GENERASI KARAKTER FANTASI BERGAYA *ART NOUVEAU* DENGAN *FINE-TUNING DREAMBOOTH* DAN REGULARISASI *DROPOUT*

LAPORAN TUGAS AKHIR

Laporan ini Disusun untuk Memenuhi Salah Satu Syarat Memperoleh Gelar Sarjana Strata 1 (S1) pada Program Studi Teknik Informatika Fakultas Teknologi Industri Universitas Islam Sultan Agung Semarang

DI SUSUN OLEH :
YUNITA ENDAH SULISTIYOWATI
32602100125

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INDUSTRI
UNIVERSITAS ISLAM SULTAN AGUNG
SEMARANG
2025

FINAL PROJECT

GENERATION OF FANTASY CHARACTERS IN ART NOUVEAU STYLE WITH FINE-TUNING DREAMBOOTH AND DROPOUT REGULARIZATION

Proposed to complete the requirement to obtain a bachelor's degree (S-1) at Informatics Engineering Departement of Industrial Technology Faculty Sultan Agung Islamic University

ARRANGED BY :
YUNITA ENDAH SULISTIYOWATI
32602100125

MAJORING OF INFORMATICS ENGINEERING
INDUSTRIAL TECHNOLOGY FACULTY
SULTAN AGUNG ISLAMIC UNIVERSITY
SEMARANG

2025

LEMBAR PENGESAHAN
TUGAS AKHIR

GENERASI KARAKTER FANTASI BERGAYA ART NOUVEAU DENGAN FINE-TUNING DREAMBOOTH DAN REGULARISASI DROPOUT

YUNITA ENDAH SULISTIYOWATI
32602100125

Telah dipertahankan di depan tim penguji ujian sarjana tugas akhir
Program Studi Teknik Informatika
Universitas Islam Sultan Agung
Pada tanggal : 17 Februari 2025

TIM PENGUJI UJIAN SARJANA :

Sam Farisa Chaerul Haviana, ST, M.Kom. NIDN. 0628028602 (Penguji 1)	.....	17 Februari 2025 .....
Ghufron, ST, M.Kom NIDN. 0602079005 (Penguji 2)	.....	17 Februari 2025 .....
Ir. Sri Mulyono, M.Eng NIDN. 0626066601 (Pembimbing)	.....	17 Februari 2025 .....

Semarang, 17 Februari 2025
Mengetahui,
Kaprodi Teknik Informatika
Universitas Islam Sultan Agung

Moch. Taufik, ST., MIT
NIDN. 0622037502

SURAT PERNYATAAN KEASLIAN TUGAS AKHIR

Yang bertanda tangan dibawah ini :

Nama : Yunita Endah Sulistiyowati

NIM : 32602100125

Judul Tugas Akhir : GENERASI KARAKTER FANTASI BERGAYA *ART NOUVEAU* DENGAN *FINE-TUNING DREAMBOOTH* DAN REGULARISASI *DROPOUT*

Dengan bahwa ini saya menyatakan bahwa judul dan isi Tugas Akhir yang saya buat dalam rangka menyelesaikan Pendidikan Strata Satu (S1) Teknik Informatika tersebut adalah asli dan belum pernah diangkat, ditulis ataupun dipublikasikan oleh siapapun baik keseluruhan maupun sebagian, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka, dan apbila di kemudian hari ternyata terbukti bahwa judul Tugas Akhir tersebut pernah diangkat, ditulis ataupun dipublikasikan, maka saya bersedia dikenakan sanksi akademis. Demikian surat pernyataan ini saya buat dengan sadar dan penuh tanggung jawab.

Semarang, 5 Maret 2025

Yang Menyatakan,

Yunita Endah Sulistiyowati

PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH

Saya yang bertanda tangan dibawah ini :

Nama : Yunita Endah Sulistiyowati

NIM : 32602100125

Program Studi : Teknik Informatika

Fakultas : Teknologi industri

Dengan ini menyatakan Karya Ilmiah berupa Tugas akhir dengan Judul : GENERASI KARAKTER FANTASI BERGAYA *ART NOUVEAU* DENGAN *FINE-TUNING DREAMBOOTH* DAN REGULARISASI *DROPOUT*. Menyetujui menjadi hak milik Universitas Islam Sultan Agung serta memberikan Hak bebas Royalti Non-Eksklusif untuk disimpan, dialihmediakan, dikelola dan pangkalan data dan dipublikasikan diinternet dan media lain untuk kepentingan akademis selama tetap menyantumkan nama penulis sebagai pemilik hak cipta. Pernyataan ini saya buat dengan sungguh-sungguh. Apabila dikemudian hari terbukti ada pelanggaran Hak Cipta/Plagiarisme dalam karya ilmiah ini, maka segala bentuk tuntutan hukum yang timbul akan saya tanggung secara pribadi tanpa melibatkan Universitas Islam Sultan agung.

Semarang, 5 Maret 2025

Yunita Endah Sulistiyowati

KATA PENGANTAR

Dengan mengucap syukur alhamdulillah atas kehadirat Allah SWT yang telah memberikan rahmat dan karunianya kepada penulis, sehingga dapat menyelesaikan Tugas Akhir dengan judul "Generasi Karakter Fantasi Bergaya *Art Nouveau* Dengan Fine-Tuning *Dreambooth* Dan Regularisasi Dropout" ini untuk memenuhi salah satu syarat menyelesaikan studi serta dalam rangka memperoleh gelar sarjana (S-1) pada Program Studi Teknik Informatika Fakultas Teknologi Industri Universitas Islam Sultan Agung Semarang.

Tugas Akhir ini disusun dan dibuat dengan adanya bantuan dari berbagai pihak, materi maupun teknis, oleh karena itu saya selaku penulis mengucapkan terima kasih kepada:

1. Rektor UNISSULA Bapak Prof. Dr. H. Gunarto, S.H., M.H yang mengizinkan penulis menimba ilmu di kampus ini.
2. Dekan Fakultas Teknologi Industri Ibu Dr. Novi Marlyana, S.T., M.T.
3. Dosen pembimbing Ir. Sri Mulyono M. Eng yang telah meluangkan waktu dan memberi ilmu.
4. Orang tua penulis yang telah mengizinkan untuk menyelesaikan laporan ini,
5. Dan kepada semua pihak yang tidak dapat saya sebutkan satu persatu.

