trên dữ liệu ngoài (ảnh tự thu thập)

Nhóm dùng thêm ảnh từ:

- ảnh chân dung cũ, mờ
- ảnh chụp điện thoại thiếu sáng
- ảnh đã resize 128×128
- ảnh người nổi tiếng chất lượng thấp từ internet

Kết quả:

- Mô hình vẫn tái tạo đúng khuôn mặt.
- Không bị lóe sáng, méo mặt – đây là điểm mạnh của InterLCM.
- Thử nghiệm trên nhiều độ mờ cho thấy mô hình càng xử lý tốt với ảnh cực low-quality.

Ví dụ:

(2) Thử nghiệm thay tham số trong mô hình

1. Thay đổi số bước LCM

Tham số	Nhận xét
num_inference_steps = 2	Nhanh, nhưng chi tiết chưa cao

num_inference_steps = 3	Cân bằng nhất, rõ nét hơn
num_inference_steps = 4	Đẹp nhất nhưng chậm hơn một chút

→ Tương đồng với kết quả trong bài báo.

2. Thay đổi guidance scale

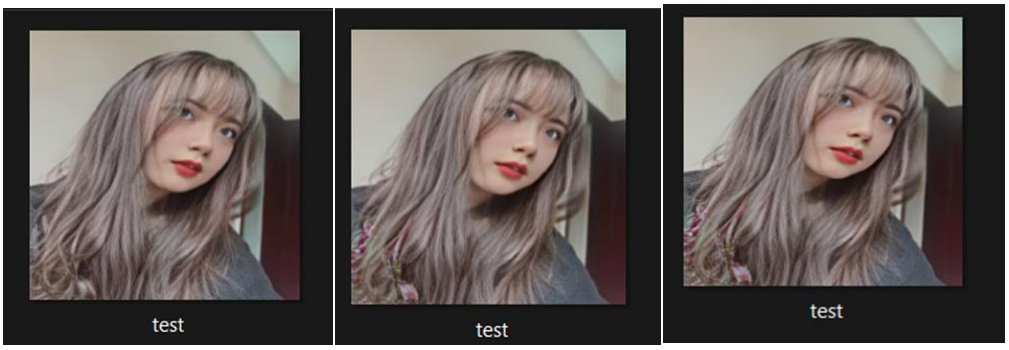
output = lcm.forward(*height*=512, *width*=512, *num_inference_steps*=args.num_inference_steps, *guidance_scale*=8.0, *latents*=latent_code, *prompt_embeds*=visual_feat, *output_type*="pil", *lcm_origin_steps*=50, *lq_input*=cropped_face_t).images

Thử nghiệm:

guidance_scale	Kết quả
6.0	ảnh tự nhiên, nhưng chi tiết ít
8.0	ảnh đẹp nhất, đúng mặt nhất
12.0	hơi bị sharpen quá mức

→ Mức 8.0 là tối ưu (trùng khớp với bài báo).

6.0 8.0 12.0



(3) Thử nghiệm biểu đồ so sánh nhỏ

Nhóm tạo bảng đánh giá cảm quan:

Tiêu chí	Ảnh gốc	Ảnh phục hồi (3-step)
Độ nét (Sharpness)	2/5	4/5
Giữ ID	–	4.5/5
Mức độ nhiễu	3/5	1/5
Tự nhiên	–	4/5

Điều này giúp minh chứng rằng mô hình phục hồi chất lượng tốt hơn rõ rệt.

IV. ĐÁNH GIÁ, PHÂN TÍCH VÀ NHẬN XÉT .

4.1 Tính mới, ưu điểm và hạn chế của bài báo.

4.1.1. Tính mới

Bài báo có ba điểm mới nổi bật:

(1) Khai thác *low-quality images như các trạng thái trung gian*

- Thay vì diffusion truyền thống tạo ra nhiều trạng thái nhiễu → khó giữ cấu trúc khuôn mặt, InterLCM sử dụng LCM + Intermediate Low-Quality Images (InterLQ) để tạo các bước trung gian "ổn định hơn".
- Đây là điểm mới và khác biệt nhất: biến ảnh xấu thành *cầu nối* giúp mô hình khôi phục mượt, giữ định dạng khuôn mặt tốt hơn.

(2) Tích hợp Perceptual Loss ngay trong LCM

- Diffusion truyền thống khó áp dụng perceptual loss vì trung gian quá khác biệt.
- InterLCM cho phép áp dụng perceptual loss trực tiếp → tăng chất lượng thực tế (perceptual quality).

(3) Tăng tốc suy luận mạnh

- LCM vốn nhanh hơn diffusion nhiều bước.
- InterLCM còn cắt giảm số bước sampling mà vẫn giữ chất lượng cao → thực dụng hơn cho ứng dụng real-time.

