 ISSN:2302-8149
e-ISSN:2540-9719

Perbandingan Model *Text-to-Image* untuk Motif Batik Nitik dengan *Prompt Enhancement LLM*

Comparing Text-to-Image Models for Nitik Batik Motifs with LLM-based Prompt Enhancement

[1]Galih Mahda Praditya Gunardi*, [2]Muhammad Iqbal Akbar
[1,2]Program Studi Teknik Informatika, Fakultas Teknik, Universitas Negeri Malang
[1,2]Jl. Semarang No. 5, Kota Malang, Jawa Timur, Indonesia
*e-mail: galihmahdal123@gmail.com

(***received*:** 11 October 2026, ***revised*:** 7 November 2026, ***accepted*:** 8 November 2026)

Abstrak

Model text-to-image pretrained memungkinkan pembuatan motif batik tanpa fine-tuning, tetapi kinerja model generik dan manfaat prompt enhancement berbasis Large Language Model (LLM) pada motif batik belum dikaji secara sistematis. Penelitian ini membandingkan Stable Diffusion 1.5 (SD 1.5) dan SDXL dalam menghasilkan motif Batik Nitik Yogyakarta secara zero-shot, serta mengevaluasi pengaruh prompt enhancement berbasis LLM. Enhanced Prompt dibuat oleh Gemini melalui satu instruksi zero-shot, tanpa contoh few-shot dan system instruction. Metode yang digunakan adalah kuantitatif-eksperimental dengan empat kombinasi pengujian (dua model dan dua jenis prompt) pada 15 motif dari dataset Batik Nitik 960, dengan 25 gambar per motif dan seed tetap. Evaluasi memakai Frechet Inception Distance (FID) terhadap gambar asli dari 15 motif yang sama dan CLIP Score terhadap satu teks acuan yang sama, dengan uji-t berpasangan pada tingkat motif. SD 1.5 dengan Enhanced Prompt menghasilkan FID terendah (296,714), tetapi selisih antar kombinasi yang berdekatan belum dapat dibedakan dengan pasti. Dengan teks acuan yang sama, Enhanced Prompt tidak meningkatkan CLIP Score (SD 1.5: -0,40; SDXL: -1,61). Pengaruhnya terhadap FID bergantung pada model: membaik pada SD 1.5 dan memburuk pada SDXL. Seluruh Enhanced Prompt melampaui batas 77 token sehingga terpotong.
Kata kunci: *batik nitik, large language model, prompt enhancement, stable diffusion, text-to-image*

Abstract

Pretrained text-to-image models enable batik motif generation without fine-tuning. However, the performance of general-purpose models and the benefits of Large Language Model (LLM)-based prompt enhancement for batik motif generation have not yet been systematically investigated. This study compares Stable Diffusion 1.5 (SD 1.5) and SDXL in generating Yogyakarta Nitik Batik motifs under zero-shot conditions and evaluates the effects of LLM-based prompt enhancement. Enhanced prompts were generated using Gemini through a single zero-shot instruction, without few-shot examples or system instructions. A quantitative experimental approach was employed, involving four experimental configurations (two models and two prompt types) across 15 motifs from the Batik Nitik 960 dataset, with 25 images generated per motif using fixed random seeds. Evaluation was conducted using the Fréchet Inception Distance (FID), calculated against original images from the same 15 motifs, and the CLIP Score, measured against a shared reference text. Paired t-tests were performed at the motif level. SD 1.5 with enhanced prompts achieved the lowest FID score (296.714); however, the differences among closely ranked configurations could not be established with sufficient certainty. Using the same reference text, prompt enhancement did not improve CLIP Scores (SD 1.5: −0.40; SDXL: −1.61). Its effect on FID was model-dependent, improving the results for SD 1.5 while worsening them for SDXL. All enhanced prompts exceeded the 77-token limit and were consequently truncated.
Keywords: *batik nitik, large language model, prompt enhancement, stable diffusion, text-to-image*

 ISSN:2302-8149
e-ISSN:2540-9719

1 Pendahuluan

Batik merupakan salah satu warisan budaya Indonesia yang sarat akan nilai filosofis dan estetika, dengan setiap motif memiliki makna simbolis yang diwariskan secara turun-temurun. Di tengah perkembangan pesat teknologi kecerdasan buatan, muncul peluang untuk melestarikan dan mengeksplorasi motif batik melalui pendekatan komputasional, salah satunya melalui teknologi text-to-image (T2I) generation.

Penelitian terdahulu pada domain text-to-image untuk motif batik umumnya berfokus pada pendekatan fine-tuning terhadap satu model generatif tertentu. Penelitian yang memanfaatkan latent diffusion model (LDM), Low-Rank Adaptation (LoRA), dan classifier-free guidance menunjukkan bahwa riset batik saat ini masih berfokus pada klasifikasi, sehingga eksplorasi model generatif untuk pembuatan motif batik masih menjadi celah tersendiri [1]. Pendekatan serupa juga diterapkan pada domain tekstil tradisional lain, seperti pada penelitian generasi motif kain Endek Bali yang membuktikan LoRA mampu meningkatkan kemampuan model dalam menangkap karakteristik struktural motif tekstil tradisional [2]. Pada domain batik etnis minoritas Tiongkok, model generatif juga terbukti kesulitan menangkap atribut semantik dan gaya yang kompleks pada motif batik Miao, sehingga dikembangkan dataset terstruktur CMBP-9 dengan pendekatan Style-Conditioned LCM-LoRA untuk mengatasi keterbatasan tersebut [3]. Sementara itu, pendekatan lain menggabungkan teknik Diffusion-GAN dengan StyleGAN2-Ada untuk menghasilkan motif batik yang realistis dan bervariasi, dengan evaluasi menggunakan metrik Frechet Inception Distance (FID) dan Kernel Inception Distance (KID) [4].

Meskipun pendekatan fine-tuning tersebut terbukti efektif, proses tersebut membutuhkan sumber daya komputasi besar, dataset ekstensif, serta waktu pelatihan yang tidak singkat. Di sisi lain, penelitian pada bidang optimasi prompt menunjukkan bahwa Large Language Model (LLM) dapat dimanfaatkan untuk melakukan prompt engineering secara otomatis guna meningkatkan kualitas prompt sekaligus gambar yang dihasilkan oleh model text-to-image [5]. Namun, penerapan pendekatan prompt enhancement berbasis LLM ini belum pernah dikaji secara khusus pada domain motif batik Indonesia, terlebih dikombinasikan dengan perbandingan performa beberapa model text-to-image pretrained secara sistematis.

Berdasarkan permasalahan tersebut, rumusan masalah penelitian ini adalah (1) bagaimana perbandingan kinerja Stable Diffusion 1.5 dan Stable Diffusion XL (SDXL) sebagai model text-to-image pretrained dalam menghasilkan motif Batik Nitik Yogyakarta secara zero-shot, dan (2) bagaimana pengaruh prompt enhancement berbasis LLM terhadap kualitas visual dan kesesuaian semantik gambar yang dihasilkan pada masing-masing model.

