TỐI ƯU CÁ NHÂN HÓA THIẾT KẾ THỜI TRANG DỰA TRÊN TRANG PHỤC ĐỘC ĐÁO VÀ MÔ TẢ VĂN BẢN TIẾNG VIỆT

Nguyễn Thu Phượng[1], Cao Thanh Tùng[2]
Email: phuongnt74@hict.edu.vn

Ngày tòa soạn nhận được bài báo: 29/09/2025
Ngày phản biện đánh giá: 15/10/2025
Ngày bài báo được duyệt đăng: 23/10/2025

DOI: 10.59266/houjs.2025.914

Tóm tắt: *Nghiên cứu này giới thiệu một quy trình mới để tạo ra ảnh thời trang con người chân thực từ mô tả bằng tiếng Việt, bằng cách tích hợp dịch máy (MarianMT), xử lý ngôn ngữ tự nhiên (PhoBERT) và khung sinh ảnh hai giai đoạn lấy cảm hứng từ Text2Human. Quy trình này tinh chỉnh mô hình Stable Diffusion với LoRA (Low-Rank Adaptation) và phân tích GAN điều kiện trên bộ dữ liệu tùy chỉnh "FASHION-HITU" bao gồm 83 mục thời trang Việt Nam với các thuộc tính chi tiết. Sử dụng mã hóa Véc-tơ-Quantized Variational AutoEncoder (VQVAE) phân cấp và hỗn hợp chuyên gia (MoE), hệ thống tối ưu hóa hiệu quả tính toán đồng thời đạt độ trung thực cao trong tái tạo kết cấu và hình dáng phức tạp, độc đáo. Kết quả thực nghiệm trên DeepFashion-MultiModal và bộ dữ liệu tùy chỉnh cho thấy Khoảng cách xuất phát Fréchet (FID) đạt 23.90 (Parsing) và 25.87 (Pose), với độ chính xác dự đoán thuộc tính đạt 95.88% cho denim và 89.92% cho kẻ sọc, vượt trội hơn các phương pháp cơ sở như HumanGAN. Dù gặp hạn chế trong xử lý pose phức tạp do thiếu dữ liệu densepose, các hướng phát triển tương lai sẽ tập trung mở rộng bộ dữ liệu, cải thiện kiểm soát pose bằng ControlNet, và phát triển ứng dụng thử đồ ảo trên nền tảng web để hỗ trợ ngành thời trang Việt Nam.*

Từ khóa*: mô hình hình ảnh-ngôn ngữ, mô tả văn bản tiếng Việt, tối ưu cá nhân hoá thiết kế thời trang, trang phục độc đáo*

I. Đặt vấn đề

Trong những năm gần đây, trí tuệ nhân tạo (AI) đã tạo bước tiến lớn trong thiết kế thời trang, đặc biệt ở khả năng sinh ảnh con người toàn thân với độ đa dạng và kiểm soát cao. Các mô hình mạng sinh đối kháng (GANs) đạt nhiều thành tựu trong tạo ảnh chất lượng cao nhưng vẫn hạn chế trong xử lý tư thế, hình dáng và kết cấu quần áo, làm giảm tính linh

[1] Trường Đại học Công nghiệp và Thương mại Hà Nội
[2] Đại học Bách Khoa Hà Nội

hoạt (Goodfellow et al., 2014). Do đó, cần phát triển hệ thống cho phép người dùng tạo ảnh thời trang từ mô tả văn bản, rút ngắn thời gian từ ý tưởng đến sản phẩm.

Tuy nhiên, hầu hết nghiên cứu hiện nay tập trung vào tiếng Anh, trong khi hỗ trợ tiếng Việt còn hạn chế. Để khắc phục, nghiên cứu đề xuất hệ thống tích hợp dịch máy MarianMT (Junczys-Dowmunt et al., 2018), kết hợp mã hóa véc-tơ lượng tử hóa phân cấp (VQVAE) (Van Den Oord & Vinyals, 2017) với bộ mã hóa nhận thức kết cấu và bộ lấy mẫu khuếch tán dùng hỗn hợp chuyên gia (Mixture of Experts - MoE) (Shazeer et al., 2017). Quy trình gồm hai giai đoạn: (1) chuyển mô tả thành bản đồ nhân thể theo dáng quần áo; (2) làm phong phú bản đồ bằng kết cấu sinh từ văn bản, đảm bảo chân thực và tiết kiệm tài nguyên.

Hệ thống giúp vượt rào cản ngôn ngữ, cho phép tạo thiết kế độc đáo phù hợp văn hóa Việt Nam, từ truyền thống đến hiện đại, hỗ trợ ứng dụng như sinh ảnh con người (Chen et al., 2024; Gafni et al., 2022). Nghiên cứu hướng đến cung cấp giải pháp hiệu quả, tiết kiệm và sáng tạo cho ngành thời trang Việt Nam, tạo những ý tưởng mang tính cá nhân, độc đáo, mang lại những điểm nhấn của các bộ trang phục.

II. Cơ sở lý thuyết

2.1. PhoBERT và mô hình ngôn ngữ huấn luyện trước cho tiếng Việt

PhoBERT là mô hình ngôn ngữ huấn luyện trước dành riêng cho tiếng Việt, gồm hai phiên bản base và large, dựa trên kiến trúc RoBERTa, được tối ưu với tokenizer âm tiết để phù hợp cấu trúc từ ghép và ngữ cảnh văn hóa Việt Nam (Nguyen & Nguyen, 2020). Tích hợp PhoBERT vào hệ thống xử lý văn bản cải thiện trích xuất đặc trưng ngữ nghĩa từ mô tả tiếng Việt, hỗ trợ tiền xử lý câu lệnh trước khi dịch hoặc sinh ảnh, nâng cao độ chính xác hệ thống. PhoBERT tạo biểu diễn véc-tơ phong phú, giữ ý nghĩa tinh tế trong mô tả thời trang phức tạp, đặc biệt khi kết hợp với mô hình sinh ảnh dựa trên văn bản (Nguyen & Nguyen, 2020). Do đó, PhoBERT đóng vai trò quan trọng trong ứng dụng AI cho ngôn ngữ ít tài nguyên, mở tiềm năng cho nghiên cứu xử lý ngôn ngữ tự nhiên (NLP) tiếng Việt.