Dengan segala kerendahan hati, penulis menyadari masih terdapat banyak kekurangan dari segi kualitas atau kuantitas maupun dari ilmu pengetahuan dalam penyusunan laporan, sehingga penulis mengharapkan adanya saran dan kritikan yang bersifat membangun demi kesempurnaan laporan ini dan masa mendatang.

Semarang,

Yunita Endah Sulistiyowati

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

ABSTRAK

Kebutuhan visualisasi karakter dalam novel fantasi sering kali sulit divisualisasikan hanya melalui deskripsi teks. Teknik fine-tuning *Dreambooth* pada model *Stable Diffusion* v1.5 mengatasi tantangan tersebut dengan menyesuaikan model terhadap *dataset* spesifik. Dalam penelitian ini, penerapan regularisasi Dropout pada CLIP *encoder* meningkatkan fidelity hasil visualisasi, meskipun mengurangi skor CLIP dari 0.80-an menjadi sekitar 0.70-an. Model ini mampu menghasilkan gambar karakter bergaya *Art Nouveau* dengan elemen khas seperti garis melengkung, motif alam, dan palet warna pastel. Sistem berbasis web memungkinkan pengguna untuk memasukkan deskripsi karakter dan memperoleh visualisasi yang sesuai. Hasil penelitian menunjukkan bahwa fine-tuning dan Dropout efektif menghasilkan representasi visual yang lebih akurat, meskipun karakter laki-laki dalam *dataset* terbatas. Penelitian ini berkontribusi pada pengembangan *Text-to-Image Generation* dalam industri kreatif.

Kata Kunci: *Stable Diffusion*, *Dreambooth* Fine-Tuning, *Dropout* Regularization, *Art Nouveau, Text-to-Image Generation.*

ABSTRACT

The need for character visualization in fantasy novels is often difficult to achieve through text descriptions alone. The fine-tuning technique using Dreambooth on the Stable Diffusion v1.5 model addresses this challenge by adapting the model to a specific dataset. In this study, the application of Dropout regularization on the CLIP encoder improved the fidelity of the generated visualizations, although it reduced the CLIP score from around 0.80 to approximately 0.7. This model is capable of generating character images in the Art Nouveau style, featuring characteristic elements such as curving lines, natural motifs, and pastel color palettes. A web-based system allows users to input character descriptions and obtain corresponding visualizations. The findings show that fine-tuning and Dropout are effective in producing more accurate visual representations, despite the limited availability of male characters in the dataset. This research contributes to the development of Text-to-Image Generation in the creative industry.

Keywords: Stable Diffusion, Dreambooth Fine-Tuning, Dropout Regularization, Art Nouveau, Text-to-Image Generation.

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Seiring dengan berkembangnya teknologi di bidang kecerdasan buatan (AI), muncul inovasi dalam pengolahan data berbasis deep learning termasuk didalamnya *text-to-Image generation*. Salah satu teknologi yang ada adalah *Diffusion Models*, yang melakukan pendekatan secara generatif untuk mengubah *input* teks menjadi gambar melalui proses yang bertahap dan berulang (Prasad *dkk.*, 2024). Model ini dikembangkan agar dapat membantu manusia dalam menciptakan visualisasi otomatis dari deskripsi teks, terutama dalam konteks kreatif seperti ilustrasi untuk sastra dan seni digital. Salah satu implementasi *Diffusion Models, Stable Diffusion* memiliki kemampuan untuk menghasilkan gambar berdasarkan deskripsi teks. Model ini menggunakan jaringan *neural* untuk secara progresif mengurangi *noise* pada gambar acak hingga membentuk representasi visual yang jelas dan sesuai dengan *input* teks.

Meski begitu tantangan terkait fidelitas (kejelasan) subjek masih menjadi perhatian utama, mendorong untuk terus berimprovisasi tentang bagaimana menjaga keutuhan karakteristik visual dalam konteks personalisasi gambar Untuk memaksimalkan performa model ini dalam konteks menggambarkan karakter dalam novel fantasi, diperlukan pendekatan khusus, salah satunya dengan menggunakan teknik *fine-tuning* (Wu *dkk.*, 2023). Model yang telah dilatih disesuaikan kembali menggunakan *Dataset* yang lebih spesifik. Dalam hal ini, *fine-tuning* dilakukan melalui *Dreambooth* untuk mengekstraksi fitur-fitur penting dari deskripsi teks untuk membantu model memahami elemen kunci yang harus divisualisasikan, seperti ciri fisik karakter, pakaian, atau atribut lain yang melekat pada latar dunia fantasi.

Dreambooth, sebagai salah satu teknik *fine-tuning*, telah memberikan pengaruh pada dunia seni digital dengan memungkinkan penciptaan gambar yang lebih mendalam dan detail dengan acuan pada deskripsi teks (Niu *dkk.*, 2024). Teknik ini memungkinkan seniman dan kreator untuk melatih *model*

diffusion agar lebih responsif terhadap karakteristik spesifik, seperti ekspresi wajah, tekstur, atau nuansa warna yang sesuai dengan deskripsi karakter atau dunia fiksi yang diinginkan.

Teknologi ini dirasa relevan untuk mengatasi tantangan yang dihadapi dalam industri kreatif, khususnya dalam novel bergenre fantasi karena pembaca sering kali kesulitan membayangkan karakter secara tepat hanya dari deskripsi teks. Tema *Art Nouveau* dipilih dalam penerapan teknologi *Diffusion Models* karena memiliki gaya visual yang khas dengan garis melengkung yang elegan, motif alam yang organik, dan desain yang sangat detail. Gaya ini dikenal dengan penggunaan elemen dekoratif yang mengalir, seperti bunga, daun, dan garis-garis melingkar yang terinspirasi dari alam, serta sering kali menampilkan wanita dengan ekspresi lembut dan penuh keindahan (Saraswati, Utami dan Pemayun, 2024). Karakter-karakter dalam gaya *Art Nouveau* tidak hanya menggambarkan keindahan visual, tetapi juga menyampaikan suasana misterius dan romantis, yang menciptakan dunia fantasi yang indah dan memikat.

