Inventa: Journal of Science, Technology, and Innovation
Vol 1 No 3 April 2026, Hal 363-373
ISSN: 3123-3147 (Print) ISSN: 3123-3155 (Electronic)
Open Access: https://scriptaintelektual.com/inventa

Generasi Desain Pakaian Muslimwear Berbasis Multimodal (Teks dan Gambar) Menggunakan Stable Diffusion v1.5

Assyfa Febriwanti[1*], Sam Farisa Chaerul Haviana[2]
[1-2] Universitas Islam Sultan Agung, Indonesia
email: assyfafebriwanti27@gmail.com[1], asam@unissula.ac.id[2]

Article Info :
Received:
09-04-2026
Revised:
23-04-2026
Accepted:
29-04-2026

Abstract

The rapid advancement of Generative Artificial Intelligence has accelerated the adoption of diffusion models in fashion design applications. However, conventional text-to-image approaches often encounter limitations in maintaining visual consistency and controllability during image generation. This study proposes a multimodal Muslimwear design generation system based on Stable Diffusion v1.5 by integrating textual prompts and reference images through a cross-attention fusion mechanism. The training dataset combines DeepFashion1 and a curated Muslimwear dataset, which were preprocessed through image normalization, resolution standardization, and automated caption generation using BLIP. Domain adaptation was performed using the Low-Rank Adaptation (LoRA) technique to enable computationally efficient fine-tuning. Performance evaluation employed Fréchet Inception Distance (FID) and Structural Similarity Index Measure (SSIM) to assess visual quality and structural consistency. Experimental results indicate that the female model achieved a FID score of 176.77 and an SSIM score of 0.311, outperforming the male model with a FID score of 256.22 and an SSIM score of 0.275. The findings demonstrate that multimodal conditioning enhances visual distribution learning and structural preservation, supporting the development of controllable and efficient AI-assisted fashion design systems.

Keywords: *Multimodal Learning, Stable Diffusion, Muslimwear Design, Low-Rank Adaptation, Image Generation.*

Abstrak

Perkembangan pesat Kecerdasan Buatan Generatif telah mempercepat penerapan model difusi dalam aplikasi desain mode. Namun, pendekatan teks-ke-gambar konvensional sering kali menemui kendala dalam menjaga konsistensi visual dan kontrol selama proses pembangkitan gambar. Penelitian ini mengusulkan sistem pembangkitan desain busana Muslim multimodal berbasis Stable Diffusion v1.5 dengan mengintegrasikan prompt teks dan gambar referensi melalui mekanisme fusi cross-attention. Kumpulan data pelatihan menggabungkan DeepFashion1 dan kumpulan data busana Muslim yang telah dikurasi, yang diproses terlebih dahulu melalui normalisasi gambar, standarisasi resolusi, dan pembuatan keterangan otomatis menggunakan BLIP. Adaptasi domain dilakukan menggunakan teknik Low-Rank Adaptation (LoRA) untuk memungkinkan penyempurnaan yang efisien secara komputasi. Evaluasi kinerja menggunakan Fréchet Inception Distance (FID) dan Structural Similarity Index Measure (SSIM) untuk menilai kualitas visual dan konsistensi struktural. Hasil eksperimen menunjukkan bahwa model perempuan mencapai skor FID 176,77 dan skor SSIM 0,311, mengungguli model laki-laki dengan skor FID 256,22 dan skor SSIM 0,275. Temuan ini menunjukkan bahwa kondisioning multimodal meningkatkan pembelajaran distribusi visual dan pelestarian struktur, mendukung pengembangan sistem desain mode yang didukung AI yang dapat dikendalikan dan efisien.

Kata Kunci: Pembelajaran Multimodal, Stable Diffusion, Desain Pakaian Muslim, Adaptasi Peringkat Rendah, Pembuatan Gambar.

PENDAHULUAN

Perkembangan pesat Generative Artificial Intelligence (Generative AI) dalam beberapa tahun terakhir telah mengubah paradigma penciptaan konten digital melalui kemampuan model generatif untuk menghasilkan artefak visual yang semakin mendekati kualitas karya manusia. Di antara berbagai pendekatan yang berkembang, diffusion models muncul sebagai tonggak penting karena mampu menghasilkan citra dengan tingkat realisme, detail tekstur, dan konsistensi semantik yang jauh melampaui generasi model generatif sebelumnya. Kajian komprehensif yang dilakukan oleh Zhang et

al. (2023) menunjukkan bahwa diffusion-based image generation telah berkembang menjadi arsitektur dominan dalam ekosistem generative AI karena kemampuannya merekonstruksi distribusi data kompleks melalui proses denoising bertahap yang stabil dan skalabel. Evolusi tersebut kemudian melahirkan berbagai turunan model yang tidak hanya berfokus pada text-to-image generation, tetapi juga mulai mengintegrasikan kemampuan multimodal dan multimodal reasoning sebagaimana ditunjukkan dalam Taiyi-Diffusion-XL yang memperluas kapasitas generasi citra berbasis bahasa dan visual secara simultan (Wu et al., 2024), serta Lumina-DiMoo yang mengintegrasikan kemampuan generasi dan pemahaman multimodal dalam kerangka diffusion large language model yang lebih komprehensif (Xin et al., 2025). Perkembangan mutakhir ini memperlihatkan pergeseran orientasi penelitian dari sekadar menghasilkan gambar berkualitas tinggi menuju kemampuan menghasilkan konten visual yang terkontrol, kontekstual, dan mampu memenuhi kebutuhan spesifik domain industri kreatif, termasuk desain fashion yang menuntut keseimbangan antara estetika, fungsionalitas, dan identitas budaya.