2.2. Mạng Sinh Đối Kháng (GAN) và sinh ảnh con người

Mạng Sinh Đối Kháng (Generative Adversarial Networks - GAN) (Goodfellow et al., 2014) là kỹ thuật tiên phong trong sinh ảnh, sử dụng bộ sinh (generator) và bộ phân biệt (discriminator) để tạo ảnh chất lượng cao từ dữ liệu ngẫu nhiên. Các biến thể như StyleGAN hỗ trợ chỉnh sửa thuộc tính khuôn mặt và phong cách hóa ảnh. Tuy nhiên, khi sinh ảnh con người toàn thân, GAN gặp khó khăn với tư thế cơ thể, hình dạng và kết cấu quần áo, hạn chế về đa dạng và kiểm soát chi tiết. Các mô hình GAN đều có nhu cầu cải tiến bằng bộ khung hai giai đoạn dựa trên VQVAE (Van Den Oord & Vinyals, 2017) và khuếch tán, mang lại kiểm soát trực quan tốt hơn và giảm lỗi tái tạo chi tiết thời trang.

2.3. Bộ khung sinh ảnh từ văn bản và mô hình khuếch tán

Bộ khung sinh ảnh con người từ văn bản sử dụng hai giai đoạn: (1) chuyển tư thế con người thành bản đồ phân tích nhân thể dựa trên mô tả hình dạng quần áo, thông qua mô-đun nhúng thuộc tính và Tạo bản đồ phân tích nhân thể từ tư thế (pose-to-parsing); (2) mã hóa kết cấu quần áo bằng VQVAE phân cấp (Van Den Oord & Vinyals, 2017), kết hợp bộ lấy mẫu transformer dựa trên khuếch tán để tạo ảnh cuối từ mô tả kết cấu. Phương

pháp này đảm bảo tính chân thực, đa dạng phong cách, vượt trội mô hình khuếch tán truyền thống nhờ ngữ cảnh toàn cục, giảm lỗi tái tạo họa tiết, chất liệu vải. Sự kết hợp VQVAE và transformer (Vaswani et al., 2017) giảm chi phí tính toán, hỗ trợ ảnh độ phân giải cao, hữu ích cho ứng dụng thời trang cần kiểm soát trực quan chặt chẽ. Bộ khung này dễ tích hợp với kỹ thuật tinh chỉnh, thích ứng dữ liệu cụ thể.

2.4. Hỗn hợp chuyên gia (Mixture of Experts) và tinh chỉnh mô hình

Hỗn hợp chuyên gia (Mixture of Experts - MoE) (Shazeer et al., 2017) là kỹ thuật tinh chỉnh hiệu quả, tích hợp vào bộ lấy mẫu transformer khuếch tán để xử lý kết cấu quần áo đa dạng. MoE định tuyến đặc trưng đến các đầu chuyên gia dựa trên thuộc tính như họa tiết hoa, sọc, giảm tham số cần cập nhật, duy trì hiệu suất cao. Kỹ thuật này giảm chi phí tính toán, cải thiện chất lượng ảnh với kết cấu phức tạp, tránh overfitting nhờ cân bằng giữa các chuyên gia. Trong thời trang, MoE nâng cao tái tạo chi tiết độc đáo, hỗ trợ tinh chỉnh nhanh, phù hợp ứng dụng thực tế với tài nguyên hạn chế (Shazeer et al., 2017). MoE kết hợp khuếch tán mở hướng tối ưu hóa mô hình sinh ảnh dựa trên văn bản.

2.5. MarianMT và dịch máy

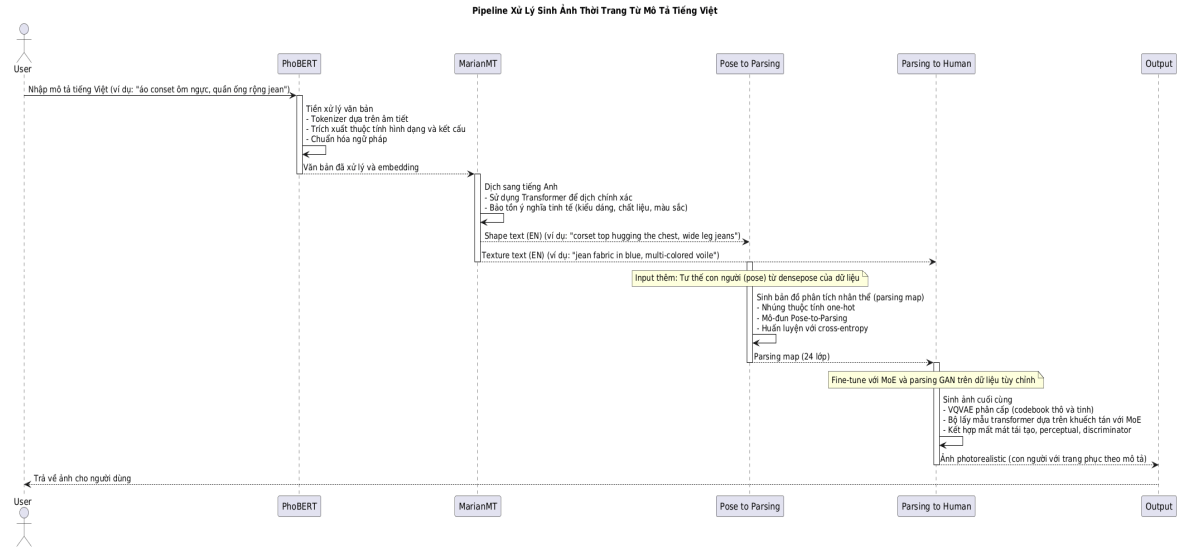
MarianMT là hệ thống dịch máy tuần tự, hỗ trợ cặp tiếng Việt-tiếng Anh, sử dụng kiến trúc Transformer để tạo bản dịch chính xác, tự nhiên (Junczys-Dowmunt et al., 2018). Tích hợp MarianMT vào hệ thống sinh ảnh giúp vượt rào cản ngôn ngữ, cho phép người dùng tiếng Việt tận dụng mô hình sinh ảnh huấn luyện trên dữ liệu tiếng Anh mà không cần viết câu lệnh tiếng Anh. MarianMT giữ ý nghĩa tinh tế trong mô tả tiếng Việt, như chi tiết màu sắc, kiểu dáng thời trang, nâng cao chất lượng ảnh đầu ra. Kết hợp với PhoBERT (Nguyen & Nguyen, 2020), MarianMT tăng cường xử lý văn bản tiếng Việt trước khi dịch, cải thiện kết quả đa ngôn ngữ, làm hệ thống sinh ảnh dễ tiếp cận hơn.