Dalam konteks novel fantasi, tema *Art Nouveau* memberikan daya tarik unik yang dapat membantu pembaca membayangkan dunia yang penuh dengan keanggunan dan keindahan alam yang magis (Gumulya, 2022). Dengan menyesuaikan generasi gambar menggunakan gaya ini, penelitian bertujuan menghasilkan visualisasi yang tidak hanya relevan secara naratif tetapi juga memikat secara estetika, menjembatani kesenjangan antara deskripsi teks dan imajinasi pembaca. Gambar yang dihasilkan diharapkan mencerminkan kesan fantasi yang elegan, penuh dengan keindahan organik, dan menyatu dengan alam.

Implementasi sistem berbasis web dalam penelitian ini dirancang untuk memudahkan pengguna menghasilkan gambar karakter bertema *Art Nouveau* hanya dengan memasukkan deskripsi teks sebagai *input* (Maulana, 2022). Pengguna dapat memasukkan detail karakter seperti sifat fisik, pakaian, hingga atribut unik lainnya, yang kemudian akan diolah oleh model *Stable Diffusion* v1.5 yang telah di-*fine-tune* menggunakan *Dreambooth*. Sistem ini

tidak hanya bertujuan untuk mendukung kebutuhan kreatif di bidang seni digital, tetapi juga untuk menunjukkan potensi teknologi AI dalam memberikan solusi inovatif dan praktis dalam memvisualisasikan ide-ide kreatif secara otomatis dan efisien.

1.2 Perumusan Masalah

Rumusan masalah dalam penelitian ini adalah sebagai berikut :

a. Bagaimana cara menerjemahkan deskripsi karakter yang berupa teks menjadi gambar menggunakan *Diffusion Models* dengan gaya *Art Nouveau*?
b. Bagaimana cara agar visualisasi yang dihasilkan mendekati interpretasi yang diinginkan penulis?

1.3 Pembatasan Masalah

Pembatasan masalah di bawah ini bertujuan untuk menghindari adanya kegiatan di luar sasaran, sehingga dalam pembuatan laporan ini perlu ditentukan suatu batasan masalah sebagai berikut

a. Penelitian ini hanya menggunakan *Stable Diffusion* v1.5 sebagai model utama untuk generasi gambar karakter.
b. Metode *Dreambooth* digunakan sebagai pendekatan untuk melakukan *fine-tuning*.
c. Sistem berbentuk aplikasi berbasis web di mana pengguna hanya dapat memasukkan deskripsi teks (*prompt*) dalam bahasa Inggris melalui antarmuka yang disediakan. *Input* lain seperti sketsa, gambar referensi, atau audio tidak termasuk dalam cakupan penelitian.
d. Hasil akhir berupa pembuatan gambar karakter berdasarkan deskripsi dan tidak termasuk lingkungan atau latar.
e. Model hanya difokuskan pada visualisasi karakter yang memiliki gaya *Art Nouveau* dengan pendekatan visual 2D, tanpa menyertakan elemen animasi atau gerakan.

f. Sistem hanya mendukung generasi gambar satu karakter per-*input*, tanpa kemampuan untuk menghasilkan beberapa karakter dalam satu proses.
g. Sistem tidak menyediakan fitur editing gambar secara langsung; pengguna hanya menerima hasil akhir sesuai dengan *prompt* yang dimasukkan.
h. Waktu pemrosesan hasil tergantung pada kompleksitas *prompt* dan kapasitas *server* yang digunakan.

1.4 Tujuan

Adapun tujuan dari penelitian ini adalah :

a. Menerjemahkan deskripsi karakter yang berupa teks menjadi gambar bergaya *Art Nouveau* menggunakan *Diffusion Models*.
b. Menghasilkan visualisasi yang dapat mendekati interpretasi yang diinginkan penulis.

1.5 Manfaat

Manfaat dari penelitian ini memberikan kemudahan bagi kreator dalam menghasilkan visualisasi karakter berdasarkan deskripsi teks, akses yang terjangkau untuk menciptakan ilustrasi berbasis AI melalui sistem berbasis web, kualitas visual yang konsisten dengan gaya *Art Nouveau* dengan dukungan personalisasi melalui *fine-tuning* menggunakan *Dreambooth*, serta membantu penulis dalam menyampaikan deskripsi karakter dari ceritanya melalui media visual.

1.6 Sistematika Penulisan

Untuk mempermudah penulisan tugas akhir ini, penulis membuat suatu sistematika yang terdiri dari:

BAB 1 : PENDAHULUAN

Bab ini menjelaskan mengenai latar belakang pemilihan judul tugas akhir "Generasi Karakter Fantasi Bergaya *Art Nouveau* Dengan *Fine-tuning Dreambooth* Dan Regularisasi *Drop-out*". Rumusan masalah,

batasan masalah, tujuan penelitian, metodologi penelitian, dan sistematika penulisan.

BAB 2 : TINJAUAN PUSTAKA DAN DASAR TEORI

Bab ini memuat dasar teori yang berfungsi sebagai sumber dalam memahami permasalahan yang dipilih.

BAB 3 : METODE PENELITIAN

Bab ini menjelaskan proses tahapan- tahapan penelitian dimulai dari analisa kebutuhan sistem, kemudian perancangan sistem hingga selesai dibuat.

BAB 4: HASIL PENELITIAN DAN IMPLEMENTASI SISTEM

Bab ini menjelaskan mengungkapkan hasil penelitian yang berupa pembuatan gambar karakter dari novel fantasi berdasarkan deskripsi teks menggunakan *Diffusion Models*.