Penelitian ini bertujuan menganalisis secara komparatif performa kedua model tersebut dan mengevaluasi pengaruh prompt enhancement berbasis LLM, dengan kualitas visual diukur melalui Frechet Inception Distance (FID) dan kesesuaian semantik melalui CLIP Score (definisi dan cara perhitungannya dijelaskan pada bagian Metode Penelitian). Kontribusi penelitian ini adalah (1) perbandingan langsung dua model pretrained tanpa adaptasi pada domain Batik Nitik, (2) evaluasi LLM sebagai pengaya prompt pada sisi masukan berbasis pengetahuan motif, dan (3) protokol evaluasi dengan seed tetap, banyak gambar per motif, dan teks acuan CLIP yang seragam, yang diuji secara statistik pada tingkat motif. Evaluasi dibatasi pada 15 dari 60 motif dalam dataset. Jumlah ini dipilih agar keempat kategori visual (bunga, buah, tanaman, dan bentuk geometris) terwakili dengan beban komputasi yang tetap terjangkau pada GPU T4 Google Colab, mengingat setiap motif dibangkitkan 25 kali pada empat kombinasi pengujian.

2 Tinjauan Literatur

Riset tentang motif batik yang menggunakan teknik generatif mulai beralih dari pendekatan berbasis Generative Adversarial Network (GAN) ke model difusi. Tim peneliti yang dipimpin oleh Octadion dan rekan-rekannya [4] menggabungkan StyleGAN2-Ada dengan teknik difusi pada dataset batik yang sudah dikumpulkan dan diatur. Mereka menguji hasilnya menggunakan berbagai metrik seperti FID, KID, precision, recall, dan non-redundancy.

Tidak lama kemudian, muncul penelitian yang menunjukkan arah baru, yakni generasi yang bisa dikendalikan lewat teks. Di antaranya adalah hasil karya Wahid dan tim [1]. Mereka melatih ulang model difusi latent (LDM) [6] dengan 20.000 gambar batik yang sudah dilengkapi caption otomatis.

 ISSN:2302-8149
e-ISSN:2540-9719

Caption tersebut dibuat oleh model multimodal seperti LLaVA dan BLIP. Mereka melakukan pelatihan penuh maupun adaptasi berbasis rank rendah (LoRA) [7]. Selain itu, mereka juga menggunakan teknik classifier-free guidance agar hasilnya sesuai dengan permintaan yang diberikan.

Pola yang sama juga terlihat pada tekstil tradisional lainnya. Hu dan timnya [3] membuat dataset CMBP-9 untuk batik Miao dan menyesuaikan Stable Diffusion v1.5 menggunakan LoRA dipadukan dengan Latent Consistency Model. Strategi penggabungan ini menyesuaikan bobot berdasarkan kompleksitas makna prompt. Purnawan dan Suparjana [2] menyesuaikan model difusi dengan LoRA pada 687 gambar kain Endek Bali yang disertai deskripsi singkat dalam bahasa Indonesia. Mereka menyatakan bahwa rata-rata skor FID turun dari 394,21 (nilai awal) menjadi 276,05 setelah menggunakan LoRA, sementara skor CLIP berada di kisaran 26-30. Tiga studi ini menunjukkan bahwa kualitas motif sangat bergantung pada penyesuaian model dengan dataset khusus domain, mulai dari ratusan hingga puluhan ribu gambar. Nilai FID awal yang tinggi pada [2] juga menandakan bahwa model generik kesulitan menangkap motif tekstil tradisional. Kemampuan model tanpa penyesuaian ini jarang diuji secara sistematis.

Studi-studi ini biasanya fokus pada adaptasi satu model dasar dari keluarga difusi laten. Menurut pencarian penulis, belum ada perbandingan langsung antara model pretrained yang belum diadaptasi di domain batik. Padahal generasi model berbeda secara arsitektur. SDXL [8], contohnya, memakai backbone lebih besar dan dua text encoder, sehingga tanggapan terhadap prompt deskriptif panjang bisa berbeda dari pendahulunya.

Dari sisi evaluasi, FID digunakan pada [2] dan [4], sementara kesesuaian antara teks dan gambar diukur menggunakan skor berbasis CLIP [9] pada [2]. Namun Jayasumana dkk. [10] menunjukkan bahwa FID memiliki kelemahan: representasi Inception yang tidak cukup baik untuk konten model modern, asumsi normalitas yang tidak selalu benar, serta hasil yang tidak konsisten ketika jumlah sampel berubah. Kelemahan ini sangat penting dalam penelitian yang memiliki jumlah gambar terbatas. FID sebaiknya tidak digunakan sebagai satu-satunya alat penilaian, tetapi sebaiknya dikombinasikan dengan metrik semantik dan uji statistik pada skor per-sampel.

Selain soal metrik, keandalan model generik pada domain budaya juga dipertanyakan. Rombach dkk. [6] memperkenalkan model difusi laten yang menjadi dasar Stable Diffusion, sedangkan SDXL [8] menambah kapasitas model dan memakai dua text encoder untuk menangkap prompt yang lebih deskriptif. Kemampuan tersebut tidak otomatis berarti ketepatan pada konteks budaya tertentu. Nayak dkk. [18] melaporkan bahwa model text-to-image secara sistematis kesulitan memenuhi ekspektasi budaya yang bersifat implisit, terutama pada komunitas yang kurang terwakili dalam data latih. Selain itu, Carlini dkk. [17] menunjukkan bahwa model difusi dapat mengingat dan mereproduksi gambar dari data latihnya, sehingga keluaran model generik perlu diperiksa dari kemungkinan artefak. Pada sisi masukan, Arzomand dkk. [19] mengusulkan kerangka kolaborasi manusia dan AI untuk rekonstruksi warisan budaya yang memadukan arahan pakar dengan rekayasa prompt yang sistematis. Temuan-temuan ini menunjukkan bahwa kualitas keluaran pada domain budaya bergantung pada dua hal, yaitu kemampuan model dasar dan ketepatan informasi domain yang disampaikan melalui prompt. Namun belum ada kajian yang menguji keduanya bersamaan pada motif batik tanpa fine-tuning.

Pemanfaatan LLM dalam studi batik saat ini masih terbatas pada data pelatihan, yaitu membuat caption untuk fine-tuning [1]. Di luar bidang budaya, LLM digunakan pada sisi masukan untuk memperbaiki prompt. Gani dkk. [11] menemukan bahwa model difusi kesulitan menangani prompt yang panjang dan rumit, sehingga mereka menggunakan LLM untuk memecah komponen prompt secara terstruktur. Li dkk. [5] memanfaatkan LLM untuk mengoptimalkan prompt secara otomatis dalam T2I dan melaporkan peningkatan kualitas prompt serta gambar. Kedua pendekatan ini bersifat umum. Namun, pengayaan prompt dengan pengetahuan spesifik motif budaya, seperti ciri visual dan makna filosofisnya, belum pernah diteliti dalam domain batik.