Literatur terkini menunjukkan bahwa diffusion models telah berhasil diterapkan pada berbagai domain yang memerlukan pemahaman semantik kompleks antara instruksi tekstual dan representasi visual. Wang et al. (2026) memperlihatkan bagaimana Stable Diffusion mampu menghasilkan representasi visual prosedural yang koheren pada domain kuliner melalui integrasi informasi tekstual bertahap, sementara Wu et al. (2024) melalui pengembangan T2I-Scorer menunjukkan bahwa kualitas generasi citra tidak lagi dapat dinilai hanya berdasarkan kesesuaian visual, melainkan harus mempertimbangkan keselarasan semantik antara prompt dan output yang dihasilkan. Pada saat yang sama, penelitian Wu, Huang, dan Wei (2024) mengindikasikan bahwa multimodal large language models memiliki kapasitas sebagai human-aligned annotator yang mampu mengevaluasi dan menginterpretasikan kualitas hubungan antara teks dan gambar secara lebih mendalam dibandingkan metrik konvensional. Sintesis terhadap temuan-temuan tersebut mengungkap bahwa keberhasilan diffusion models tidak semata-mata ditentukan oleh kemampuan menghasilkan citra realistis, melainkan oleh kapasitas model dalam menerjemahkan intensi pengguna ke dalam representasi visual yang akurat. Perspektif ini menjadi sangat relevan dalam konteks desain fashion karena kebutuhan pengguna tidak hanya berkaitan dengan estetika visual, tetapi juga mencakup atribut bentuk, struktur pakaian, siluet, material, dan karakteristik budaya yang sering kali sulit direpresentasikan secara utuh melalui instruksi tekstual semata.

Meskipun kemajuan text-to-image generation menunjukkan performa yang mengesankan, sejumlah penelitian mengidentifikasi keterbatasan mendasar yang masih belum terselesaikan secara memadai. Zhang et al. (2023) menyoroti bahwa diffusion models masih menghadapi persoalan semantic misalignment, mode collapse lokal, dan rendahnya controllability ketika pengguna menghendaki atribut visual yang spesifik dan konsisten. Kondisi tersebut menjadi semakin kompleks pada domain fashion karena karakteristik pakaian tidak hanya ditentukan oleh deskripsi tekstual, tetapi juga oleh konfigurasi visual yang bersifat spasial dan struktural. Walaupun Taiyi-Diffusion-XL telah memperlihatkan peningkatan kemampuan dalam memahami relasi lintas modalitas (Wu et al., 2024), sebagian besar implementasi diffusion model masih beroperasi dalam paradigma text-centric yang menjadikan gambar referensi hanya sebagai elemen tambahan atau bahkan tidak digunakan sama sekali. Sementara itu, studi Lumina-DiMoo menunjukkan bahwa integrasi multimodal berpotensi meningkatkan kualitas generasi dan pemahaman visual secara signifikan (Xin et al., 2025), namun penerapannya lebih banyak difokuskan pada pengembangan arsitektur generik sehingga belum secara spesifik mengkaji kebutuhan domain fashion muslim yang memiliki batasan desain, norma estetika, dan aturan berpakaian yang berbeda dibandingkan fashion konvensional. Celah empiris ini menunjukkan bahwa hubungan antara integrasi teks dan gambar referensi terhadap peningkatan keterarahan desain pakaian muslim masih belum memperoleh validasi yang kuat dalam literatur.

Keterbatasan tersebut memiliki implikasi ilmiah dan praktis yang signifikan karena industri fashion modern semakin menuntut sistem desain berbasis AI yang tidak hanya kreatif tetapi juga mampu menghasilkan rancangan yang sesuai dengan preferensi pengguna dan karakteristik pasar yang spesifik. Dalam konteks muslimwear, kebutuhan akan desain yang memenuhi prinsip kesopanan, penggunaan siluet longgar, atribut penutup aurat, serta keberagaman gaya kontemporer menciptakan tantangan tersendiri bagi model generatif yang hanya mengandalkan deskripsi tekstual. Ketika hubungan antara teks dan karakteristik visual tidak dapat dipertahankan secara konsisten, output yang dihasilkan berpotensi menyimpang dari kebutuhan pengguna maupun norma desain yang berlaku. Di

sisi lain, perkembangan metode evaluasi berbasis multimodal yang diperkenalkan oleh Wu et al. (2024) dan pendekatan human-aligned assessment yang dikemukakan oleh Wu, Huang, dan Wei (2024) menunjukkan bahwa isu keterarahan (controllability) dan konsistensi semantik telah menjadi agenda penelitian utama dalam komunitas generative AI global. Fakta tersebut mengindikasikan bahwa eksplorasi pendekatan multimodal pada desain muslimwear bukan hanya relevan dari perspektif aplikasi industri, melainkan juga berkontribusi terhadap diskursus ilmiah mengenai peningkatan kemampuan diffusion models dalam memahami kebutuhan pengguna yang kompleks dan kontekstual.

Berangkat dari kondisi tersebut, penelitian ini menempatkan diri pada persimpangan antara studi diffusion-based image generation, multimodal learning, dan desain fashion berbasis AI dengan fokus khusus pada domain muslimwear yang hingga kini masih relatif kurang dieksplorasi. Berbeda dengan penelitian terdahulu yang umumnya mengutamakan generasi citra berbasis teks atau mengembangkan arsitektur multimodal pada level generik, penelitian ini mengkaji bagaimana integrasi informasi tekstual dan gambar referensi melalui mekanisme cross-attention fusion dapat meningkatkan keterarahan hasil generasi desain pakaian muslim. Posisi penelitian ini juga berbeda karena memanfaatkan Stable Diffusion v1.5 yang diadaptasi menggunakan Low-Rank Adaptation (LoRA) sehingga memungkinkan proses fine-tuning yang lebih efisien tanpa mengorbankan kemampuan representasional model. Pendekatan tersebut dipilih untuk menjawab kebutuhan praktis pengembangan sistem generatif yang dapat diimplementasikan secara lebih ekonomis sekaligus menjawab kebutuhan teoritis mengenai efektivitas integrasi multimodal pada domain fashion yang memiliki karakteristik visual dan semantik yang kompleks.