BAB 5: KESIMPULAN DAN SARAN

Bab ini memuat kesimpulan dari keseluruhan uraian bab-bab sebelumnya dan saran-saran dari hasil yang diperoleh dan diharapkan dapat bermanfaat dalam penelitian selanjutnya.

BAB II
TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Generative Artificial Intelligence (Gen AI) telah merevolusi berbagai bidang dengan kemampuannya menghasilkan media yang menyerupai karya manusia, tetapi juga menghadirkan tantangan terkait transparansi, prediktabilitas, dan etika. Dalam penelitian berjudul "*Custom Concept Text-to-Image Using Stable Diffusion Model in Generative Artificial Intelligence*" mengenai penggunaan *model diffusion* untuk menghasilkan gambar dari teks, meskipun hasil gambar realistis, terdapatkan beberapa faktor yang memengaruhi hasil, seperti kebutuhan *dataset* , waktu komputasi dan kualitas uji FID. *Frechet Inception Distance* (FID), menunjukkan nilai tinggi sebesar 1284.4430 bergantung pada ketersediaan *dataset* yang cukup besar untuk pelatihan serta kebutuhan komputasi yang tinggi menjadi kendala signifikan dalam meningkatkan performa model (Rahmatulloh, 2024).

Pengembangan sistem pengenalan teks saat ini bergantung pada sintesis dan augmentasi gambar karena sulitnya mencakup kompleksitas dan keragaman dunia nyata melalui pengumpulan dan anotasi gambar teks nyata. Dalam penelitian dengan judul "*Conditional Text Image Generation with Diffusion Models* (CTIG-DM)" digunakan tiga kondisi (gambar, teks, dan gaya) untuk dapat mengontrol atribut, isi, dan gaya dalam proses generasi gambar teks. Melalui konfigurasi ketiga kondisi ini, empat mode generasi gambar teks dapat diperoleh: mode sintesis, augmentasi, pemulihan, dan imitasi (Zhu *dkk.*, 2023). Metode ini menawarkan fleksibilitas lebih besar dalam mengontrol atribut visual dari gambar yang dihasilkan. Namun, aplikasi dunia nyata dari metode ini masih perlu diuji lebih lanjut untuk memastikan kelayakannya di berbagai skenario kompleks.

Penelitian lain berjudul *"Denoising Diffusion Probabilistic Models (DDPM) Dynamics: Unraveling Change Detection in Evolving Environments*" mengkaji aplikasi *Denoising Diffusion Probabilistic Models* (DDPM) dalam deteksi perubahan pada lingkungan yang berkembang secara

dinamis. Model ini dapat dipergunakan dalam menangkap perubahan halus dan variasi visual, sehingga dapat diandalkan untuk tugas-tugas seperti pengawasan, pemantauan lingkungan, dan deteksi anomali. Meskipun lebih berfokus pada dinamika temporal, metode ini relevan untuk mendeteksi dan menyesuaikan perubahan halus dalam karakter novel yang mungkin muncul dalam variasi kecil namun signifikan, misalnya dalam hal pencahayaan atau pose (Anderson dan Akram, 2024).

Model *text-to-Image* berskala besar telah memungkinkan sintesis gambar berkualitas tinggi dan beragam dari *prompt* teks, tetapi masih memiliki keterbatasan dalam mereplikasi tampilan subjek tertentu dan menghasilkan versi baru dalam konteks yang berbeda. Untuk mengatasi hal ini, dalam penelitian berjudul "*Dreambooth*: *Fine tuning text-to-Image Diffusion Models for subject-driven generation*" dilakukan pendekatan baru untuk "personalization" pada model difusi *text-to-Image* diperkenalkan, di mana model yang telah dilatih disesuaikan menggunakan beberapa gambar subjek sebagai referensi. (Ruiz *dkk.*, 2023). Teknik ini memanfaatkan prior semantik dalam model dan menambahkan fungsi *class-specific prior preservation Loss* untuk menjaga fitur utama subjek sembari memungkinkan variasi kontekstualisasi, termasuk rekonfigurasi subjek, sintesis tampilan berdasarkan teks, dan rendering artistik.

Dalam konteks pengenalan gambar medis, kekurangan data pelatihan berkualitas menjadi tantangan terutama ketika model CNN membutuhkan jumlah data besar untuk mencapai akurasi yang diinginkan. Untuk mengatasi hal ini, algoritma *Dreambooth* dapat dioptimalkan untuk mempelajari fitur spesifik dari gambar MRI otak, memungkinkan generasi data yang lebih realistis dan bermanfaat bagi pelatihan model. Bahwa generasi sekitar 800 gambar tambahan mampu meningkatkan akurasi pengenalan CNN hingga 60%, sebuah peningkatan sebesar 3% dibandingkan dengan set data yang lebih kecil (Zhang, 2023).

Dalam penelitian lain yang berjudul "*A New Chinese Landscape Paintings Generation Model based on Stable Diffusion using Dreambooth*"

diperkenalkan metode gabungan antara *Model Stable Diffusion* (SDM) dan *Parameter-Efficient Fine-tuning* untuk menghasilkan lukisan lanskap Tiongkok. Kombinasi SDM dengan *Dreambooth* memberikan kiner ja yang superior, mengungguli model lainnya, termasuk SDM yang dilatih sebelumnya secara umum dan SDM dengan *fine-tuning* berbasis LoRA (Gu, Fang dan Deng, 2024). Meskipun demikian, pendekatan ini memerlukan sumber daya komputasi yang lebih besar, menjadi kekurangan untuk aplikasi lebih luas,

Kemajuan dalam model *text-to-Image*, seperti *Stable Diffusion*, telah memungkinkan sintesis gambar dari *prompt* teks, sementara teknik personalisasi seperti *Dreambooth* memungkinkan model di-*fine-tune* untuk mengaitkan pengenal teks unik dengan beberapa gambar subjek tertentu (Park, Ko dan Jang, 2023). Meskipun teknik ini telah menunjukkan keberhasilan dalam menghasilkan gambar sesuai gaya tertentu, ada tantangan dalam mempelajari gaya seni yang abstrak dan luas—meliputi garis, bentuk, tekstur, dan warna.