 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 1 Perbandingan penelitian terdahulu

Penelitian	Domain	Pendekatan	Data	Fokus Perbandingan	LLM saat inferensi
Octadion dkk. [4]	Batik	Diffusion-GAN + StyleGAN2-Ada	Batik terkurasi	Arsitektur Diffusion-GAN	Tidak
Wahid dkk. [1]	Batik	LDM + LoRA + CFG (fine-tuning)	20.000 gambar	Fine-tuning penuh vs LoRA	Tidak (LLM hanya untuk caption data)
Hu dkk. [3]	Batik Miao	SD v1.5 + LoRA + LCM	649 gambar (CMBP-9)	Fusi LoRA–LCM	Tidak
Purnawan & Suparjana [2]	Tekstil Endek	Difusi + LoRA	687 gambar	Baseline vs LoRA	Tidak
Penelitian ini	Batik Nitik	Model pretrained tanpa adaptasi	240 gambar dari 15 motif (hanya evaluasi)	SD 1.5 vs SDXL × jenis prompt	Ya (gemini)

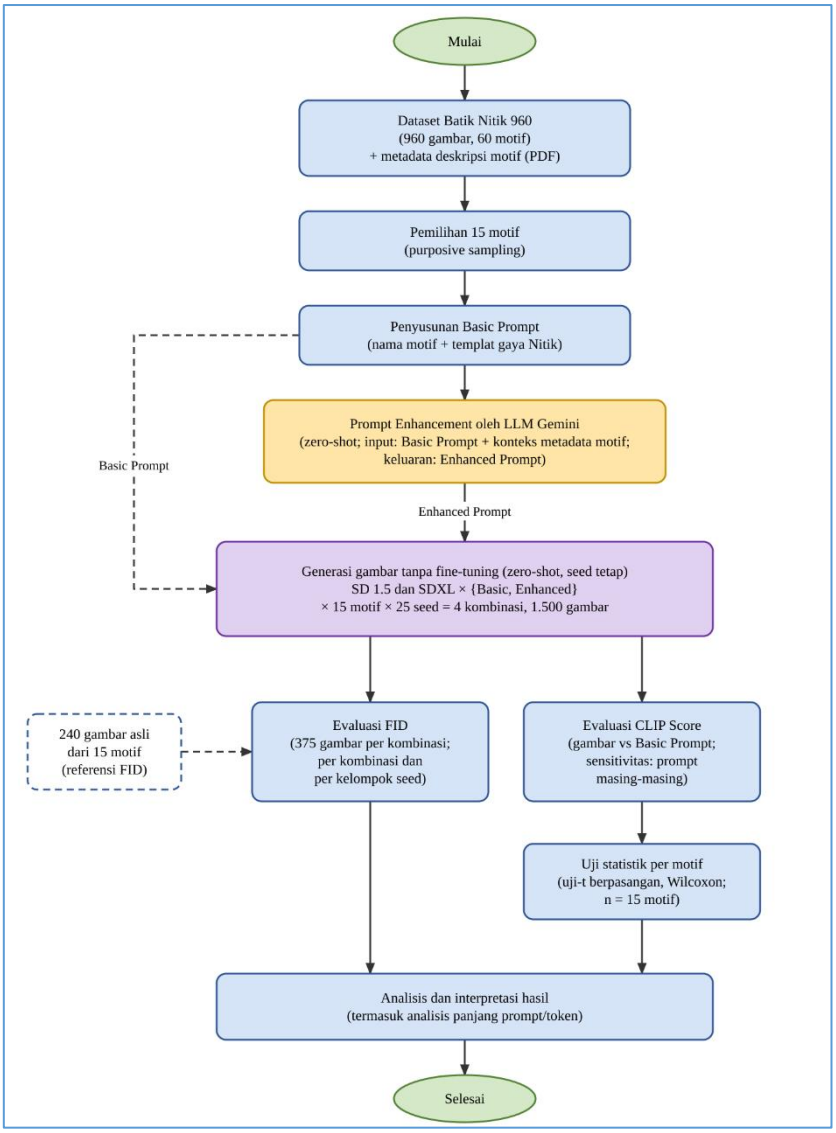
Berdasarkan penelitian yang dilakukan, ada tiga hal baru yang dihasilkan. Pertama, penelitian ini mengecek kemampuan dua model yang sudah dilatih sebelumnya, yaitu SD 1.5 dan SDXL, tanpa melakukan penyesuaian khusus pada motif Batik Nitik Yogyakarta. Kedua, penelitian ini menggunakan LLM di bagian masukan agar bisa menambahkan informasi visual dan makna dari motif, bukan untuk membuat deskripsi data latih. Ketiga, dalam evaluasi, digunakan FID bersama dengan skor CLIP dan uji-t berpasangan untuk setiap motif, sehingga tidak hanya mengandalkan FID yang bisa terpengaruh oleh jumlah sampel [10].

3 Metode Penelitian

Penelitian ini menggunakan metode kuantitatif-eksperimental untuk mengevaluasi kemampuan model text-to-image dalam menghasilkan motif Batik Nitik khas Yogyakarta. Berbeda dengan penelitian serupa lainnya yang umumnya menggunakan proses fine-tuning pada model generatif tunggal dengan data dalam jumlah besar, penelitian ini menggunakan model yang telah dilatih sebelumnya (pre-trained models) dengan pendekatan zero-shot. Dengan demikian, hasil penelitian akan menunjukkan kemampuan model dalam menghasilkan gambar berdasarkan prompt yang memuat deskripsi motif budaya spesifik tanpa memerlukan pelatihan tambahan.

Secara umum, penelitian ini membandingkan dua model text-to-image yang mewakili dua generasi arsitektur model berbeda sehingga memiliki mekanisme dasar yang berlainan yaitu Stable Diffusion 1.5 dan SDXL. Di sisi lain, sebagai pertanyaan penelitian tersendiri, studi ini juga mengevaluasi efektivitas penggunaan Large Language Models (LLM) untuk menyempurnakan prompt awal. LLM Gemini digunakan untuk memperkaya prompt awal dengan deskripsi filosofis dan visual motif guna memperoleh hasil yang lebih relevan. Dengan demikian, penelitian ini mencakup empat kombinasi pengujian (dua model dan dua jenis prompt) serta dua metrik evaluasi kuantitatif yaitu Frechet Inception Distance (FID) dan CLIP Score. Seluruh tahapan penelitian, mulai dari persiapan hingga analisis, dijelaskan lebih lanjut dalam makalah ini. Gambaran umum penelitian disajikan pada Gambar 1 di bawah ini.