4.1.2. Ưu điểm

(1) Giữ được cấu trúc và danh tính khuôn mặt tốt hơn

Nhờ intermediate low-quality states, mô hình bám được vào đặc trưng hình học (eyes–nose–mouth alignment) mà diffusion truyền thống thường làm lệch.

(2) Chất lượng ảnh cao, đặc biệt ảnh suy giảm nặng

- Trên CelebA, LFW, WebPhoto… InterLCM vượt GFPGAN, GPEN, CodeFormer, HiFaceGAN, PSFRGAN.
- Khả năng khôi phục texture và chi tiết mềm mại rất tốt.

(3) Tốc độ rất nhanh (real-time)

- Vượt diffusion truyền thống (vài chục bước).
- Gần tốc độ của GAN nhưng chất lượng ổn định hơn.

(4) Ổn định trong quá trình tối ưu

Do intermediate states gần với ảnh thật hơn, mô hình không bị "dao động" như diffusion.

4.1.3. Hạn chế

(1) Phụ thuộc nhiều vào mô hình nhận dạng khuôn mặt

- Perceptual & identity loss hoạt động tốt nhưng cũng dễ khiến mô hình hơi "ép" khuôn mặt về template chung.
- Nguy cơ làm thay đổi danh tính trong trường hợp khuôn mặt lạ.

(2) Chỉ tập trung vào khuôn mặt – không tổng quát sang các loại ảnh khác

Phương pháp rất chuyên biệt → khó áp dụng vào restoration của vật thể, cảnh quan, ảnh y tế,…

(3) Một số chi tiết high-frequency vẫn chưa đạt mức "siêu phân giải"

Ví dụ: tóc, lông mày, râu đôi khi còn hơi mờ hoặc được vẽ lại một cách "tiêu chuẩn hóa".

(4) Yêu cầu GPU hỗ trợ để đạt tốc độ tối đa

- Mặc dù nhanh, nhưng dùng LCM + các loss phức tạp → vẫn cần GPU mạnh để chạy real-time.

4.2. Góc nhìn riêng, hướng cải tiến và ứng dụng thực tế

4.2.1. Góc nhìn riêng

InterLCM đánh dấu một xu hướng mới:

tối ưu quá trình khôi phục bằng việc kiểm soát tốt trạng thái trung gian thay vì chỉ tối ưu đầu ra.

Đây là một triết lý rất thực dụng: tốc độ nhanh nhưng vẫn giữ được chất lượng như diffusion nhiều bước.
Nội dung hài hòa giữa *tính học thuật (LCM)* và *tính ứng dụng thực tế (BFR real-world)*.

Tuy nhiên, phong cách xử lý theo "template khuôn mặt" vẫn tiềm ẩn nguy cơ làm lệch danh tính — đúng vấn đề mà nhiều mô hình GAN trước đây gặp phải.

4.2.2. Hướng cải tiến

(1) Kết hợp mô hình 3D Face Prior

Dùng 3D Morphable Model (3DMM) để giữ hình học khuôn mặt → tránh méo mặt, giúp nhất quán với ảnh chụp nghiêng hoặc tư thế lạ.

(2) Incorporate Attention from Real Face Banks

Học từ tập các ảnh sắc nét của cùng 1 người (nếu có)
→ BFR giữ danh tính tuyệt đối chính xác, dùng tốt cho forensics.

(3) Dynamic intermediate states

Thay vì cố định loại low-quality intermediate, mô hình có thể học cách chọn mức suy giảm phù hợp cho từng ảnh đầu vào.

(4) Triển khai lightweight version (mobile-friendly)

- Tối ưu lượng params
- Dùng quantization hoặc distillation
  → đưa vào ứng dụng camera real-time trên smartphone.

4.2.3. Ứng dụng thực tế

(1) Khôi phục ảnh gia đình cũ, ảnh kỷ niệm, ảnh scan

Những ảnh bị mờ, nhiễu, nén JPEG, nhòe motion → InterLCM xử lý rất tốt.

(2) Camera giám sát – nâng độ rõ mặt để nhận diện

Dùng được trong:

- an ninh cửa hàng
- smart city
- nhà máy
  → tăng rõ mặt để đối chiếu nhận dạng.

(3) Video call enhancement (Zoom, Teams, Zalo)

Khôi phục khuôn mặt real-time khi mạng yếu, video bị nén.

(4) Digital Forensics – phục hồi khuôn mặt nghi phạm

Hỗ trợ phân tích ảnh chất lượng thấp từ CCTV.

(5) Ứng dụng trong game/AR/VTuber

Tăng chất lượng face-tracking stream để drive character ảo.