Sebuah penelitian yang memanfaatkan model *Stable Diffusion* dan *Dreambooth* dihasilkan gambar bergaya 'Facebook Alegria' dan ikon siluet hitam untuk kebutuhan *user interface*. Hasil dari implementasi ini menunjukkan pengurangan besar dalam jam kerja yang dihabiskan untuk pengembangan frontend, hingga 81.65%, serta penurunan biaya sebesar 22.80% (Chávez dan Ticona, 2024). Hal ini menyoroti manfaat langsung dari penggunaan generator gambar berbasis teks dimana tim pengembang dapat mengurangi waktu dan biaya yang terkait dengan pembuatan aset visual.

Tantangan terkait generasi subjek yang dipersonalisasi terletak pada kesulitan menjaga keseimbangan antara mempelajari konsep subjek dengan mempertahankan keunggulan kemampuan generasi dari model yang telah dilatih sebelumnya. Untuk mengatasi kendala ini, penelitian berjudul "*DreamTuner: Single Image is Enough for Subject-Driven Generation*" melakukan pendekatan yang didasarkan pada *Dreambooth* bernama DreamTurner. Bertujuan untuk meningkatkan proses generasi gambar

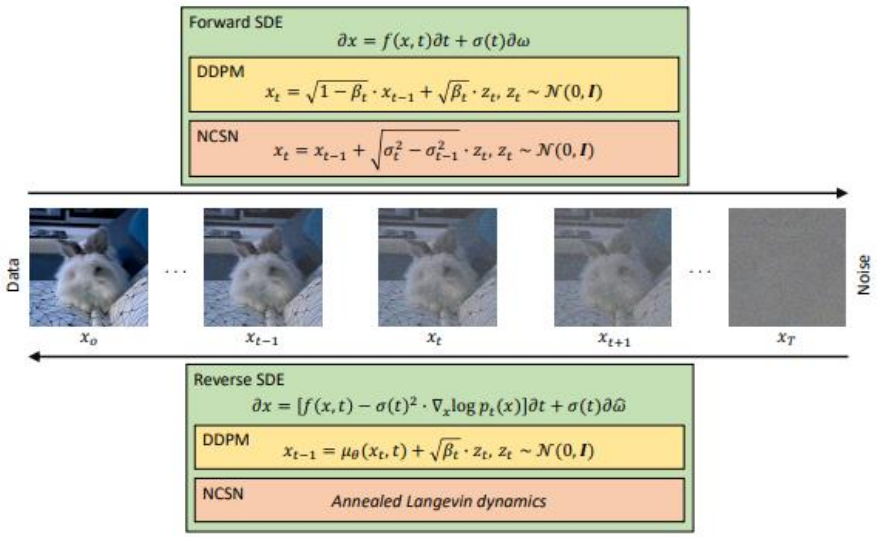
berbasis subjek secara efisien, tanpa mengorbankan detail visual penting dengan memanfaatkan pengenalan *subject-encoder* untuk menjaga identitas subjek secara kasar dan memperkenalkan fitur subjek terkompresi melalui lapisan atensi sebelum proses cross-attention visual-teks (Hua *dkk.*, 2023)

Dalam upaya mempertahankan identitas subjek banyak metode yang memerlukan optimasi yang memakan waktu atau melibatkan *encoder* tambahan. Untuk mengatasi batasan ini, penelitian berjudul "*DreamSalon: A Staged Diffusion Framework for Preserving Identity-Context in Editable Face Generation*" melakukan pendekatan yang didasarkan pada *Dreambooth* Bernama DreamSalon dengan kerangka kerja pengeditan berstaging yang dipandu oleh *noise* . Metode ini dirancang untuk menangani manipulasi detail dengan tetap menjaga konteks dan identitas asli subjek (Lin, 2024).

2.2 Dasar Teori

2.2.1 *Stable Diffusion*

Model generatif adalah model statistik yang mempelajari distribusi dari suatu data untuk menghasilkan sampel baru, dalam konteks pengolahan gambar, model generatif bertujuan untuk menciptakan gambar yang realistis berdasarkan *input* tertentu, seperti teks atau gambar. Sedangkan *model diffusion* adalah jenis model generatif yang bekerja dengan membalik proses penyebaran kebisingan secara bertahap pada data.



Gambar 2. 1 *Forward process* dan *Reverse*/Backward Process (F.-A. Croitoru *dkk.*, 2023)

Seperti yang ditunjukkan pada Gambar 2.1, *model diffusion* terdiri dari dua proses utama, yaitu *Forward SDE* dan *Reverse SDE*. Terdapat dua pendekatan utama dalam *model diffusion*, yaitu DDPM (*Denoising Diffusion Probabilistic Models*) dan NCSN (*Noise Conditional Score Networks*). DDPM lebih stabil dan probabilistik, tetapi membutuhkan banyak langkah *denoising* untuk merekonstruksi gambar dari *noise*. Di sisi lain, NCSN lebih fleksibel karena menggunakan *score-based modeling* yang membimbing proses rekonstruksi gambar secara iteratif dengan estimasi gradien *log-likelihood* (F. A. Croitoru *dkk.*, 2023).

Forward Stochastic Differential Equation (SDE) menggambarkan bagaimana data secara bertahap dicampur dengan *noise* hingga menjadi pure *noise* pada waktu T. Ini adalah tahap destruktif dalam *model diffusion*, yang bertujuan untuk membuat data asli X_0 tidak dapat dikenali secara bertahap.

$$\partial x = f(x,t)dt + \sigma(t)d\omega \quad (1)$$

x: Data atau gambar yang sedang mengalami transformasi (misalnya gambar asli yang sedang dicampur dengan *noise*).

t: Waktu dalam proses difusi, di mana $t = 0$ adalah gambar asli dan $t = T$ adalah *noise* murni.

f(x,t) adalah fungsi drift yang dapat mengubah data,

$\sigma\,(t)$ adalah skala *noise* yang berubah sesuai waktu t,

$d\omega$ adalah *Brownian motion*, yang melambangkan *noise* acak *Gaussian*.