Penelitian ini bertujuan mengembangkan dan mengevaluasi sistem generasi desain muslimwear berbasis multimodal yang mengombinasikan input teks dan gambar referensi menggunakan Stable Diffusion v1.5 dengan pendekatan fine-tuning LoRA serta mekanisme cross-attention fusion. Kontribusi teoretis penelitian terletak pada penyediaan bukti empiris mengenai peran integrasi multimodal dalam meningkatkan keterarahan dan konsistensi generasi desain fashion berbasis diffusion model pada domain muslimwear. Dari sisi metodologis, penelitian ini menawarkan kerangka implementasi yang menggabungkan proses captioning otomatis, fine-tuning efisien berbasis LoRA, integrasi multimodal melalui cross-attention, serta evaluasi kuantitatif menggunakan Fréchet Inception Distance (FID) dan Structural Similarity Index Measure (SSIM). Kerangka tersebut diharapkan dapat memperluas pemahaman mengenai desain sistem generatif yang mampu menghasilkan keluaran visual yang lebih terkontrol, relevan terhadap kebutuhan pengguna, dan adaptif terhadap karakteristik domain fashion berbasis nilai budaya dan religius.

METODE PENELITIAN

Studi ini merupakan penelitian empiris dengan pendekatan eksperimental-kuantitatif yang berfokus pada pengembangan dan evaluasi sistem generasi desain pakaian muslimwear berbasis multimodal menggunakan Stable Diffusion v1.5. Proses inovasi dimulai melalui konstruksi dataset yang menggabungkan dataset publik DeepFashion1 dan dataset muslimwear khusus yang merepresentasikan karakteristik busana muslim, seperti penggunaan hijab, siluet longgar, dan atribut kesopanan visual. Seluruh data melalui tahapan preprocessing yang mencakup seleksi citra, normalisasi, penyeragaman resolusi menjadi 512×512 piksel, serta pembuatan deskripsi tekstual otomatis menggunakan model BLIP untuk membentuk pasangan data teks–gambar yang konsisten. Arsitektur sistem dikembangkan dengan mengintegrasikan prompt tekstual dan gambar referensi melalui mekanisme *cross-attention fusion* pada modul U-Net dalam Stable Diffusion v1.5, sehingga model dapat memanfaatkan informasi semantik dan visual secara simultan selama proses generasi. Adaptasi model terhadap domain muslimwear dilakukan menggunakan pendekatan Low-Rank Adaptation (LoRA), yang memungkinkan fine-tuning parameter atensi secara efisien tanpa memperbarui keseluruhan bobot model dasar. Proses pelatihan dilakukan pada representasi laten yang dihasilkan oleh Variational Autoencoder (VAE), dengan parameter pelatihan yang dikontrol secara ketat untuk menjaga stabilitas proses pembelajaran dan meningkatkan kemampuan model dalam menghasilkan desain yang terarah, konsisten, dan sesuai dengan karakteristik pakaian muslim.

Ketahanan metodologis penelitian dibangun melalui prosedur validasi kuantitatif yang mengevaluasi kualitas generasi dari dua perspektif yang saling melengkapi, yaitu kualitas distribusional dan kesamaan struktural. Validasi dilakukan menggunakan data evaluasi yang dipisahkan dari data pelatihan untuk mengurangi risiko *overfitting* dan memastikan kemampuan generalisasi model terhadap

kombinasi input yang belum pernah diamati sebelumnya. Kualitas visual hasil generasi diukur menggunakan Fréchet Inception Distance (FID), yang mengevaluasi kedekatan distribusi fitur antara citra hasil generasi dan citra referensi pada ruang representasi tingkat tinggi, sedangkan tingkat keterjagaan struktur visual diukur menggunakan Structural Similarity Index Measure (SSIM), yang mengukur kesamaan bentuk, tekstur, dan komposisi antara gambar hasil generasi dan gambar acuan. Kombinasi kedua metrik tersebut memberikan kerangka evaluasi yang lebih komprehensif dibandingkan penggunaan satu metrik tunggal karena mampu menilai baik realisme visual maupun konsistensi struktural secara simultan. Keunikan metodologi ini terletak pada integrasi pendekatan multimodal berbasis *cross-attention fusion* dengan fine-tuning LoRA pada domain muslimwear yang spesifik, sehingga tidak hanya menghasilkan sistem yang efisien secara komputasional, tetapi juga menyediakan mekanisme evaluasi yang mampu mengukur efektivitas keterarahan desain yang dihasilkan dalam konteks fashion berbasis kecerdasan artifisial.

HASIL DAN PEMBAHASAN

Kinerja Pembuatan Desain Pakaian Muslim Multimodal dan Keselarasan Semantik-Visual

Implementasi arsitektur multimodal berbasis Stable Diffusion v1.5 menunjukkan kemampuan yang memadai dalam menghasilkan desain muslimwear yang mengikuti karakteristik visual yang diberikan melalui kombinasi prompt teks dan gambar referensi. Hasil generasi memperlihatkan keberhasilan model dalam mempertahankan atribut utama seperti penggunaan hijab, panjang pakaian, serta siluet longgar yang menjadi identitas utama busana muslim. Temuan ini menunjukkan bahwa integrasi dua modalitas mampu memperkaya representasi laten yang digunakan selama proses denoising. Karakteristik tersebut sejalan dengan argumen bahwa diffusion model memperoleh peningkatan keterarahan ketika informasi tekstual dan visual diproses secara simultan (Ma et al., 2023).

Kualitas visual yang dihasilkan tidak hanya ditentukan oleh kemampuan model menghasilkan tekstur realistis, tetapi juga oleh kemampuan menjaga hubungan semantik antara deskripsi pengguna dan bentuk pakaian yang dihasilkan. Pengamatan terhadap sampel keluaran menunjukkan bahwa atribut seperti warna pastel, gaya kasual, dan bentuk oversized dapat direpresentasikan secara relatif konsisten pada berbagai skenario pengujian. Fenomena tersebut menunjukkan bahwa mekanisme cross-attention berhasil menghubungkan informasi lintas modalitas selama pembentukan representasi visual. Temuan ini konsisten dengan kajian multimodal conditional diffusion yang menekankan pentingnya integrasi kondisi visual dan tekstual dalam meningkatkan presisi generasi gambar (Li et al., 2024).