 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 1 Alur penelitian

Data dan objek penelitian. Objek penelitian adalah pola Batik Nitik Yogyakarta. Dataset yang digunakan adalah Batik Nitik 960, data sekunder terbuka yang diterbitkan oleh Minarno dkk. di repositori Mendeley Data (versi 3, 2023; DOI 10.17632/sgh484jxzy.3) [12] dan didukung publikasi data descriptor di jurnal Data [20]. Gambar berasal dari kain koleksi Batik Winotosastro milik Paguyuban Pecinta Batik Indonesia (PPBI) Sekar Jagad Yogyakarta. Dataset berisi 960 gambar dalam 60 kategori pola (16 gambar per kategori: empat foto asli yang masing-masing diperbanyak melalui rotasi 90°, 180°, dan 270°), dilengkapi metadata bilingual yang memuat nama dan makna filosofis setiap pola. Gambar asli dipakai sebagai acuan FID dan metadata sebagai sumber konteks prompt; dataset tidak dipakai untuk melatih model apa pun.

Pengumpulan dan penyiapan data. Nama pola diambil dari nama file gambar, lalu disesuaikan dengan teks metadata yang diambil dari berkas PDF menggunakan pustaka pdfplumber. Karena ada perbedaan penulisan nama motif antara file gambar dan PDF, kecocokan dilakukan dengan memperhatikan variasi ejaan itu. Dari proses ini, 57 dari 60 pola berhasil sejalan dengan deskripsinya. Kemudian 15 motif dipilih dengan teknik purposive sampling berdasarkan kriteria variasi kategori visual (bunga, buah, tanaman, dan bentuk geometris) untuk memastikan representasi berbagai bentuk visual dalam evaluasi. Motif yang dipilih adalah Sekar Kemuning, Sekar Duren, Cakar Ayam, Sekar Jagung, Sekar Tanjung, Sekar Nangka, Kawung Nitik, Sekar Mlathi, Manggar, Gedhangan, Sekar Lintang, Sekar Kenanga, Sekar Jeruk, Sekar Kepel, dan Jayakusuma. Jumlah 15 motif (25% dari 60

 ISSN:2302-8149
e-ISSN:2540-9719

pola) ditetapkan sebagai kompromi antara keterwakilan kategori visual dan keterbatasan komputasi GPU T4 Google Colab. Gambar asli dari 15 motif tersebut (15 × 16 = 240 gambar, termasuk versi rotasi) digunakan sebagai acuan FID.

Alat dan bahan. Eksperimen dilaksanakan dalam lingkungan komputasi awan Google Colab menggunakan GPU NVIDIA Tesla T4 dan bahasa pemrograman Python. Model yang dibandingkan adalah Stable Diffusion 1.5 [6] dan SDXL base 1.0 [8], yang keduanya diakses melalui pustaka Diffusers [14] dengan presisi float16 dan parameter bawaan pustaka tanpa fine-tuning, yaitu guidance scale 7,5 dengan resolusi 512 × 512 piksel untuk SD 1.5, serta guidance scale 5,0 dengan resolusi 1024 × 1024 piksel untuk SDXL. Pemilihan antara SD 1.5 dan SDXL didasarkan pada perbedaan generasi arsitektur mereka: SD 1.5 menggambarkan generasi awal dari model difusi laten dengan satu pengode teks, sedangkan SDXL mencerminkan generasi yang lebih maju dengan backbone lebih besar dan dua pengode teks [8], sehingga perbandingan antara keduanya dapat mengungkapkan dampak kompleksitas arsitektur terhadap efektivitas peningkatan prompt. Pengayaan prompt menggunakan LLM Gemini (gemini-3.6-flash) yang diakses melalui API [15] dan dijalankan secara zero-shot melalui satu instruksi pengguna tanpa contoh few-shot dan tanpa system instruction, dengan parameter pembangkitan bawaan API. Penilaian kecocokan semantik memakai model CLIP ViT-B/32 [9].

Rancangan penelitian. Penyusunan Enhanced Prompt dalam penelitian ini mengadopsi prinsip bahwa prompt engineering yang didasarkan pada pengetahuan domain spesifik dapat meningkatkan ketepatan representasi visual pada konten budaya, sebagaimana ditunjukkan pada kerangka kerja rekonstruksi warisan budaya berbasis bukti arkeologis dan ikonografis [19]. Basic Prompt disusun dari templat berbahasa Inggris: "traditional Indonesian batik pattern, {nama motif} motif, Nitik style, intricate detailed pattern, symmetrical, textile design, high quality" (dalam bahasa Indonesia: corak batik Indonesia tradisional, motif {nama motif}, gaya Nitik, pola yang kaya detail, simetris, desain tekstil, kualitas tinggi). Enhanced Prompt dibuat satu kali untuk setiap motif dengan mengarahkan Gemini untuk berfungsi sebagai pakar seni batik dan teknik prompt, melalui satu instruksi zero-shot (tanpa contoh few-shot dan tanpa system instruction). Masukannya berupa Basic Prompt, nama motif, serta 500 karakter pertama dari deskripsi metadata sebagai konteks visual dan filosofis, dan keluarannya adalah satu prompt berbahasa Inggris dengan batas instruksi kurang dari 75 kata. Setiap model menghasilkan 25 gambar untuk setiap motif dari masing-masing jenis prompt, dengan seed tetap 1000 sampai 1024 yang sama untuk keempat kombinasi, sehingga setiap kombinasi terdiri atas 375 gambar dan seluruh eksperimen menghasilkan 1.500 gambar. Rincian parameter eksperimen ditunjukkan pada Tabel 2.

Tabel 2 Parameter eksperimen

Parameter	Nilai
Model	SD 1.5 dan SDXL base 1.0 (tanpa fine-tuning)
Jenis prompt	Basic dan Enhanced (LLM)
Jumlah motif	15
Seed	1000–1024 (tetap; sama untuk semua kombinasi)
Gambar per motif	25
Gambar per kombinasi	375 (total 1.500)
Langkah inferensi	25
Guidance scale	7,5 (SD 1.5); 5,0 (SDXL)
Resolusi	512 × 512 (SD 1.5); 1024 × 1024 (SDXL)
Perangkat keras	Google Colab, GPU Tesla T4

Menurut Tabel 2, perpaduan antara dua model dan dua jenis prompt menghasilkan empat kelompok pengujian yang menggunakan motif, seed, langkah inferensi, dan perangkat keras yang sama, sehingga perbedaan hasil di antara kelompok mengindikasikan pengaruh model dan jenis prompt. Perbedaan antar model juga mencakup perbedaan resolusi dan guidance scale bawaan masing-masing, sehingga perbandingan SD 1.5 dan SDXL perlu ditafsirkan dengan hati-hati.

Definisi operasional variabel. Variabel penelitian, meliputi variabel bebas, terikat, dan kontrol, didefinisikan secara operasional pada Tabel 3.