Pada model *Denoising Diffusion Probabilistic Models* (DDPM) adalah model generatif berbasis proses difusi yang dikembangkan untuk menghasilkan gambar berkualitas tinggi (Huberman-Spiegelglas, Kulikov dan Michaeli, 2023). Model ini bekerja dengan menambahkan *noise* secara bertahap ke dalam data asli hingga menjadi distribusi *Gaussian* murni, kemudian melatih model untuk membalik proses tersebut guna merekonstruksi kembali data asli dari *noise* tersebut.. Pada DDPM, *forward process* diformulasikan sebagai berikut:

$$x_t = \sqrt{1-\beta_t}\,.\,x_{t-1} + \sqrt{\beta_t}\,.\,z_t\,,\quad z_t \sim N\,(0,1) \tag{2}$$

β_t adalah parameter yang mengontrol seberapa besar *noise* yang ditambahkan di tiap langkah t,

x_t adalah hasil dari kombinasi linier antara x_{t-1} dan *noise Gaussian* z_t

x_{t-1} adalah data pada waktu sebelumnya sebelum *noise* ditambahkan.

z_t adalah *Noise Gaussian* acak dengan distribusi normal $N\,(0,I)$ yang bertanggung jawab untuk membuat gambar semakin tidak dapat dikenali.

Noise Conditional Score Network (NCSN) adalah model generatif berbasis *score-based generative modeling*, yang bertujuan untuk mempelajari gradien logaritma dari distribusi data (dikenal sebagai *score function*), alih-alih langsung memodelkan distribusinya (Jung *dkk.*, 2024). Sementara itu, dalam NCSN, *forward process* dituliskan sebagai:

$$x_t = x_{t-1} + \sqrt{\sigma_t^2 - \sigma_{t-1}^2}\,.\,z_t,\quad z_t \sim N\,(0,1) \tag{3}$$

σ_t^2 adalah skala *noise* pada waktu, yang meningkat seiring waktu

σ_{t-1}^2 adalah skala *noise* pada waktu sebelumnya.

Berbeda dari DDPM, model ini tidak menggunakan parameter β_t, tetapi memakai skala *noise* σ_t yang bersifat kontinu, Proses ini lebih fleksibel karena memungkinkan penyesuaian *noise* dengan distribusi yang lebih luas.

Setelah gambar terdegradasi menjadi pure *noise* pada tahap *forward*, tahap *reverse* berusaha mengembalikan gambar asli dengan membalik proses penyebaran *noise*. Secara umum, *reverse* SDE dirumuskan sebagai:

$$\partial x = \left|\dot{f}(x,t) - \sigma(t)^2 \nabla_x log p_t(x)\right| dt + \sigma(t)\, d\bar{\omega} \quad (4)$$

$\nabla_x log p_t(x)$ adalah *score function*, yang merupakan gradien dari *log-likelihood* data pada waktu

$d\bar{\omega}$ adalah *Brownian motion* pada proses *denoising*, yang dapat mengoreksi *noise* agar kembali ke bentuk aslinya.

Pada DDPM, proses rekonstruksi gambar dituliskan sebagai:

$$x_{t-1} = \mu_0(x_t, t) + \sqrt{\beta_t}\, .z_t,\ z_t \sim N\,(0,1) \quad (5)$$

$\mu_0(x_t, t)$ adalah prediksi distribusi data sebelumnya berdasarkan parameter,

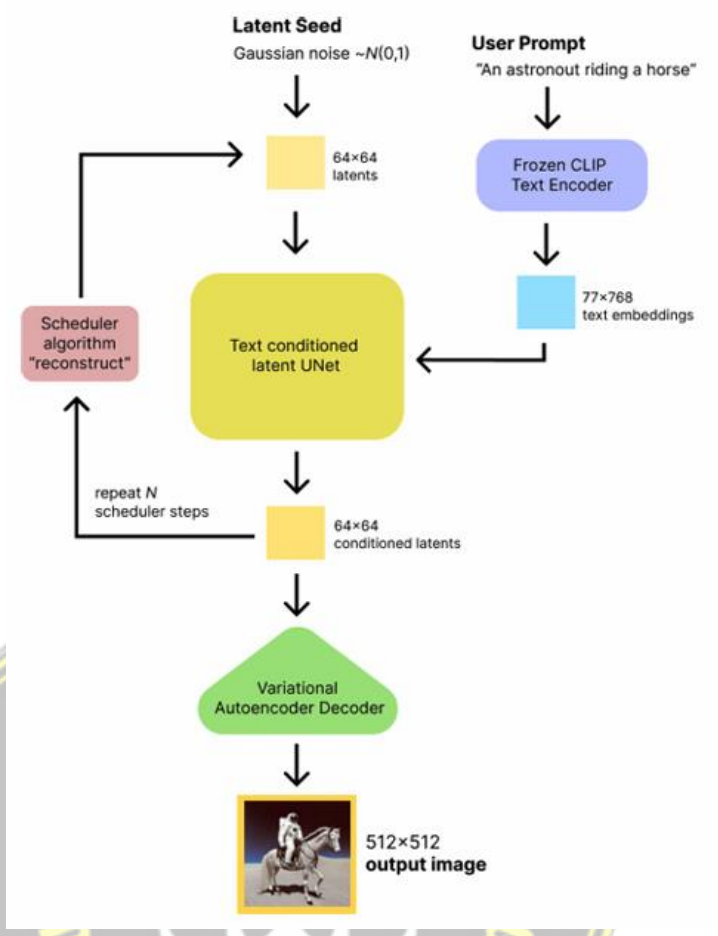
$\sqrt{\beta_t}$ adalah faktor *noise* yang dikurangi dalam langkah *denoising*.