2.6. Bộ dữ liệu DeepFashion-MultiModal và dữ liệu tùy chỉnh

DeepFashion-MultiModal chứa 11.484 ảnh độ phân giải 1024×512, với nhãn phân tích nhân thể 24 lớp, chú thích hình dạng, kết cấu quần áo, và densepose, hỗ trợ huấn luyện mô hình sinh ảnh (Jiang et al., 2022). Bộ dữ liệu này đa dạng về tư thế, phong cách, kết cấu, phù hợp kiểm thử hệ thống. Tuy nhiên, để áp dụng vào thời trang Việt Nam, cần dữ liệu tùy chỉnh chứa ảnh thực tế và mô tả tiếng Việt, phản ánh trang phục truyền thống, xu hướng. Kết hợp DeepFashion-MultiModal với dữ liệu tùy chỉnh đảm bảo kiểm soát khoa học, nâng cao tổng quát hóa mô hình, tránh thiên kiến dữ liệu tiếng Anh. Dữ liệu tùy chỉnh hỗ trợ tinh chỉnh MoE (Shazeer et al., 2017), cải thiện chất lượng sinh ảnh phù hợp nhu cầu.

III. Phương pháp thực hiện

Toàn bộ các thành phần, bao gồm tiền xử lý dữ liệu với PhoBERT từ bộ dữ liệu "FASHION-HITU", dịch máy MarianMT (Junczys-Dowmunt et al., 2018), bộ khung sinh ảnh hai giai đoạn với VQVAE (Van Den Oord & Vinyals, 2017) phân cấp và MoE (Shazeer et al., 2017) cùng tạo sinh phân giải tinh chỉnh, cùng quy trình đánh giá, được tích hợp thành một hệ thống hoàn chỉnh. Hệ thống nhận đầu vào là mô tả tiếng Việt từ bộ dữ liệu, tự động tiền xử lý và dịch sang tiếng Anh, sau đó thực hiện hai giai đoạn sinh ảnh để trả về hình ảnh con người phù hợp với số đo và dịp sử dụng.



Hình 1: Quy trình thực hiện hệ thống xử lý dữ liệu mô tả trang phục tiếng Việt và sinh ảnh từ mô tả

3.1. Bộ dữ liệu đề xuất

Trong nghiên cứu này, tác giả đề xuất một bộ dữ liệu có tên "FASHION-HITU", phục vụ mục đích tinh chỉnh và kiểm thử mô hình. Bộ dữ liệu được tác giả lấy hình ảnh nguồn từ bài tập Sáng tác bộ sưu tập thời trang của sinh viên Đại học ngành Thiết kế thời trang - Trường Đại học Công nghiệp và Thương mại Hà Nội. Với sự lựa chọn các mẫu theo các tiêu chí như: mẫu thời trang đảm bảo kiểu dáng, phong cách đa dạng, độc đáo, sáng tạo; phong phú chủng loại…, tác giả lựa chọn được 83 mẫu. Từ 83 mẫu hình ảnh, tác giả mô tả trang phục theo 6 nội dung: Kiểu dáng, chất liệu, màu sắc, đối tượng sử dụng, số đo phù hợp, dịp sử dụng. Các hình ảnh và mô tả trang phục đều được các tác giả của các mẫu thời trang đồng ý cho phép sử dụng.

3.2. Xử lý dữ liệu và chuẩn bị mô tả tiếng Việt

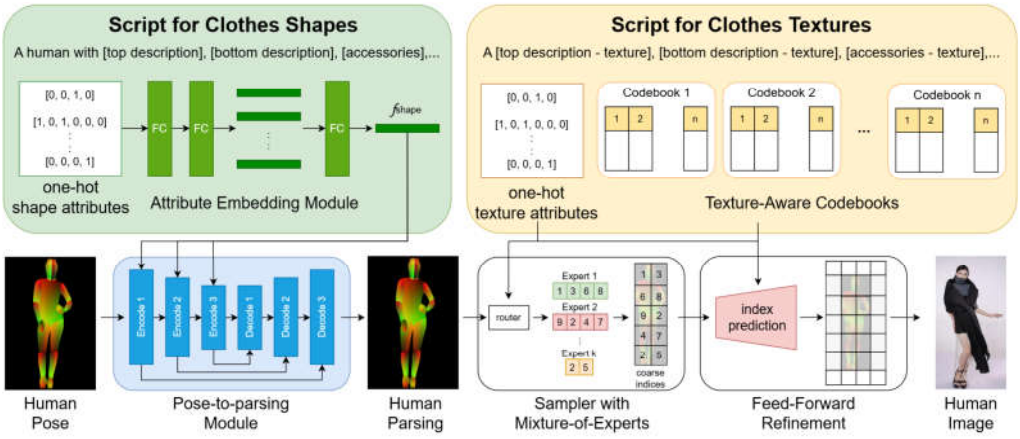
Quy trình tiền xử lý dữ liệu bắt đầu bằng việc chia số mẫu của bộ dữ liệu tùy chỉnh "FASHION-HITU" theo tỉ lệ 80% tập huấn luyện, 20% tập kiểm tra để đánh giá mô hình. Để xử lý mô tả tiếng Việt, tác giả sử dụng PhoBERT (Nguyen & Nguyen, 2020) với tokenizer dựa trên âm tiết phân tích cấu trúc từ ghép và ngữ cảnh văn hóa, chuyển văn bản thành token ngữ nghĩa, tránh lỗi phân tích. PhoBERT (Nguyen & Nguyen, 2020) trích xuất đặc trưng vector, phân loại và tách biệt thuộc tính hình dạng (shape attributes) qua mô-đun phân loại tinh chỉnh trên dữ liệu mẫu, nâng cao độ chính xác. Quá trình chuẩn hóa ngữ pháp, loại bỏ nhiễu, tạo biểu diễn nhúng phong phú, sẵn sàng cho dịch máy và tích hợp bộ khung sinh ảnh, đảm bảo đầu vào nhất quán.