 ISSN:2302-8149
e-ISSN:2540-9719

Tabel 3 Variabel penelitian

Variabel	Jenis	Definisi operasional
Model text to image	Bebas	SD 1.5 atau SDXL
Jenis prompt	Bebas	Basic Prompt atau Enhanced Prompt
FID	Terikat	Jarak Frechet antar fitur Inception-v3 dari 375 gambar hasil generasi per kombinasi dan 240 gambar asli dari 15 motif yang sama [13]; makin rendah makin baik
CLIP Score	Terikat	Kemiripan kosinus antara embedding gambar dan embedding Basic Prompt motif yang sama (satu teks acuan untuk semua gambar) dikali 100 [9], [16], dirata-rata dari 25 seed per motif lalu dari 15 motif; makin tinggi makin baik
Motif, seed, langkah inferensi, dataset referensi, perangkat keras	Kontrol	Dijaga sama pada seluruh kombinasi

Metrik evaluasi. Penelitian ini menerapkan dua metrik yang saling melengkapi. Metrik pertama, Frechet Inception Distance (FID), mengukur seberapa mirip sekumpulan gambar hasil generasi dengan sekumpulan gambar yang sebenarnya [13]. Kedua set gambar dimasukkan ke jaringan Inception-v3 untuk mengekstrak fitur visualnya, yaitu serangkaian angka yang menggambarkan tekstur, pola, dan bentuk gambar. Sebaran fitur dari gambar asli dan gambar yang dihasilkan dimodelkan sebagai distribusi Gaussian, dan jarak Frechet antara kedua distribusi dihitung sebagai nilai FID. Semakin rendah FID, semakin mirip kualitas visual gambar hasil generasi dengan gambar aslinya. Gambar acuan yang dipakai adalah 240 gambar asli dari 15 motif yang sama dengan motif yang dibangkitkan (bukan seluruh 960 gambar), sedangkan gambar yang dihasilkan sebanyak 375 untuk setiap kombinasi pengujian. FID hanya mengevaluasi kesamaan visual tanpa memperhatikan isi teks prompt, dan nilainya dipengaruhi jumlah sampel serta representasi fitur yang dipakai [10].

Metrik kedua, CLIP Score, mengukur relevansi antara gambar dan teks [16]. Model CLIP [9] dilatih dengan sekitar 400 juta pasangan gambar dan teks untuk mengonversi gambar maupun teks menjadi vektor dalam ruang yang sama. Skor dihitung dari kesamaan kosinus antara vektor gambar dan vektor teks, dikali 100, dengan model CLIP ViT-B/32; semakin besar nilainya, semakin sesuai isi gambar dengan teks. Karena Enhanced Prompt jauh lebih panjang daripada Basic Prompt, penilaian dengan prompt masing-masing akan membandingkan gambar dengan teks yang berbeda. Oleh karena itu, CLIP Score utama dalam penelitian ini dihitung untuk semua gambar terhadap satu teks acuan yang sama, yaitu Basic Prompt dari motif yang bersangkutan. Sebagai analisis sensitivitas, CLIP Score juga dihitung terhadap prompt yang dipakai untuk membangkitkan masing-masing gambar (dengan teks dipotong pada batas 77 token CLIP). Skor setiap motif diperoleh dengan merata-ratakan 25 gambarnya, kemudian dirata-ratakan pada 15 motif.

Teknik analisis. Penelitian ini menghitung dua metrik evaluasi kuantitatif pada empat kombinasi pengujian. Pertama, FID dihitung untuk setiap kombinasi dengan membandingkan fitur 375 gambar hasil generasi dengan fitur 240 gambar asli dari 15 motif yang sama. Untuk menilai kestabilan, FID juga dihitung pada lima kelompok seed yang masing-masing terdiri atas lima seed (75 gambar per kombinasi); nilai FID kelompok ini lebih tinggi karena sampel lebih sedikit sehingga hanya dipakai untuk melihat variabilitas dan konsistensi peringkat, bukan untuk dibandingkan dengan FID utama. Kedua, CLIP Score dihitung untuk setiap gambar terhadap Basic Prompt motifnya, lalu dirata-ratakan per motif (25 seed) dan antar 15 motif; analisis sensitivitas memakai prompt masing-masing gambar.

 ISSN:2302-8149
e-ISSN:2540-9719

Ketiga, uji-t berpasangan dan uji peringkat bertanda Wilcoxon dilakukan pada CLIP Score per motif (n = 15) untuk membandingkan Basic dan Enhanced pada setiap model, serta SD 1.5 dan SDXL pada setiap jenis prompt, dengan ukuran efek Cohen's dz. Uji statistik tidak dilakukan pada FID karena FID dihitung untuk satu himpunan gambar per kombinasi. Keempat, jumlah token Basic Prompt dan Enhanced Prompt dihitung dengan tokenizer CLIP untuk mengetahui prompt yang melampaui batas 77 token.

4 Hasil dan Pembahasan

Hasil generasi mencakup 1.500 gambar, yaitu 15 motif Batik Nitik × 25 seed × 4 kombinasi (dua model dan dua jenis prompt), dengan 375 gambar per kombinasi. Gambar 2 menampilkan contoh hasil generasi motif Sekar Kemuning dengan seed yang sama pada keempat kombinasi, dibandingkan dengan salah satu gambar asli dari dataset. Contoh ini bersifat ilustratif dan tidak dimaksudkan sebagai bukti; penilaian dilakukan secara kuantitatif seperti pada Tabel 4 dan Tabel 5.



Gambar 2 Contoh hasil generasi motif sekar kemuning pada empat kombinasi model dan prompt

Hasil evaluasi FID dan CLIP Score. Hasil evaluasi kuantitatif pada keempat kombinasi ditunjukkan pada Tabel 4.

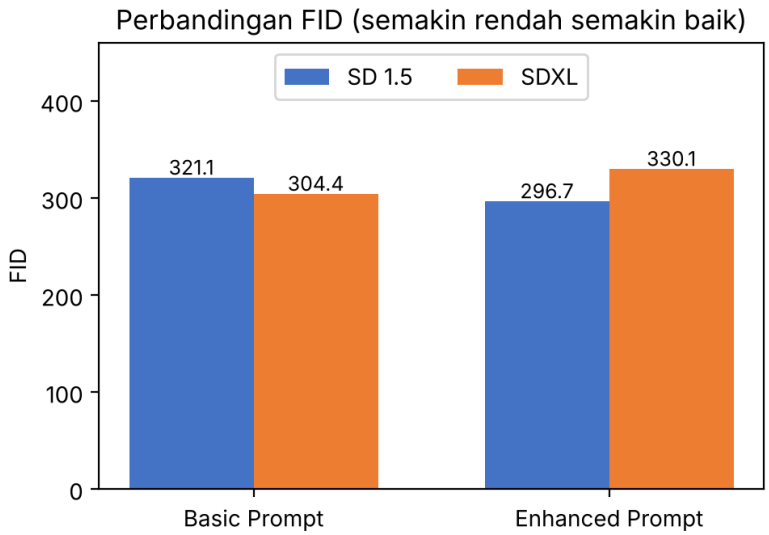
Tabel 4 Hasil evaluasi FID dan CLIP score per kombinasi