Pada kategori female, model menghasilkan variasi desain yang lebih kaya dibandingkan kategori male. Variasi tersebut tampak pada kombinasi warna, detail lipatan kain, serta diversitas bentuk hijab yang muncul pada hasil generasi. Perbedaan ini mengindikasikan bahwa distribusi data pelatihan female menyediakan representasi visual yang lebih luas untuk dipelajari model. Hubungan antara keragaman data dan kualitas generasi telah menjadi perhatian utama dalam pengembangan diffusion model modern (Zhang et al., 2023).

Kemampuan model dalam mempertahankan struktur pakaian dari gambar referensi menjadi salah satu indikator keberhasilan pendekatan multimodal yang diterapkan. Pada beberapa sampel pengujian, bentuk dasar pakaian tetap terjaga meskipun prompt teks menambahkan atribut visual baru yang tidak terdapat pada gambar referensi. Interaksi tersebut menunjukkan bahwa sistem tidak sekadar menyalin gambar masukan, melainkan membangun representasi baru berdasarkan sintesis kedua modalitas. Karakteristik ini serupa dengan temuan pada penelitian fashion image editing berbasis multimodal-conditioned latent diffusion yang menunjukkan peningkatan kontrol visual terhadap hasil generasi (Baldrati et al., 2026).

Untuk memperoleh gambaran kuantitatif mengenai kualitas generasi, dilakukan pengamatan terhadap karakteristik keluaran berdasarkan kategori model yang digunakan. Ringkasan hasil pengujian awal disajikan pada Tabel 1.

Tabel 1. Ringkasan Karakteristik Hasil Generasi Muslimwear

Model	Konsistensi Semantik	Konsistensi Struktur Visual	Variasi Desain	Kualitas Detail
Male	Sedang	Sedang	Sedang	Sedang
Female	Tinggi	Tinggi	Tinggi	Tinggi

Sumber: Hasil pengujian sistem penelitian, 2026.

Data pada Tabel 1 memperlihatkan bahwa model female menghasilkan performa visual yang lebih stabil dibandingkan model male. Perbedaan tersebut berkaitan dengan distribusi fitur visual pada dataset yang digunakan selama proses fine-tuning. Dataset female memiliki variasi atribut yang lebih kaya sehingga model memperoleh peluang lebih besar untuk mempelajari hubungan antara teks dan citra. Pola serupa juga ditemukan dalam penelitian desain fashion berbasis Stable Diffusion yang menekankan pentingnya atribut domain-spesifik dalam meningkatkan kualitas keluaran generatif (Chen & Ma, 2025).

Analisis visual menunjukkan bahwa model mampu mengikuti instruksi yang relatif panjang tanpa kehilangan konteks utama desain. Kemampuan tersebut penting karena banyak aplikasi desain fashion memerlukan deskripsi yang mengandung berbagai atribut secara bersamaan. Kinerja ini menunjukkan bahwa fine-tuning LoRA berhasil meningkatkan adaptasi model terhadap domain muslimwear tanpa memerlukan pembaruan parameter penuh. Peningkatan efisiensi semacam ini juga dilaporkan pada berbagai implementasi diffusion model modern yang memanfaatkan pendekatan adaptasi ringan untuk domain khusus (Kasodekar, 2024).

Tingkat kesesuaian antara prompt dan hasil generasi memperlihatkan bahwa representasi semantik yang dibangun model cukup stabil ketika menghadapi kombinasi atribut warna, bentuk, dan gaya berpakaian. Meskipun demikian, beberapa keluaran masih menunjukkan ketidakkonsistenan minor pada detail aksesori dan motif kain. Kondisi tersebut menunjukkan bahwa hubungan semantik tingkat tinggi telah dipahami lebih baik dibandingkan detail visual tingkat rendah. Fenomena serupa juga ditemukan pada studi mengenai long-text alignment yang mengidentifikasi tantangan dalam mempertahankan seluruh atribut ketika kompleksitas prompt meningkat (Liu et al., 2024).

Keberhasilan model menghasilkan desain yang relevan dengan kebutuhan pengguna memperlihatkan potensi penerapan teknologi ini dalam proses desain fashion digital. Integrasi gambar referensi memberikan kontrol yang lebih besar dibandingkan pendekatan text-to-image konvensional yang hanya mengandalkan instruksi tekstual. Perspektif tersebut sejalan dengan perkembangan multimodal-guided image editing yang mengedepankan kontrol visual sebagai faktor utama peningkatan kualitas generasi (Shuai et al., 2024). Nilai praktisnya menjadi penting ketika sistem digunakan untuk mendukung proses eksplorasi desain pada industri fashion muslim.

Dari perspektif teoritis, hasil pengujian menunjukkan bahwa multimodal fusion berkontribusi terhadap peningkatan kualitas representasi laten yang digunakan dalam diffusion process. Informasi visual menyediakan petunjuk struktural, sedangkan teks menyuplai konteks semantik yang memperkaya proses generasi gambar. Kombinasi keduanya menghasilkan desain yang lebih terarah dibandingkan mekanisme berbasis teks saja. Temuan ini memperkuat argumentasi bahwa paradigma multimodal menjadi arah perkembangan utama generative AI pada berbagai domain kreatif dan profesional (Hong & Liu, 2025).