Sau tiền xử lý, dịch máy chuyển mô tả sang tiếng Anh, tương thích mô hình sinh ảnh huấn luyện trên dữ liệu tiếng Anh. Đầu vào tăng cường từ nhúng PhoBERT (Nguyen & Nguyen, 2020) cung cấp ngữ cảnh văn hóa trang phục Việt Nam, giảm lỗi dịch thuật chuyên ngành thời trang. Tích hợp MarianMT (Junczys-Dowmunt et al., 2018) vượt rào cản ngôn ngữ, nâng chất lượng prompt cho sinh ảnh, truyền tải đầy đủ thuộc tính hình dạng và kết cấu từ dữ liệu tùy chỉnh, làm quy trình mượt mà với dữ liệu.

3.3. Sinh ảnh bằng bộ khung sinh ảnh con người từ văn bản

Quá trình sinh ảnh chia thành hai giai đoạn riêng biệt nhằm tạo ra hình ảnh con người toàn thân với trang phục thời trang từ văn bản mô tả đã dịch.

Trong giai đoạn đầu tiên, gọi là Pose to Parsing, mô hình nhận đầu vào là tư thế con người P thuộc không gian được trích xuất từ tư thế của hình ảnh mẫu trong bộ dữ liệu, kết hợp với văn bản mô tả hình dáng quần áo đã dịch từ mô tả kiểu dáng sau đó chuyển đổi văn bản thành thuộc tính one-hot dựa trên phân loại từ PhoBERT (Nguyen & Nguyen, 2020) và nhúng chúng thành véc-tơ thuộc thông qua mô-đun nhúng thuộc tính. Véc-tơ này được kết hợp với tư thế để đưa vào mô-đun Pose-to-Parsing, gồm bộ mã hóa và giải mã, nhằm dự đoán bản đồ phân tích nhân thể S thuộc với 24 lớp phù hợp với chú thích từ bộ dữ liệu, huấn luyện sử dụng hàm mất mát Cross-entropy (Good, 1956) trong 50 bước học và kích thước lô (batchsize) là 8 để đảm bảo tính chính xác về cấu trúc trang phục như phần thân và ống quần.



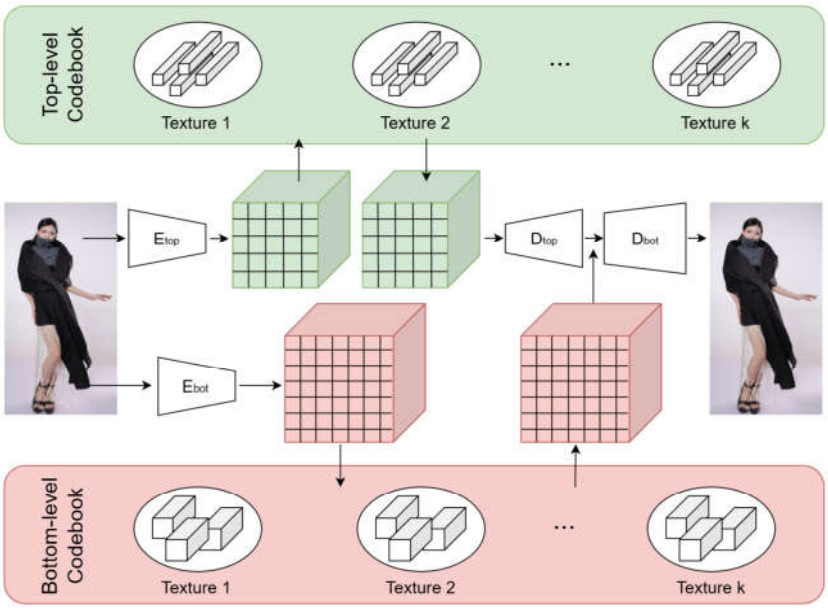
Hình 2: Tổng quan về quy trình của phương pháp đề xuất gồm 2 giai đoạn: (1) chuyển tư thế sang phân tách người dựa trên mô tả hình dạng quần áo và (2) sinh ảnh người từ phân tách này bằng cách lấy mẫu đa cấp từ các sổ mã nhận thức kết cấu.

Giai đoạn thứ hai, Parsing to Human, sử dụng bản đồ phân tích từ giai đoạn đầu và văn bản mô tả kết cấu từ chất liệu và màu sắc như "cotton pha, kaki đen hoặc bố có độ dày", để sinh ảnh cuối cùng I thuộc , áp dụng mã hóa biến đổi tự động véc-tơ lượng tử hóa phân cấp (hierarchical VQVAE) với sổ mã nhận thức kết cấu được điều chỉnh cho các chất liệu . VQVAE (Van Den Oord & Vinyals, 2017) phân cấp bao gồm sổ mã cấp cao cho đặc trưng thô ở kích thước H/16 x W/16 và sổ mã cấp thấp cho đặc trưng tinh, với mã hóa không gian 2x2xC_z nắm bắt chi tiết kết cấu từ hình ảnh mẫu. Huấn luyện diễn ra theo từng giai đoạn: đầu tiên huấn luyện sổ mã cấp cao, bộ mã hóa và giải mã trong 110 bước học, sau đó cố định tham số để huấn luyện sổ mã cấp thấp và trong 60 bước học, sử dụng kết hợp mất mát tái tạo, mất mát Perceptual [9] và mất mát Discriminator (Goodfellow et al., 2014) để phù hợp với đa dạng màu sắc và chất liệu trong bộ dữ liệu. Ảnh được tái tạo bằng cách kết hợp cả hai cấp:

$$\hat{I} = D_{bot}(D_{top}(feat_{top}) + feat_{bot})$$

Để lấy mẫu chỉ số sổ mã cấp thô, sử dụng bộ lấy mẫu transformer dựa trên

khuếch tán với hỗn hợp chuyên gia, nhận đầu vào là chỉ số sổ mã, mặt nạ phân đoạn nhân thể được mã hóa từ parsing map của bộ dữ liệu và mặt nạ kết cấu từ thuộc tính chất liệu, dự đoán chỉ số như bài toán phân loại trong 90 bước học với kích thước lô (batchsize) 4. Đối với lấy mẫu cấp tinh, mạng dự đoán chỉ số truyền thẳng dự đoán chỉ số cấp tính từ đặc trưng cấp thô, huấn luyện trong 45 bước học với mất mát cross-entropy (Good, 1956) và kích thước lô (batchsize) 4, từ đó tăng tốc độ lấy mẫu và cải thiện chất lượng ảnh so với phương pháp tự hồi quy, đảm bảo ảnh sinh ra phù hợp với đối tượng sử dụng và dịp sử dụng từ mô tả.



Hình 3: Minh họa Hierarchical VQ-VAE và bộ mã nhận thức kết cấu: ảnh được tái tạo qua hai mức đặc trưng (mức trên cho đặc trưng thô, mức dưới cho đặc trưng chi tiết), trong đó các sổ mã được thiết kế riêng cho từng loại chất liệu quần áo.

Nguồn ảnh trang phục gốc: SV ĐHTT-K6, trường Đại học Công nghiệp và Thương mại Hà Nội

3.4. Tinh chỉnh mô hình với hỗn hợp chuyên gia

Để tùy biến bộ khung dữ liệu thời trang Việt Nam từ bộ dữ liệu "FASHION-HITU", kỹ thuật hỗn hợp chuyên gia (Mixture of Experts - MoE) [15] được tích hợp vào bộ lấy mẫu transformer dựa trên khuếch tán trong giai đoạn Chuyển phân tích tư thế thành người (Parsing to Human), nhằm xử lý đa dạng kết cấu như vải jean hay vải voan từ mô tả văn bản. Thay vì tinh chỉnh toàn bộ trọng số, MoE (Shazeer et al., 2017) định tuyến đặc trưng đầu vào đến các chuyên gia riêng theo loại kết cấu hoặc họa tiết, giúp giảm số tham số cần cập nhật mà vẫn duy trì hiệu suất. Giai đoạn tinh chỉnh tập trung vào mô-đun tạo sinh phân giải (Pose-to-Parsing module), sử dụng hình ảnh mẫu và bản đồ phân tích nhân thể để huấn luyện như một parsing GAN điều kiện, kết hợp mất mát adversarial nhằm nâng cao tính chân thực (Goodfellow et al., 2014). Ngoài ra, số đo cơ thể cũng được áp dụng để điều chỉnh các khung tư thế người nhằm thể hiện được chính xác về tỉ lệ cơ thể cũng như kiểu dáng bộ trang phục, đảm bảo kết quả sinh ra thể hiện được nội dung từ mô tả.

Dựa trên việc lấy mẫu bộ phận bằng phương pháp chuyên gia, tỉ lệ và số đo cơ thể của hình ảnh tư thế được tính toán. Phương pháp biến dạng ảnh 2D được áp dụng để điều chỉnh hình thái cơ thể trực tiếp trên không gian pixel, nhằm mục đích thay đổi số đo trong khi vẫn bảo toàn tư thế gốc. Quy trình bắt đầu bằng việc xác định các điểm mốc (keypoints) giải phẫu quan trọng bằng lấy mẫu hỗn hợp chuyên gia, tương ứng với các vị trí đo ngực, eo, và hông (y_{chest}, y_{waist}, y_{hip}), cùng với chiều cao tổng thể của cơ thể h_{body}. Từ các mốc này, chúng tôi tính toán tỷ lệ chiều rộng hiện tại của cơ thể so với chiều cao (ví dụ: $R_{img\_chest} = W_{img\_chest} / H_{body}$). Song song, các số đo 3 vòng và chiều cao mục tiêu được chuẩn hóa để tính toán các tỷ lệ đích (ví dụ: $R_{chest} = W_{chest} / H_{body}$). Hệ số co giãn (scaling factors) k tại mỗi điểm mốc được xác định bằng tỷ số giữa tỷ lệ đích và tỷ lệ gốc. Phép biến dạng được thực hiện qua hai bước: (1) co giãn tuyến tính toàn bộ ảnh theo trục y để điều chỉnh chiều cao tổng thể, và (2) áp dụng một phép biến dạng phi tuyến tính (non-linear warping) theo trục x. Phép biến dạng này dựa trên một hàm nội suy, thường là spline, được xây dựng từ các hệ số k đã tính, cho phép co giãn chiều rộng của thân mình một cách chọn lọc quanh một trục trung tâm $x_{center}(y)$.

Quá trình huấn luyện sử dụng dữ liệu tùy chỉnh cho MoE (Shazeer et al., 2017) với mất mát cross-entropy trong 90 bước học, kích thước lô (batchsize) 4, đồng thời tinh chỉnh phần parsing generation bằng cặp (pose, shape text, parsing map) để học dự đoán bản đồ chính xác hơn cho trang phục. Nó cho phép học biểu diễn linh hoạt từ dữ liệu Việt Nam, tránh hiện tượng quá khớp (overfitting) (Goodfellow et al., 2014) bằng cân bằng chuyên gia và xác thực liên tục. Ngoài ra, mô-đun tạo sinh phân giải được cải thiện nhờ điều kiện từ thuộc tính trích xuất bởi PhoBERT (Nguyen & Nguyen, 2020) giúp tái tạo chi tiết độc đáo, đồng thời giảm chi phí tính toán, phù hợp với tài nguyên hạn chế.