Model	Jenis Prompt	FID	FID per kelompok seed (rerata ± SD)	CLIP Score (rerata ± SD antar motif)
Stable Diffusion 1.5	Enhanced	296,714	316,7 ± 7,8	29,62 ± 0,40
SDXL	Basic	304,446	328,1 ± 10,9	29,05 ± 0,31
Stable Diffusion 1.5	Basic	321,053	348,5 ± 14,4	30,02 ± 0,34
SDXL	Enhanced	330,069	350,2 ± 15,1	27,43 ± 0,64

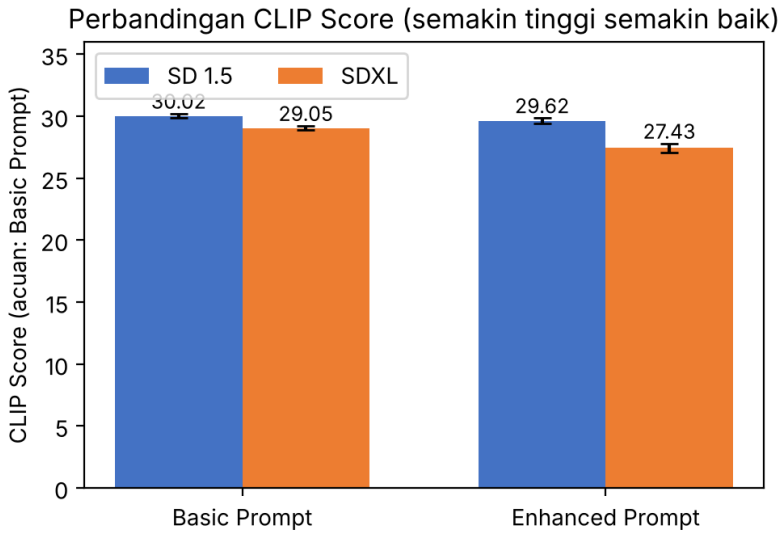
Tabel 4 menyajikan FID untuk setiap kombinasi dan CLIP Score yang dihitung terhadap Basic Prompt sebagai teks acuan. SD 1.5 dengan Enhanced Prompt memiliki FID terendah (296,714), diikuti SDXL-Basic (304,446), SD 1.5-Basic (321,053), dan SDXL-Enhanced (330,069). Urutan ini sesuai dengan peringkat rata-rata pada lima kelompok seed (berturut-turut 1,2; 2,0; 3,0; dan 3,8). SD 1.5-Enhanced menempati peringkat pertama pada empat dari lima kelompok, dan SDXL-Enhanced peringkat terakhir pada empat dari lima kelompok. Namun, rentang FID antar kelompok seed saling tumpang tindih: SD 1.5-Enhanced (309,9-329,7) dengan SDXL-Basic (311,3-339,3), serta SD 1.5-Basic (333,0-370,6) dengan SDXL-Enhanced (326,6-365,2). Dengan demikian, selisih antar kombinasi yang berdekatan tidak dapat ditafsirkan sebagai perbedaan yang pasti; yang terlihat hanyalah kecenderungan pemisahan antara dua kombinasi dengan FID lebih rendah (SD 1.5-Enhanced dan SDXL-Basic) dan dua kombinasi dengan FID lebih tinggi (SD 1.5-Basic dan SDXL-Enhanced).

Pada CLIP Score yang dihitung terhadap Basic Prompt, SD 1.5-Basic memperoleh skor tertinggi (30,02), diikuti SD 1.5-Enhanced (29,62), SDXL-Basic (29,05), dan SDXL-Enhanced (27,43). Dengan teks acuan yang sama, Enhanced Prompt tidak meningkatkan CLIP Score pada kedua model. Perlu dicatat bahwa Basic Prompt sebagai acuan cenderung menguntungkan gambar yang dibangkitkan dari prompt tersebut, sehingga hasil ini perlu dibaca bersama analisis sensitivitas pada Tabel 5. Perbandingan visual disajikan pada Gambar 3 untuk FID dan Gambar 4 untuk CLIP Score.

 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 3 Perbandingan FID antar kombinasi



Gambar 4 Perbandingan CLIP Score terhadap basic prompt antar kombinasi

Sebagaimana ditunjukkan pada Gambar 3, Enhanced Prompt menurunkan FID pada SD 1.5 (321,053 menjadi 296,714; turun 24,3 poin atau sekitar 7,6%), yang berarti distribusi visual gambar hasil generasi menjadi sedikit lebih dekat dengan batik asli. Sebaliknya, pada SDXL FID naik (304,446 menjadi 330,069; naik 25,6 poin atau sekitar 8,4%). Pola yang berlawanan ini mengindikasikan bahwa pengaruh prompt enhancement terhadap kualitas visual bersifat spesifik terhadap model, bukan efek universal. Pada Gambar 4, CLIP Score terhadap Basic Prompt menurun pada kedua model ketika memakai Enhanced Prompt, dengan penurunan yang lebih besar pada SDXL (-1,61) dibandingkan SD 1.5 (-0,40); garis galat pada gambar tersebut menunjukkan interval kepercayaan 95% antar motif.

Nilai FID pada seluruh kombinasi tergolong tinggi (297 hingga 330). Sebagai pembanding, penelitian sintesis motif kain Endek melaporkan FID baseline 394,21 yang turun menjadi 276,05 setelah adaptasi LoRA [2], sehingga FID yang tinggi pada model generik tanpa adaptasi tampaknya merupakan karakteristik umum pada domain tekstil tradisional dan sejalan dengan pendekatan zero-shot yang dipakai. Nilai FID antar penelitian tidak dapat dibandingkan langsung karena dataset acuannya berbeda, dan estimasi FID peka terhadap jumlah sampel karena 375 gambar hasil generasi

 ISSN:2302-8149
e-ISSN:2540-9719

dan 240 gambar acuan lebih sedikit daripada dimensi fitur Inception (2.048) [10]. Karena itu, pembahasan difokuskan pada pola perubahan akibat perlakuan, bukan pada besaran nilai absolut.

Hasil uji statistik. Uji-t berpasangan pada tingkat motif (n = 15; skor setiap motif adalah rerata 25 seed) dilakukan untuk menguji perbedaan CLIP Score antara Basic Prompt dan Enhanced Prompt pada tiap model, dengan dua teks acuan: Basic Prompt (analisis utama) dan prompt masing-masing gambar (analisis sensitivitas). Hasilnya ditunjukkan pada Tabel 5.