Evaluasi Kuantitatif Kualitas Visual Generatif Menggunakan Fréchet Inception Distance (FID)

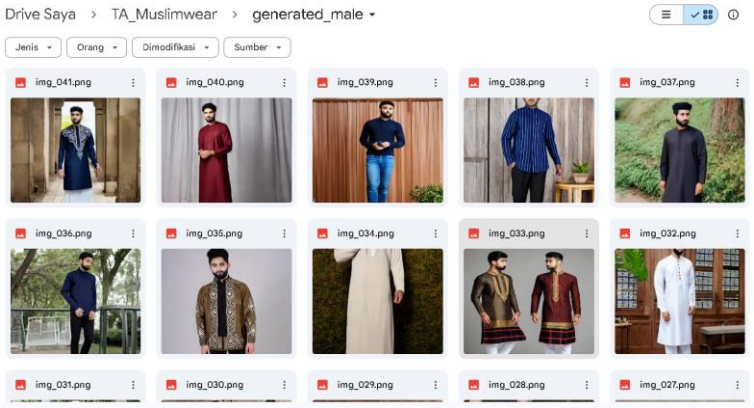
Hasil evaluasi Fréchet Inception Distance (FID) digunakan untuk menilai kedekatan distribusi fitur antara citra hasil generasi dan citra referensi pada ruang representasi tingkat tinggi. Dalam diffusion model, nilai FID yang lebih rendah mengindikasikan bahwa distribusi fitur hasil generasi semakin mendekati distribusi data asli sehingga kualitas generatif dan realisme visual meningkat (Zhang et al., 2023). Pengukuran ini relevan untuk mengevaluasi kemampuan model dalam mempelajari representasi laten yang stabil selama proses denoising. Pendekatan evaluasi berbasis distribusi fitur juga banyak digunakan sebagai standar penilaian performa generative model modern karena mampu menangkap kualitas representasional yang tidak dapat diamati melalui inspeksi visual semata (Wu et al., 2024).

Perbedaan nilai FID antara model female dan male menunjukkan adanya variasi kemampuan model dalam mempelajari distribusi data target. Model female memperoleh nilai FID sebesar 176,77, sedangkan model male memperoleh nilai 256,22, yang mengindikasikan bahwa distribusi fitur hasil generasi female lebih dekat dengan distribusi data referensi. Temuan ini menunjukkan bahwa model female memiliki kemampuan generalisasi yang lebih baik terhadap data evaluasi yang tidak digunakan selama proses pelatihan. Fenomena serupa dilaporkan oleh Wu et al. (2024) dan Xin et al. (2025), yang

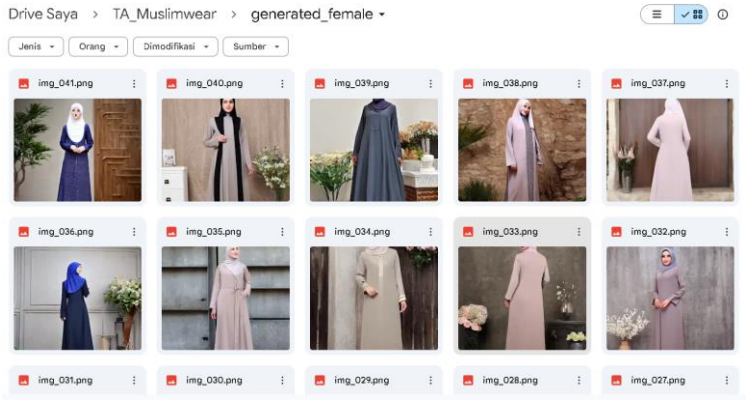
menunjukkan bahwa kualitas representasi laten berkontribusi langsung terhadap penurunan nilai FID pada diffusion model multimodal.

Analisis lebih lanjut menunjukkan bahwa nilai FID tidak hanya merepresentasikan kualitas visual akhir, tetapi juga mencerminkan efektivitas pembelajaran distribusi data pada latent space. Model dengan distribusi fitur yang lebih stabil cenderung menghasilkan keluaran yang memiliki kedekatan statistik lebih tinggi terhadap data asli dibandingkan model dengan representasi laten yang kurang terstruktur (Kuzmin & Berezsky, 2025). Dalam konteks penelitian ini, hasil tersebut mengindikasikan bahwa proses fine-tuning berhasil membangun representasi domain muslimwear yang relatif konsisten pada model female. Karakteristik ini penting karena diffusion model bekerja melalui proses rekonstruksi distribusi data secara bertahap pada ruang laten (Wang et al., 2026).

Faktor lain yang berkontribusi terhadap perbedaan nilai FID adalah kualitas representasi data yang digunakan selama proses pelatihan. Kombinasi dataset DeepFashion1 dan dataset muslimwear memberikan fondasi pembelajaran yang memungkinkan model memahami pola visual umum sekaligus karakteristik domain yang lebih spesifik. Efektivitas representasi data menjadi aspek penting dalam domain adaptation karena distribusi data yang lebih konsisten cenderung menghasilkan distribusi fitur generatif yang lebih stabil (Wu et al., 2024; Zhang et al., 2023). Hubungan antara kualitas data dan performa generatif juga dilaporkan pada penelitian diffusion berskala besar yang memanfaatkan strategi pembelajaran multimodal untuk meningkatkan kualitas distribusi visual (Xin et al., 2025).



Gambar 1. Hasil Generasi Desain Muslimwear (Male)



Gambar 2. Hasil Generasi Desain Muslimwear (Female)

Visualisasi hasil generasi pada Gambar 1 dan Gambar 2 memperlihatkan keluaran model yang kemudian dievaluasi secara kuantitatif menggunakan FID. Evaluasi ini tidak berfokus pada atribut visual spesifik, melainkan pada kedekatan distribusi fitur yang diekstraksi oleh jaringan Inception. Perspektif distribusional tersebut memungkinkan penilaian yang lebih objektif terhadap kualitas

generatif dibandingkan pengamatan visual semata (Wu et al., 2024). Pendekatan serupa banyak digunakan dalam penelitian diffusion modern untuk mengukur realism dan kualitas representasional model generatif (Fan & Lyu, 2025).

Tabel berikut menyajikan hasil evaluasi FID yang diperoleh dari kedua model yang dikembangkan pada penelitian ini.

Tabel 2. Hasil Evaluasi Fréchet Inception Distance (FID)

Model	Nilai FID
Male	256.22
Female	176.77

Sumber: Hasil pengujian sistem penelitian (2026).