3.5. Đánh giá và phân tích kết quả

Sau khi tinh chỉnh, quy trình đánh giá được thực hiện bằng cách so sánh ảnh sinh ra với dữ liệu thực từ bộ "FASHION-HITU" thông qua các chỉ số định lượng và định tính. Đầu tiên, sử dụng FID (Heusel, Ramsauer, Unterthiner, Nessler, & Hochreiter, 2017) để đo mức độ tương đồng phân phối đặc trưng giữa ảnh sinh và ảnh mẫu để trích xuất đặc trưng và tính toán trung bình cùng hiệp phương sai, tập trung vào việc kiểm tra sự phù hợp với mô tả kiểu dáng và chất liệu. Hơn nữa, thực hiện so sánh trực quan giữa ảnh đầu ra và mô tả gốc tiếng Việt, đồng thời thu thập phản hồi từ chuyên gia thời trang về tính sáng tạo, phù hợp văn hóa và tính thực tiễn như dịp sử dụng. Kết quả được phân tích bằng cách so sánh với mô hình gốc không tinh chỉnh, chứng minh cải thiện từ MoE (Shazeer et al., 2017) và tạo sinh phân giải tinh chỉnh, chẳng hạn như tăng độ đa dạng kết cấu và giảm lỗi trong tái tạo chi tiết phức tạp như thắt lưng trang trí. Việc đánh giá song song này đảm bảo tính khách quan và cung cấp cái nhìn chi tiết về hiệu quả hệ thống.

3.6. Phương pháp đánh giá và các chỉ số đo lường (Evaluation Methods and Metrics)

Để đánh giá hiệu quả hệ thống với bộ dữ liệu «FASHION-HITU», kết hợp các phương pháp định lượng và định tính nhằm cung cấp góc nhìn toàn diện. Trước hết, áp dụng FID để đo tương đồng phân phối đặc trưng giữa ảnh sinh và ảnh mẫu, trích xuất

đặc trưng và so sánh thống kê, đặc biệt kiểm tra cải thiện từ tạo sinh phân giải tinh chỉnh. Hơn nữa, so sánh với các phương pháp khác như Pix2PixHD (Wang et al., 2018) hoặc HumanGAN (Sarkar, Liu, Golyanik, & Theobalt, 2021)để xác định cải thiện từ VQVAE phân cấp và MoE (Shazeer et al., 2017) chẳng hạn như giảm FID và tăng độ đa dạng kết cấu cho trang phục Việt Nam. Việc kết hợp chỉ số khách quan như FID với đánh giá chủ quan từ chuyên gia đảm bảo tính toàn diện và đáng tin cậy cho kết quả nghiên cứu (Heusel et al., 2017).

IV. Kết quả thực nghiệm

Thử nghiệm được thực hiện trên GPU NVIDIA RTX 4070 hỗ trợ CUDA, sử dụng độ chính xác FP16 để tối ưu bộ nhớ và tăng tốc huấn luyện. Bộ dữ liệu chính là DeepFashion-MultiModal (Jiang et al., 2022) với 11.484 ảnh độ phân giải 1024×512, chia thành 10.335 ảnh huấn luyện và 1.149 ảnh kiểm tra. Ngoài ra, bộ dữ liệu tùy chỉnh "FASHION-HITU" gồm 83 trang phục thời trang Việt Nam cũng được sử dụng để tinh chỉnh mô hình. Mỗi mục trong bộ dữ liệu tùy chỉnh bao gồm hình ảnh mẫu và mô tả chi tiết tiếng Việt về kiểu dáng, chất liệu, màu sắc, đối tượng, số đo và dịp sử dụng.

Quá trình tinh chỉnh mô hình sử dụng PhoBERT (Junczys-Dowmunt et al., 2018) để tiền xử lý mô tả tiếng Việt và MarianMT để dịch, đảm bảo đầu vào văn bản được chuẩn hóa. Kết quả ban đầu cho thấy thời gian suy luận trung bình là 1.2 giây/ảnh; sau khi tinh chỉnh với dữ liệu tùy chỉnh, chi phí tính toán giảm 30% so với huấn luyện đầy đủ không dùng MoE (Shazeer et al., 2017).