Tabel 5 Hasil Uji-t berpasangan terhadap CLIP score

Model	Teks acuan	Rerata Basic	Rerata Enhanced	Selisih	t (db = 14)	p
Stable Diffusion 1.5	Basic Prompt	30,02	29,62	−0,40	−3,091	0,0080
SDXL	Basic Prompt	29,05	27,43	−1,61	−8,486	< 0,001
Stable Diffusion 1.5	Prompt masing-masing	30,02	31,97	+1,95	11,858	< 0,001
SDXL	Prompt masing-masing	29,05	31,53	+2,49	8,214	< 0,001

Berdasarkan Tabel 5, dengan Basic Prompt sebagai teks acuan yang sama, Enhanced Prompt menurunkan CLIP Score pada kedua model. Pada SD 1.5 penurunannya kecil tetapi signifikan (selisih -0,40; p = 0,008; uji Wilcoxon p = 0,008; Cohen's dz = -0,80), dan pada SDXL lebih besar (selisih -1,61; p < 0,001; dz = -2,19). Sebaliknya, ketika setiap gambar dinilai dengan prompt yang membangkitkannya, Enhanced Prompt tampak meningkatkan skor (SD 1.5: +1,95; SDXL: +2,49; keduanya p < 0,001). Perbedaan arah ini menunjukkan bahwa kenaikan tersebut tidak dapat dibaca sebagai gambar yang lebih sesuai, karena teks pembandingnya berbeda. Dengan acuan Basic Prompt, skor SDXL lebih rendah daripada SD 1.5 pada Basic Prompt (selisih -0,98; p < 0,001) maupun Enhanced Prompt (selisih -2,19; p < 0,001). Nilai p yang kecil sebagian berasal dari perataan 25 seed per motif; inferensi hanya berlaku untuk 15 motif yang diuji dan tidak dikoreksi untuk pengujian berganda, sehingga p = 0,008 pada SD 1.5 perlu ditafsirkan dengan hati-hati.

Panjang prompt. Basic Prompt memiliki 29–31 token (rerata 30,0) sehingga tidak terpotong. Sebaliknya, seluruh 15 Enhanced Prompt (100%) melampaui batas 77 token encoder teks CLIP, dengan panjang 83 hingga 100 token (rerata 90,6 token atau sekitar 59,5 kata). Batas kurang dari 75 kata pada instruksi Gemini tidak menjamin batas token, karena tanda baca dan kata bertanda hubung dihitung sebagai token tersendiri. Sebanyak 6 sampai 23 token (rerata 13,6) di akhir setiap Enhanced Prompt terpotong saat diproses oleh kedua model. Bagian yang terpotong umumnya berupa deskripsi tekstur kain dan penutup kualitas yang generik (misalnya highly detailed atau seamless pattern), sedangkan deskripsi motif dan warna di bagian awal prompt sebagian besar tetap terbaca; pada beberapa prompt, kata warna di bagian akhir (misalnya ivory) ikut terpotong.

Pembahasan. Hasil penelitian ini menunjukkan dua temuan utama. Pertama, pengaruh prompt enhancement berbasis LLM bergantung pada model dan metrik. Dengan teks acuan yang sama, Enhanced Prompt tidak meningkatkan kesesuaian semantik (CLIP Score) pada kedua model, sedangkan peningkatan pada penilaian dengan prompt masing-masing terutama mencerminkan perbedaan teks pembanding. Pada kualitas visual (FID), Enhanced Prompt menurunkan FID pada SD 1.5 (sekitar 7,6%) tetapi menaikkannya pada SDXL (sekitar 8,4%), dengan catatan bahwa rentang FID antar kelompok seed saling tumpang tindih sehingga selisih tersebut bersifat indikatif. Pola ini tidak sejalan dengan penelitian optimasi prompt pada domain umum yang melaporkan peningkatan kualitas gambar [5]. Penyebab penurunan pada SDXL tidak dapat dipastikan dari desain penelitian ini; dua dugaan yang perlu diuji adalah (i) pemotongan Enhanced Prompt pada batas 77 token pada kedua text encoder SDXL [8] dan (ii) perbedaan resolusi dan guidance scale bawaan antar model. Pada SD 1.5, penurunan FID sejalan dengan prinsip bahwa pengetahuan domain yang terstruktur pada penyusunan prompt dapat mengarahkan hasil mendekati otentisitas budaya [19], tetapi prinsip ini tidak berlaku pada SDXL sehingga tidak dapat digeneralisasi tanpa memperhatikan model.

 ISSN:2302-8149
e-ISSN:2540-9719

Kedua, tingginya FID pada seluruh kombinasi memperkuat argumen bahwa model text-to-image pretrained generik terbatas dalam merepresentasikan domain budaya spesifik seperti motif Batik Nitik tanpa penyesuaian tambahan, sejalan dengan penelitian serupa pada tekstil tradisional lain [1], [3] dan dengan temuan bahwa model text-to-image sistematis kesulitan memenuhi ekspektasi budaya implisit pada komunitas yang kurang terwakili dalam data latih [18]. Dibandingkan dengan penelitian yang melakukan fine-tuning pada ribuan hingga puluhan ribu gambar [1], [2], pengayaan prompt dengan LLM hanya memberi perbaikan FID terbatas pada SD 1.5 dan tidak memperbaiki kesesuaian semantik, sehingga belum dapat menggantikan fine-tuning. Bagi pengguna dengan sumber daya komputasi terbatas, seperti pelaku UMKM atau desainer independen, prompt yang lebih panjang dan rinci tidak otomatis menghasilkan gambar yang lebih baik, sehingga efeknya perlu diperiksa pada model yang dipakai. Keunikan penelitian ini terletak pada kombinasi tiga hal yang belum dikaji bersamaan: perbandingan dua generasi model pretrained secara zero-shot, pemanfaatan LLM sebagai pengaya prompt berbasis pengetahuan motif budaya, dan protokol evaluasi dengan seed tetap, banyak gambar per motif, serta pengujian statistik pada tingkat motif.

Rekomendasi panjang prompt. Berdasarkan temuan bahwa seluruh Enhanced Prompt terpotong, penulis menyarankan agar prompt enhancement dibatasi dengan jumlah token, bukan jumlah kata. Instruksi ke LLM sebaiknya menetapkan batas sekitar 70-75 token (kira-kira 45-50 kata, dengan rasio sekitar 1,5 token per kata seperti pada penelitian ini), menempatkan ciri visual dan warna khas motif di awal prompt, dan menaruh deskripsi umum seperti kualitas dan tekstur di akhir. Panjang prompt sebaiknya diperiksa dengan tokenizer CLIP sebelum dipakai. Rekomendasi ini berasal dari pengamatan pada penelitian ini dan belum diuji secara eksperimental; pengaruh prompt yang lebih pendek terhadap FID dan CLIP Score perlu diuji pada penelitian lanjutan.

Keterbatasan. Pertama, penelitian hanya mencakup 15 motif dari satu dataset Batik Nitik sehingga hasil belum tentu berlaku untuk motif atau jenis batik lain. Kedua, FID dihitung dari 375 gambar per kombinasi dan 240 gambar acuan, lebih sedikit daripada dimensi fitur Inception (2.048), sehingga nilainya kurang stabil [10], selisih kecil antar kombinasi tidak boleh ditekankan, dan FID tidak diuji secara inferensial. Ketiga, kedua teks acuan CLIP Score tidak netral dan CLIP ViT-B/32 hanya memproses 77 token pertama. Keempat, Enhanced Prompt dihasilkan satu kali per motif sehingga variasi keluaran LLM tidak dievaluasi. Kelima, SD 1.5 dan SDXL dijalankan pada resolusi dan guidance scale bawaan masing-masing sehingga perbedaan antar model tidak murni berasal dari arsitektur. Keenam, 1.500 gambar tidak diperiksa secara sistematis untuk artefak seperti watermark yang dapat muncul akibat memorization [17], dan evaluasi hanya memakai metrik otomatis tanpa penilaian pakar batik.