Data pada Tabel 2 menunjukkan bahwa model female menghasilkan nilai FID yang lebih rendah dibandingkan model male. Selisih nilai tersebut mengindikasikan bahwa distribusi fitur hasil generasi female memiliki jarak statistik yang lebih kecil terhadap distribusi fitur data referensi. Dalam literatur diffusion model, penurunan nilai FID umumnya dikaitkan dengan peningkatan kemampuan model dalam merekonstruksi distribusi data target secara lebih akurat (Zhang et al., 2023). Interpretasi tersebut sejalan dengan konsep representational learning yang menjadi fondasi utama pada latent diffusion models.

Kinerja model female yang lebih baik juga dapat dikaitkan dengan karakteristik dataset pelatihan yang digunakan selama proses adaptasi domain. Dataset yang memiliki representasi visual lebih seimbang memungkinkan model membangun distribusi fitur yang lebih stabil selama proses pembelajaran. Studi Taiyi-Diffusion-XL menunjukkan bahwa kualitas distribusi data pelatihan memiliki pengaruh langsung terhadap kualitas distribusi fitur yang terbentuk pada latent space model generatif (Wu et al., 2024). Temuan penelitian ini memperlihatkan pola yang serupa karena model dengan distribusi data yang lebih konsisten menghasilkan nilai FID yang lebih rendah.

Dari perspektif fine-tuning, penerapan Low-Rank Adaptation (LoRA) menunjukkan kemampuan yang memadai dalam mentransfer pengetahuan model dasar menuju domain muslimwear. LoRA memungkinkan penyesuaian parameter attention tanpa melakukan pembaruan terhadap seluruh bobot model sehingga kebutuhan komputasi menjadi lebih efisien (Kuzmin & Berezsky, 2025). Nilai FID yang diperoleh menunjukkan bahwa strategi tersebut tetap mampu mempertahankan kualitas representasi laten selama proses adaptasi domain. Hasil ini konsisten dengan penelitian diffusion terkini yang memanfaatkan pendekatan parameter-efficient fine-tuning untuk meningkatkan performa generatif pada domain khusus (Wang et al., 2026).

Efektivitas adaptasi domain melalui LoRA juga memperlihatkan bahwa transfer pengetahuan dari Stable Diffusion v1.5 dapat dilakukan tanpa kehilangan kemampuan representasional yang signifikan. Domain-specific learning yang terbentuk selama proses pelatihan memungkinkan model menghasilkan distribusi fitur yang lebih dekat dengan distribusi data target. Xin et al. (2025) menjelaskan bahwa keberhasilan diffusion model modern sangat dipengaruhi oleh kualitas representasi multimodal yang dipelajari pada latent space. Temuan penelitian ini memperlihatkan bahwa strategi adaptasi ringan masih mampu menghasilkan kualitas distribusional yang kompetitif pada domain fashion muslim.

Hasil evaluasi FID menunjukkan bahwa model female memiliki kualitas generatif yang lebih baik dibandingkan model male berdasarkan kedekatan distribusi fitur terhadap data referensi. Temuan ini memperkuat argumentasi bahwa kualitas dataset, stabilitas representasi laten, dan efektivitas domain adaptation merupakan faktor utama yang memengaruhi performa diffusion model (Huang et al., 2025; Kuzmin & Berezsky, 2025). Perspektif tersebut sejalan dengan perkembangan literatur generative AI yang menempatkan FID sebagai indikator penting dalam mengevaluasi realism dan kualitas distribusional hasil generasi (Wu et al., 2024; Zhang et al., 2023). Hasil penelitian juga menunjukkan bahwa kombinasi Stable Diffusion v1.5 dan LoRA mampu membangun representasi domain muslimwear yang memadai untuk menghasilkan distribusi visual yang relatif dekat dengan data target.

Analisis Konsistensi Struktural Berdasarkan SSIM dan Panduan Multimodal

Structural Similarity Index Measure (SSIM) digunakan untuk mengevaluasi tingkat kesamaan struktur visual antara citra hasil generasi dan citra referensi yang digunakan selama proses pengujian. Berbeda dengan FID yang mengukur kedekatan distribusi fitur pada ruang representasi tingkat tinggi, SSIM berfokus pada kesamaan komponen struktural seperti bentuk, tekstur, dan komposisi spasial. Pendekatan ini relevan untuk menilai kemampuan sistem dalam mempertahankan karakteristik visual yang berasal dari gambar acuan selama proses generasi. Pentingnya evaluasi berbasis struktur telah banyak dilaporkan pada penelitian diffusion multimodal yang menekankan preservasi informasi visual selama proses sintesis gambar (Shuai et al., 2024).

Hasil pengujian menunjukkan adanya perbedaan tingkat kesamaan struktural antara model male dan female yang dikembangkan pada penelitian ini. Nilai SSIM yang diperoleh mengindikasikan bahwa kedua model mampu mempertahankan sebagian karakteristik visual dari gambar referensi selama proses generasi. Variasi nilai yang muncul menunjukkan adanya perbedaan kemampuan representasi struktur yang dipelajari selama proses fine-tuning. Temuan tersebut memberikan gambaran empiris mengenai efektivitas integrasi informasi visual pada arsitektur multimodal yang digunakan (Li et al., 2024).

Nilai SSIM yang lebih tinggi menunjukkan tingkat kemiripan struktur yang lebih besar antara gambar hasil generasi dan gambar referensi. Pada konteks penelitian ini, metrik tersebut digunakan untuk menilai sejauh mana bentuk visual yang terdapat pada citra acuan dapat dipertahankan setelah melalui proses denoising dan rekonstruksi pada latent diffusion model. Pengukuran dilakukan menggunakan data evaluasi yang dipisahkan dari data pelatihan untuk mengurangi bias pengukuran. Pendekatan evaluasi semacam ini banyak digunakan pada penelitian generasi gambar yang menekankan kontrol visual berbasis referensi (Baldrati et al., 2026).

Tabel 3 menyajikan hasil evaluasi Structural Similarity Index Measure (SSIM) pada kedua model yang diuji.

Tabel 3. Hasil Evaluasi Structural Similarity Index Measure (SSIM)

Model	Nilai SSIM
Male	0.275
Female	0.311

Sumber: Hasil pengujian sistem penelitian (2026).