Pada NCSN, *reverse* process menggunakan pendekatan *Annealed Langevin Dynamics*, yang dituliskan sebagai:

$$x_{t-1} = x_t + \epsilon \nabla_x log p_t(x) \quad (6)$$

ϵ adalah *learning rate* yang mengontrol seberapa besar perubahan yang dilakukan pada setiap iterasi *denoising*.

Stable Diffusion menggabungkan prinsip dari keduanya, tetapi bekerja di ruang laten untuk meningkatkan efisiensi sampling (Anderson dan Akram, 2024). Model ini bekerja dengan cara mengubah data berisik (*noise*) secara bertahap menjadi data yang lebih terstruktur.



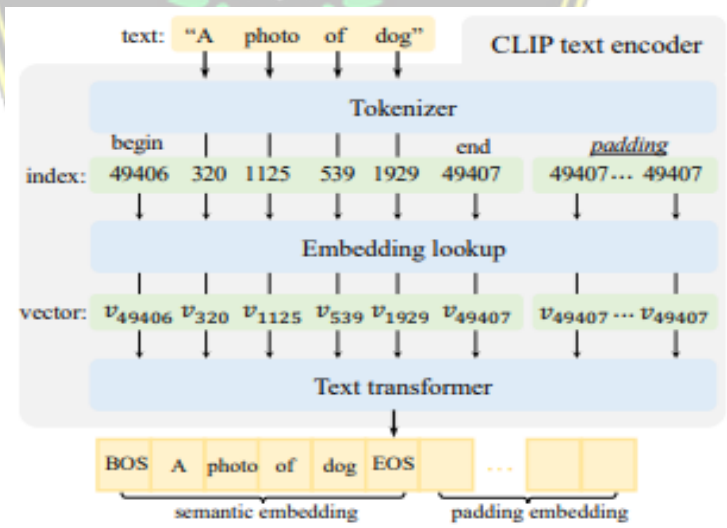
Gambar 2. 2 Ilustrasi Cara Kerja *Stable Diffusion*

Pada gambar 2.2, proses *Stable Diffusion* dimulai dengan pengguna memberikan *prompt* berupa teks, misalnya "*An astronaut riding a horse*", yang kemudian diproses menggunakan *Frozen CLIP Text Encoder* untuk mengubahnya menjadi vektor *embedding* berukuran 77×768 sebagai panduan dalam merekonstruksi gambar dari *noise* awal. Sementara itu, model menginisialisasi latent seed berupa *Gaussian noise* yang mengikuti distribusi dalam ruang laten berukuran 64×64. *Noise* laten ini kemudian diproses melalui *Text-Conditioned Latent U-Net*, yang bertugas menghilangkan *noise* secara bertahap menggunakan informasi dari *embedding teks*. Setelah melalui beberapa iterasi *denoising*, model menghasilkan 64×64 *conditioned latents*, yang kemudian diterjemahkan menggunakan *Variational Autoencoder (VAE) Decoder*, hingga akhirnya menghasilkan *output* gambar.

Stable Diffusion v1.5 dipilih dalam penelitian ini karena lebih efisien dalam hal penggunaan memori dan waktu inferensi dibandingkan dengan versi 2.1 dan SDXL. Meskipun versi 2.1 dan SDXL menawarkan beberapa perbaikan kualitas gambar, v1.5 tetap menjadi pilihan yang optimal untuk eksperimen yang membutuhkan efisiensi sumber daya. Penggunaan memori yang lebih rendah dan waktu inferensi yang lebih cepat pada v1.5 memungkinkan pemrosesan gambar yang lebih efisien tanpa mengorbankan kualitas visual yang signifikan, menjadikannya lebih praktis untuk aplikasi yang membutuhkan banyak iterasi atau pengolahan data dalam skala besar (Krojer *dkk.*, 2023).

2.2.2 *Text to Image Generation*

Text-to-Image generation adalah salah satu cabang dalam *artificial intelligence* yang bertujuan untuk menerjemahkan deskripsi teks menjadi gambar yang akurat dan sesuai. Teknik ini melibatkan beberapa komponen penting, seperti *embedding teks* ke dalam representasi vektor yang dapat dipahami oleh model, serta penggunaan teknik generatif seperti GAN (*Generative Adversarial Networks*) atau *Diffusion Models* untuk menghasilkan gambar dari representasi tersebut (Tao *dkk.*, 2022).



Gambar 2. 3 Text *Encoder* pada CLIP (Yu *dkk.*, 2024)

Untuk mengevaluasi kualitas gambar yang dihasilkan terhadap teks *input*, digunakan CLIP Score, yang dihitung menggunakan prinsip *cosine similarity* antara *embedding teks* dan *embedding* gambar dari model yang telah di*finetune*,seperti yang ditunjukkan pada gambar 2.3. Secara umum Clip Score dihitung dengan rumus :

$$S = \frac{E_t \cdot E_i}{\|E_t\| \cdot \|E_i\|} \quad (7)$$

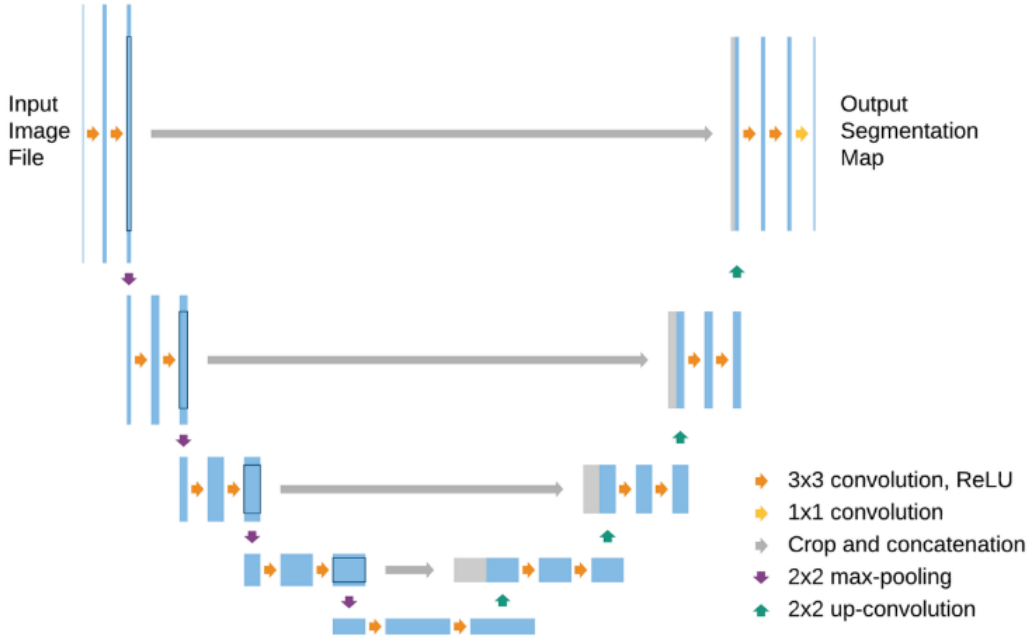
S adalah Clip Score, nilai kesesuaian antara teks dan gambar,