5 Kesimpulan

Penelitian ini membandingkan Stable Diffusion 1.5 dan SDXL dalam membuat motif Batik Nitik Yogyakarta tanpa fine-tuning, serta menguji pengaruh prompt enhancement berbasis Large Language Model terhadap kualitas visual dan kesesuaian semantik. Hasilnya menunjukkan bahwa manfaat prompt enhancement bersifat spesifik model dan metrik. Pada SD 1.5, prompt enhancement mendekatkan distribusi visual hasil generasi ke batik asli, sedangkan pada SDXL justru menjauhkannya. Dengan teks acuan yang sama, prompt enhancement tidak meningkatkan kesesuaian semantik pada kedua model; kenaikan skor yang tampak ketika setiap gambar dinilai dengan prompt masing-masing lebih mencerminkan perbedaan teks pembanding daripada gambar yang lebih sesuai. Seluruh Enhanced Prompt juga melampaui batas 77 token sehingga terpotong, sehingga panjang prompt perlu dikendalikan dengan jumlah token. Nilai FID yang tinggi pada semua kombinasi menunjukkan bahwa model pretrained generik masih terbatas dalam merepresentasikan motif batik tanpa penyesuaian lebih lanjut. Penelitian berikutnya disarankan memperluas jumlah motif dan jenis batik, menguji Enhanced Prompt yang dibatasi pada jumlah token tertentu, menyamakan resolusi dan guidance scale antar model, serta menambahkan evaluasi pakar batik dan perbandingan dengan fine-tuning berskala kecil.

Referensi

[1] R. D. I. Wahid, N. Yudistira, C. Dewi, I. N. Sari, D. Pradhikta, and Fatmawati, "*Prompt*

Conditioned Batik Pattern Generation using LoRA Weighted Diffusion Model with Classifier-Free Guidance," IEEE Access, Vol. 13, pp. 2436–2448, 2025, DOI: 10.1109/ACCESS.2024.3523494.

[2] I. K. A. Purnawan and I. K. G. J. Suparjana, "*Fine-Tuning Text-to-Image Diffusion Models with LoRA for Generative Synthesis of Balinese Endek Textile Patterns*," Jurnal Nasional Pendidikan Teknik Informatika: JANAPATI, Vol. 15, No. 1, pp. 46–65, 2026, DOI: 10.23887/janapati.v15i1.111817.

[3] Q. Hu, Y. Peng, J. Xu, Z. Shao, Z. Tian, and J. Chen, "*Adaptive Stylized Image Generation for Traditional Miao Batik using Style-Conditioned LCM-LoRA Enhanced Diffusion Models*," Mathematics, Vol. 13, No. 12, 2025, DOI: 10.3390/math13121947.

[4] O. Octadion, N. Yudistira, and D. Kurnianingtyas, "*Synthesis of Batik Motifs using a Diffusion-Generative Adversarial Network*," Multimedia Tools and Applications, Vol. 84, pp. 3407–3438, 2025, Doi: 10.1007/s11042-025-20620-9.

[5] W. Li, J. Wang, and X. Zhang, "*PROMPTIST: Automated Prompt Optimization for Text-to-Image Synthesis*," in Natural Language Processing and Chinese Computing, NLPCC 2024, LNCS Vol. 15360, Springer, Singapore, 2025, DOI: 10.1007/978-981-97-9434-8_23.

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "*High-Resolution Image Synthesis with Latent Diffusion Models*," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2022, pp. 10684–10695.

[7] E. J. Hu et al., "*LoRA: Low-Rank Adaptation of Large Language Models*," in Proc. Int. Conf. Learn. Represent. (ICLR), 2022.

[8] D. Podell et al., "*SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis*," in Proc. Int. Conf. Learn. Represent. (ICLR), 2024.

[9] A. Radford et al., "*Learning Transferable Visual Models from Natural Language Supervision*," in Proc. 38th Int. Conf. Mach. Learn. (ICML), PMLR Vol. 139, 2021, pp. 8748–8763.

[10] S. Jayasumana, S. Ramalingam, A. Veit, D. Glasner, A. Chakrabarti, and S. Kumar, "*Rethinking FID: Towards a Better Evaluation Metric for Image Generation*," in Proc. IEEE/CVF CVPR, 2024, pp. 9307–9315, DOI: 10.1109/CVPR52733.2024.00889.

[11] H. Gani, S. F. Bhat, M. Naseer, S. Khan, and P. Wonka, "*LLM Blueprint: Enabling Text-to-Image Generation with Complex and Detailed Prompts*," in Proc. ICLR, 2024.

[12] A. E. Minarno, H. A. Nugroho, and I. Soesanti, "*Batik Nitik 960*," Mendeley Data, V3, 2023, DOI: 10.17632/sgh484jxzy.3.

[13] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "*GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium*," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Vol. 30, 2017.

[14] Hugging Face, "*Diffusers: State-of-the-Art Diffusion Models*," GitHub repository, 2022. [Online]. Available: https://github.com/huggingface/diffusers

[15] Google DeepMind, "*Gemini: A Family of Highly Capable Multimodal Models*," arXiv:2312.11805, 2023.

[16] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "*CLIPScore: A Reference-Free Evaluation Metric for Image Captioning*," in Proc. 2021 Conf. Empirical Methods in Natural Language Processing (EMNLP), 2021, pp. 7514–7528, DOI: 10.18653/v1/2021.emnlp-main.595.

[17] N. Carlini et al., "*Extracting Training Data from Diffusion Models*," in Proc. 32nd USENIX Security Symposium, 2023, arXiv:2301.13188.

[18] S. Nayak, M. Bhatia, X. Zhang, V. Rieser, L. A. Hendricks, S. van Steenkiste, Y. Goyal, K. Stańczak, and A. Agrawal, "*CulturalFrames: Assessing Cultural Expectation Alignment in Text-to-Image Models and Evaluation Metrics*," in Findings of the Association for Computational Linguistics: EMNLP 2025, 2025, pp. 20918–20953.

[19] K. Arzomand, T. Kalganova, and M. J. F. Rustell, "*HARF: A Human–AI Collaborative Framework for Cultural Heritage Reconstruction with Expert-Guided Multi-Platform Generative AI and Systematic Prompt Engineering*," Digital Applications in Archaeology and Cultural Heritage, Vol. 42, e00554, 2026, DOI: 10.1016/j.daach.2026.e00554.

[20] A. E. Minarno, I. Soesanti, and H. A. Nugroho, "*Batik Nitik 960 Dataset for Classification, Retrieval, and Generato*r," Data, Vol. 8, No. 4, Art. No. 63, 2023, DOI: 10.3390/data8040063.