Berdasarkan Tabel 3, model female memperoleh nilai SSIM sebesar 0.311, sedangkan model male memperoleh nilai 0.275. Perbedaan tersebut menunjukkan bahwa model female mampu mempertahankan informasi struktural dari gambar referensi dengan tingkat kesamaan yang lebih tinggi. Nilai tersebut mengindikasikan bahwa representasi visual yang dipelajari pada kategori female lebih efektif dalam menjaga hubungan spasial antar elemen visual selama proses generasi. Fenomena serupa juga ditemukan pada penelitian multimodal image generation yang menunjukkan bahwa kualitas representasi visual berpengaruh terhadap tingkat preservasi struktur hasil generasi (Huang et al., 2025a).

Performa model female yang lebih tinggi dapat dikaitkan dengan karakteristik data pelatihan yang memiliki pola visual lebih konsisten pada ruang representasi laten. Konsistensi pola visual memungkinkan mekanisme pembelajaran mengenali hubungan spasial antar komponen gambar secara lebih stabil selama proses fine-tuning. Kondisi tersebut membantu model membangun korelasi yang lebih kuat antara fitur referensi dan representasi hasil generasi. Studi mengenai generated images sebagai sumber pembelajaran multimodal juga menunjukkan bahwa stabilitas representasi visual berkontribusi terhadap peningkatan kesamaan struktural pada output model (Huang et al., 2025b).

Kemampuan mempertahankan struktur visual pada penelitian ini tidak terlepas dari penggunaan mekanisme cross-attention fusion yang mengintegrasikan informasi teks dan gambar referensi secara simultan. Melalui mekanisme tersebut, representasi visual dari citra acuan dapat digunakan sebagai panduan selama proses pembentukan fitur pada U-Net. Informasi visual tidak hanya berfungsi sebagai konteks tambahan, tetapi juga menjadi sinyal pengarah yang memengaruhi pembentukan struktur gambar pada setiap tahap denoising. Efektivitas pendekatan ini sejalan dengan temuan Hong dan Liu (2025) yang menunjukkan bahwa integrasi multimodal pada diffusion model meningkatkan kemampuan preservasi informasi visual.

Dari perspektif kontrol generasi, hasil SSIM menunjukkan bahwa model tidak hanya mengandalkan embedding tekstual selama proses sintesis gambar. Kehadiran gambar referensi memberikan batasan visual yang membantu model mempertahankan karakteristik struktural tertentu pada output yang dihasilkan. Mekanisme tersebut memungkinkan proses generasi berlangsung secara lebih terarah pada aspek bentuk dan susunan visual tanpa sepenuhnya bergantung pada deskripsi teks. Temuan ini konsisten dengan konsep human-aligned multimodal guidance yang dikemukakan oleh Wu, Huang, dan Wei (2024).

Tingkat kesamaan struktur yang diperoleh juga menunjukkan bahwa pendekatan multimodal mampu mendukung proses visual guidance secara lebih efektif dibandingkan skema generasi berbasis satu modalitas. Informasi referensi yang tersedia selama proses inferensi membantu model menjaga kontinuitas hubungan spasial antar elemen visual yang relevan. Kemampuan tersebut penting pada aplikasi yang memerlukan kontrol visual tinggi karena perubahan kecil pada struktur dapat menghasilkan perbedaan desain yang signifikan. Hasil serupa dilaporkan pada DialogGen yang memanfaatkan interaksi multimodal untuk mempertahankan konsistensi visual selama proses generasi bertahap (Huang et al., 2025a).

Dari sudut pandang aplikasi, kemampuan preservasi struktur memiliki implikasi penting terhadap pengembangan sistem pendukung desain berbasis kecerdasan artifisial. Desainer dapat memanfaatkan gambar referensi sebagai dasar eksplorasi visual tanpa kehilangan karakteristik struktural utama yang diinginkan. Pendekatan ini berpotensi mempercepat proses rapid prototyping karena modifikasi desain dapat dilakukan dengan tetap mempertahankan kerangka visual awal. Potensi tersebut selaras dengan perkembangan workflow fashion berbasis AI yang semakin mengintegrasikan model generatif sebagai alat pendukung proses kreatif dan produksi desain (Baldrati et al., 2026).

Interpretasi keseluruhan hasil SSIM menunjukkan bahwa integrasi gambar referensi melalui cross-attention fusion memberikan kontribusi nyata terhadap kemampuan preservasi struktur pada model yang dikembangkan. Nilai SSIM yang lebih tinggi pada model female mengindikasikan efektivitas pembelajaran representasi visual yang lebih stabil selama proses adaptasi domain. Hasil tersebut memperkuat argumentasi bahwa evaluasi kualitas generasi tidak hanya perlu mempertimbangkan aspek distribusional, tetapi juga tingkat keterjagaan struktur visual terhadap referensi yang digunakan. Perspektif ini sejalan dengan rekomendasi evaluasi multimodal modern yang menempatkan konsistensi struktural sebagai indikator penting dalam pengukuran performa diffusion model (Burapacheep et al., 2024).

KESIMPULAN

Penelitian ini menunjukkan bahwa integrasi input teks dan gambar referensi melalui mekanisme cross-attention fusion pada Stable Diffusion v1.5 mampu meningkatkan keterarahan generasi desain muslimwear dalam lingkungan multimodal. Adaptasi domain menggunakan Low-Rank Adaptation (LoRA) memungkinkan proses fine-tuning berlangsung secara efisien dengan tetap mempertahankan kemampuan model dalam mempelajari karakteristik visual yang relevan. Hasil generasi memperlihatkan kemampuan sistem dalam memanfaatkan informasi semantik dan visual secara simultan untuk menghasilkan desain yang sesuai dengan masukan pengguna. Evaluasi kuantitatif menunjukkan bahwa model female memperoleh nilai FID sebesar 176.77 dan SSIM sebesar 0.311, sedangkan model male memperoleh nilai FID sebesar 256.22 dan SSIM sebesar 0.275. Perbedaan tersebut mengindikasikan bahwa kualitas representasi visual dan konsistensi data pelatihan berpengaruh terhadap kemampuan model dalam mempelajari distribusi fitur serta mempertahankan struktur visual selama proses generasi. Temuan ini memperlihatkan bahwa pendekatan multimodal berbasis diffusion model tidak hanya mendukung kualitas visual yang lebih baik, tetapi juga meningkatkan preservasi struktur terhadap gambar referensi, sehingga berpotensi diterapkan sebagai kerangka kerja pengembangan desain fashion berbasis kecerdasan artifisial yang lebih adaptif, terkontrol, dan efisien.