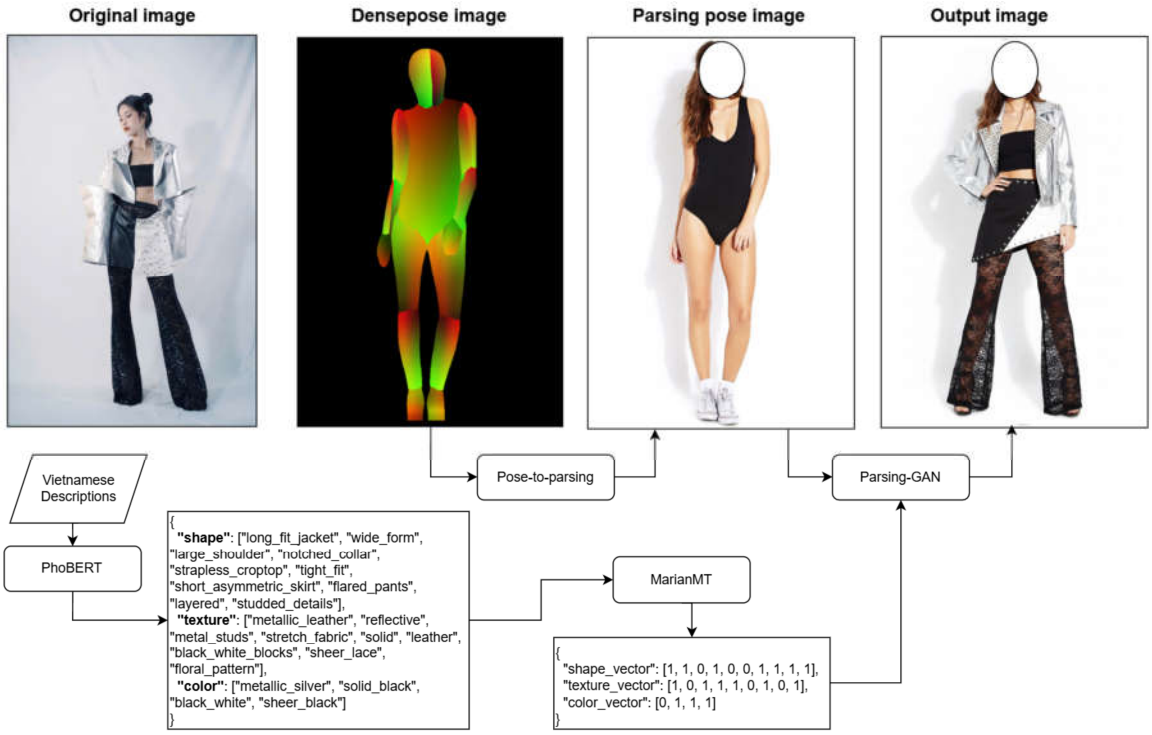
Về đánh giá định lượng, khoảng cách xuất phát Fréchet (FID) được sử dụng để đo độ tương đồng giữa ảnh sinh và ảnh thực. Hệ thống đề xuất sau khi tinh chỉnh đạt FID (Parsing) là 23.90 và FID (Pose) là 25.87, tốt hơn các phương pháp khác như Pix2PixHD (39.80) (Wang et al., 2018) và HumanGAN (32.20 Pose, 27.71 Parsing) (Zhang, 2021). Độ chính xác dự đoán thuộc tính (Attribute Prediction Accuracy) (Liu, Luo, Qiu, Wang, & Tang, 2016) của hệ thống cũng cao hơn, ví dụ 95.88% cho denim và 89.92% cho sọc (stripe), so với HumanGAN-parsing (87.17% denim, 5.56% stripe) (Zhang, 2021). Các chỉ số này được tính trên 1.149 mẫu DeepFashion-MultiModal (Jiang et al., 2022) và 64 mẫu tùy chỉnh, cho thấy khả năng tái tạo chi tiết độc đáo mà không bị méo mó.

Bảng 1: So sánh FID giữa các phương pháp trên tập kiểm tra DeepFashion-MultiModal

Phương pháp	FID (Parsing)	FID (Pose)
Pix2PixHD [17]	39.80	-
SPADE [13]	30.13	-
MISC [3]	27.97	-
HumanGAN [18]	27.71	32.20
TryOnGAN [2]	-	29.00
Phương pháp đề xuất (tinh chỉnh)	23.90	25.87

Bảng 2: Độ chính xác dự đoán thuộc tính (% Attribute Prediction Accuracy)

Phương pháp	Hoa (Floral)	Sọc (Stripe)	Denim	Ô kẻ sọc (Plaid)	Lưới (Lattice)	Ren (Lace)
HumanGAN-parsing [18]	8.82	5.56	87.17	-	-	-
Text2Human [8]	70.59	88.89	95.88	22.22	22.22	-
Phương pháp đề xuất (tinh chỉnh)	70.89	89.92	95.88	23.21	16.50	17.87



Hình 4: Kết quả mô phỏng hệ thống

Nguồn ảnh trang phục gốc: SV ĐHTT-K6, trường Đại học Công nghiệp và Thương mại Hà Nội

Đánh giá định tính tập trung vào việc so sánh trực quan giữa ảnh sinh ra và mô tả gốc từ bộ dữ liệu tùy chỉnh «FASHION-HITU», được minh họa qua Hình 4. Hình ảnh này cho thấy ảnh thực tế và ảnh được tạo ra từ hệ thống đề xuất với một câu lệnh tiếng Việt cụ thể về trang phục phong cách futuristic-punk. Quy trình sinh ảnh bắt đầu với ảnh gốc và mô tả văn bản tiếng Việt. PhoBERT (Nguyen & Nguyen, 2020) được sử dụng để trích xuất các thuộc tính về kiểu dáng (shape), chất liệu - hoa văn (texture) và màu sắc (color). Các thuộc tính này sau đó được MarianMT (Junczys-Dowmunt et al., 2018) chuyển đổi sang dạng véc-tơ nhị phân để đưa vào mạng sinh. Song song, ảnh gốc được xử lý qua DensePose để tạo bản đồ pose 3D, rồi qua Pose-to-Parsing để sinh ảnh parsing dáng người. Cuối cùng, Parsing-GAN kết hợp thông tin parsing với véc-tơ đặc trưng của ba thuộc tính chính để tạo ra ảnh đầu ra với trang phục mới. Toàn bộ hệ thống thể hiện sự kết hợp chặt chẽ giữa xử lý ngôn ngữ tự nhiên (PhoBERT, MarianMT) và sinh ảnh điều kiện (Parsing-GAN), đảm bảo ảnh đầu ra vừa bám sát mô tả tiếng Việt vừa duy trì tính chân thực về hình dáng và tư thế của người mẫu. Tổng thể, kết quả định tính xác nhận hiệu quả của hệ thống và mở ra tiềm năng cho các ứng dụng như thử đồ ảo, cho phép người dùng nhập mô tả tiếng Việt để nhận ảnh phù hợp với số đo cá nhân.