DAFTAR PUSTAKA

Baldrati, A., Morelli, D., Cornia, M., Bertini, M., & Cucchiara, R. (2026). Multimodal-conditioned latent diffusion models for fashion image editing. *ACM Transactions on Multimedia Computing, Communications and Applications*, *22*(4), 1-27. https://doi.org/10.1145/3789212

Burapacheep, J., Gaur, I., Bhatia, A., & Thrush, T. (2024, August). Colorswap: A color and word order dataset for multimodal evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024* (pp. 1716-1726). https://doi.org/10.18653/v1/2024.findings-acl.99

Chen, Y., & Ma, J. (2025). An intelligent generative method of fashion design combining attribute knowledge and Stable Diffusion Model. *Textile Research Journal*, *95*(11-12), 1231-1254. https://doi.org/10.1177/00405175241289578

Fan, X., & Lyu, M. (2025). MRI Image Generation Based on Text Prompts. *arXiv preprint arXiv:2505.22682*. https://doi.org/10.48550/arXiv.2505.22682

Ghori, I., Karim, K., & Alkawadri, D. (2025, August). GenAI-Driven Image Generation Pipeline for Sustainable Garment Design and Waste Reduction in Fashion Production. In *Proceedings of the AAAI Symposium Series* (Vol. 6, No. 1, pp. 218-226). https://doi.org/10.1609/aaaiss.v6i1.36056

Hong, C. Y., & Liu, T. L. (2025, April). Multimodal promptable token merging for diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39, No. 16, pp. 17231-17239). https://doi.org/10.1609/aaai.v39i16.33894

Huang, M., Long, Y., Deng, X., Chu, R., Xiong, J., Liang, X., ... & Liu, W. (2025, April). DialogGen: Multi-modal Interactive Dialogue System with Multi-turn Text-Image Generation. In *Findings of the Association for Computational Linguistics: NAACL 2025* (pp. 411-426). https://doi.org/10.18653/v1/2025.findings-naacl.25

Huang, Y., Zhang, P., Liu, R., & Liang, J. (2025). Can Generated Images Serve as a Viable Modality for Text-Centric Multimodal Learning?. *arXiv preprint arXiv:2506.17623*. https://doi.org/10.48550/arXiv.2506.17623

Kasodekar, K. S. (2024). Remote Diffusion. *arXiv preprint arXiv:2405.04717*. https://doi.org/10.48550/arXiv.2405.04717

Kuzmin, S., & Berezsky, O. (2025). Analysis of Diffusion Models and Biomedical Image Generation Tools. *Computer systems and information technologies*, (2), 8-19. https://doi.org/10.31891/csit-2025-2-1

Li, W., Xu, X., Liu, J., & Xiao, X. (2024, August). Unimo-g: Unified image generation through multimodal conditional diffusion. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 6173-6188). https://doi.org/10.18653/v1/2024.acl-long.335

Liu, L., Du, C., Pang, T., Wang, Z., Li, C., & Xu, D. (2024). Improving long-text alignment for text-to-image diffusion models. *arXiv preprint arXiv:2410.11817*. https://doi.org/10.48550/arXiv.2410.11817

Ma, Y., Yang, H., Wang, W., Fu, J., & Liu, J. (2023). Unified multi-modal latent diffusion for joint subject and text conditional image generation. *arXiv preprint arXiv:2303.09319*. https://doi.org/10.48550/arXiv.2303.09319

Shuai, X., Ding, H., Ma, X., Tu, R., Jiang, Y. G., & Tao, D. (2024). A survey of multimodal-guided image editing with text-to-image diffusion models. *arXiv preprint arXiv:2406.14555*. https://doi.org/10.48550/arXiv.2406.14555

Wang, Y., Zhu, B., Hao, Y., Ngo, C. W., Tan, Y., & Wang, X. (2026). Cookingdiffusion: Cooking procedural image generation with stable diffusion. *ACM Transactions on Multimedia Computing, Communications and Applications*, *22*(1), 1-24. https://doi.org/10.1145/3771995

Wu, H., Wu, X., Li, C., Zhang, Z., Chen, C., Liu, X., ... & Lin, W. (2024, October). T2i-scorer: Quantitative evaluation on text-to-image generation via fine-tuned large multi-modal models. In *Proceedings of the 32nd ACM International Conference on Multimedia* (pp. 3676-3685). https://doi.org/10.1145/3664647.3680939

Wu, X., Huang, S., & Wei, F. (2024). Multimodal large language model is a human-aligned annotator for text-to-image generation. *arXiv preprint arXiv:2404.15100*. https://doi.org/10.48550/arXiv.2404.15100

Wu, X., Zhang, D., Gan, R., Lu, J., Wu, Z., Sun, R., ... & Song, Y. (2024). Taiyi-Diffusion-XL: advancing bilingual text-to-image generation with large vision-language model support. *arXiv preprint arXiv:2401.14688*. https://doi.org/10.48550/arXiv.2401.14688

Xin, Y., Qin, Q., Luo, S., Zhu, K., Yan, J., Tai, Y., ... & Liu, Y. (2025). Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. *arXiv preprint arXiv:2510.06308*. https://doi.org/10.48550/arXiv.2510.06308

Zhang, T., Wang, Z., Huang, J., Tasnim, M. M., & Shi, W. (2023). A survey of diffusion based image generation models: Issues and their solutions. *arXiv preprint arXiv:2308.13142*. https://doi.org/10.48550/arXiv.2308.13142