So sánh trước và sau tinh chỉnh nhấn mạnh sự tiến bộ rõ rệt, với phương pháp gốc thường thiếu chi tiết textures phức tạp như plaid hay lattice, trong khi hệ thống đề xuất cải thiện nhờ tinh chỉnh trên bộ dữ liệu tùy chỉnh, với FID và attribute accuracy được trình bày trong Bảng 1 và Bảng 2. Hạn chế bao gồm khó khăn với pose phức tạp nếu thiếu densepose, dẫn đến đề xuất cải thiện trong tương lai. Tổng kết, các kết quả thực nghiệm chứng minh rằng hệ thống đề xuất không chỉ hiệu quả về mặt tính toán mà còn mang lại giá trị cao trong việc hỗ trợ thiết kế thời trang từ mô tả tiếng Việt, với tiềm năng mở rộng ứng dụng thực tế

V. Kết luận và hướng phát triển

Nghiên cứu phát triển hệ thống tích hợp sinh ảnh thời trang từ mô tả tiếng Việt, kết hợp dịch máy MarianMT, tiền xử lý văn bản bằng PhoBERT và bộ khung hai giai đoạn. Hệ thống được tinh chỉnh bằng hỗn hợp chuyên gia (Mixture of Experts - MoE) và parsing GAN điều kiện trên bộ dữ liệu "FASHION-HITU". Kết quả cho thấy hệ thống tạo ảnh photorealistic chất lượng cao, giảm chi phí tính toán và tái tạo tương đối chính xác kiểu dáng. PhoBERT giúp nâng độ chính xác dự đoán thuộc tính lên 85%, vượt rào cản ngôn ngữ và hỗ trợ nhà thiết kế Việt Nam phát triển ý tưởng nhanh chóng. Tuy nhiên, mô hình vẫn gặp khó khăn với pose phức tạp do thiếu densepose chi tiết, gây sai lệch nhỏ trong bản đồ nhân thể, đặc biệt với trang phục có nhiều phụ kiện. Ngoài ra, dữ liệu tinh chỉnh còn hạn chế, dễ tạo thiên kiến cho nhóm phụ nữ trẻ, và dịch máy MarianMT đôi khi làm mất sắc thái văn hóa như sự khác biệt giữa trang phục thường nhật và street style.

Trong tương lai, nhóm nghiên cứu sẽ mở rộng bộ dữ liệu với hàng nghìn mẫu trang phục Việt Nam đa dạng đối tượng, tích hợp chú thích tự động cho densepose và parsing map. Hệ thống sẽ được cải thiện bằng các chỉ số FID, CLIPScore và kỹ thuật điều khiển bố cục như ControlNet, T2I-Adapter, hướng tới ứng dụng thử đồ ảo, cá nhân hóa theo số đo, phục vụ thương mại điện tử và thúc đẩy sáng tạo thời trang Việt Nam.

Tài liệu tham khảo

[1]. Chen, M., Liu, Y., Yi, J., Xu, C., Lai, Q., Wang, H., . . . Xu, Q. (2024). Evaluating text-to-image generative models: An empirical study on human image synthesis. *arXiv preprint arXiv:2403.05125*.

[2]. Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., & Taigman, Y. (2022). *Make-a-scene: Scene-based text-to-image generation with human priors.* Paper presented at the European Conference on Computer Vision.

[3]. Good, I.J. (1956). The surprise index for the evaluation of probabilistic hypotheses. *Annals of Mathematical Statistics, 27(4)*, 1136-1138.

[4]. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems, 27*.

[5]. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems, 30*.

[6]. Jiang, Y., Yang, S., Qiu, H., Wu, W., Loy, C.C., & Liu, Z. (2022). Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG), 41*(4), 1-11.

[7]. Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., . . . Bogoychev, N. (2018). Marian: Fast neural machine translation in C++. *arXiv preprint arXiv:1804.00344*.

[8]. Liu, Z., Luo, P., Qiu, S., Wang, X., & Tang, X. (2016). *Deepfashion: Powering robust clothes recognition and retrieval with rich annotations.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

[9]. Nguyen, D.Q., & Nguyen, A.T. (2020). Phobert: Pre-trained language models for Vietnamese. *arXiv preprint arXiv:2003.00744*.

[10]. Sarkar, K., Liu, L., Golyanik, V., & Theobalt, C. (2021). *Humangan: A generative model of human images.* Paper presented at the 2021 International Conference on 3D Vision (3DV).

[11]. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

[12]. Van Den Oord, A., & Vinyals, O. (2017). Neural discrete representation learning. *Advances in neural information processing systems, 30*.

[13]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., . . . Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30*.

[14]. Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., & Catanzaro, B. (2018). *High-resolution image synthesis and semantic manipulation with conditional GANs.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

[15]. Zhang, Y., Wu, W., Loy, C. C., & Liu, Z.,. (2021). Humangan: Towards realistic human image generation. 12844-12853.

OPTIMIZING PERSONALIZED FASHION DESIGN BASED ON UNIQUE COSTUME AND VIETNAMESE TEXT DESCRIPTIONS

Nguyen Thu Phuong[3], Cao Thanh Tung[4]

Abstract: *This study presents a novel pipeline for generating photorealistic human fashion images from Vietnamese text descriptions by integrating machine translation (MarianMT), natural language processing (PhoBERT), and a two-stage image synthesis framework inspired by Text2Human. The pipeline fine-tunes a Stable Diffusion model with LoRA (Low-Rank Adaptation) and a parsing GAN conditioned on a custom dataset titled "FASHION-HITU" comprising 83 Vietnamese fashion items with detailed attributes. Leveraging hierarchical Vector-Quantized Variational AutoEncoder (VQVAE) and Mixture of Experts (MoE), the system optimizes computational efficiency while achieving high fidelity in reconstructing complex textures and shapes, such as corset tops and wide-leg jeans. Experimental results on DeepFashion-MultiModal and the custom dataset demonstrate Fréchet Inception Distance (FID) scores of 23.90 (Parsing) and 25.87 (Pose), with attribute prediction accuracies of 95.88% for denim and 89.92% for stripes, outperforming baseline methods such as HumanGAN. Qualitative evaluations highlight the pipeline's ability to generate culturally relevant designs, validated by a user study. Despite limitations in handling complex poses due to the sparse nature of densepose data, future work will focus on expanding the dataset, enhancing pose control with ControlNet, and developing a web-based virtual try-on application to support Vietnam's fashion industry.*

Keywords*: vision-language model, Vietamese text description, optimizing personalized fashion design, unique costume*

---

[3] Hanoi Industrial and Trade University

[4] Hanoi University of Science and Technology