E_t adalah vector *embedding* dari teks (*prompt input*)

E_i adalah vector *embedding* dari gambar hasil dari model yang sudah di *finetune*

CLIP Score menjadi metrik evaluasi yang penting dalam *Text-to-Image Generation*, karena dapat mengukur sejauh mana model memahami dan menerjemahkan deskripsi teks ke dalam elemen visual yang relevan. Semakin tinggi nilai CLIP Score, semakin baik kesesuaian antara teks dan gambar yang dihasilkan, sehingga model dapat lebih diandalkan dalam menghasilkan gambar yang sesuai dengan berbagai variasi deskripsi teks (Li *dkk.*, 2023).

2.2.3 U-Net



Gambar 2. 4 Arsitektur U-Net (Weng dan Zhu, 2021)

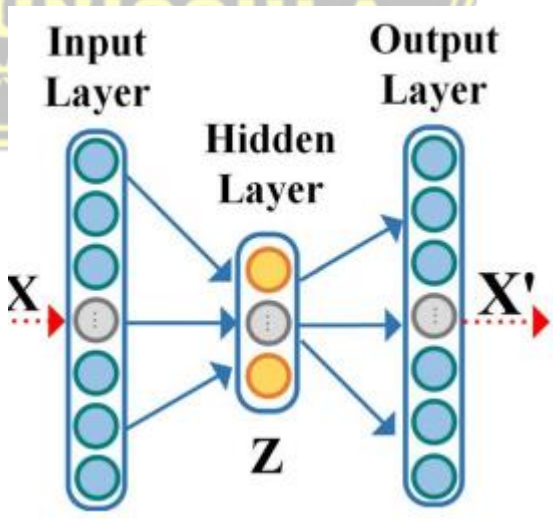
U-Net adalah arsitektur jaringan yang secara khusus dirancang untuk segmentasi citra, proses pemisahan objek atau klasifikasi setiap piksel dalam citra ke dalam beberapa kelas seperti yang ditujunkkan pada Gambar 2. 4. Teknik ini didasarkan pada *Convolutional Neural Network* (CNN), dengan berfokus pada ekstraksi fitur spasial dari gambar dengan menggunakan lapisan-lapisan konvolusi. CNN biasanya efektif dalam mendeteksi fitur penting dalam citra, seperti tepi atau tekstur, melalui penggunaan kernel yang bergerak di atas gambar *input* (Siddique *dkk.*, 2021). U-Net menggunakan struktur simetris berbentuk "U" yang terdiri dari "*contracting path"* dan "*expanding path*".

- *Contracting path* bertujuan untuk mengecilkan ukuran citra sambil menangkap fitur.
- *Expanding path* memperluas citra kembali ke ukuran asli sambil menggabungkan informasi fitur dari lapisan-lapisan sebelumnya.

Pendekatan ini memungkinkan U-Net untuk mempertahankan resolusi tinggi pada hasil segmentasi untuk mendeteksi detail kecil dan mempertahankan informasi spasial. Dalam praktiknya, U-Net memiliki kemampuan untuk menyesuaikan segmentasi terhadap struktur yang kompleks dan berbeda-beda.

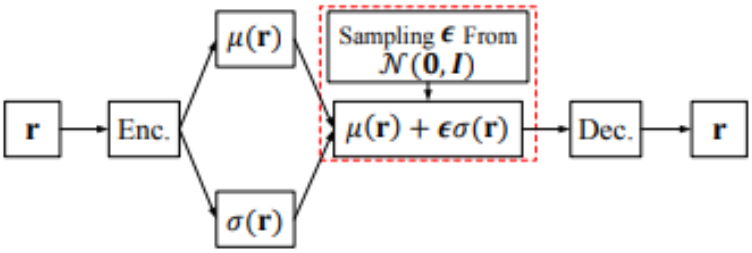
2.2.4 *Variational Autoencoders* (VAE)

Autoencoders (AEs) adalah jenis jaringan saraf yang dirancang untuk melakukan kompresi data, seperti gambar, dengan cara mengonversi *input* menjadi representasi laten yang lebih sederhana dan terkompresi, kemudian mendekodekannya kembali untuk menghasilkan *output* yang sedekat mungkin dengan *input* asli (Chen dan Guo, 2023). Proses ini dimulai dengan *encoder*, yang berfungsi untuk mengubah *input* menjadi vektor laten berdimensi lebih rendah menggunakan lapisan konvolusi dan *pooling*, seperti yang ditunjukkan pada gamabr 2. 5. Setelah itu, *decoder* berfungsi untuk melakukan operasi seperti dekonvolusi dan upsampling guna merekonstruksi *input* dari vektor laten yang telah dikompresi. Keduanya dikendalikan oleh fungsi kehilangan *Mean Squared Error* (MSE), yang mengukur perbedaan antara *input* asli dan hasil rekonstruksi pada tingkat piksel.



Gambar 2. 5 Arsitektur Auto-*encoder* (Berahmand *dkk.*, 2024)

Sementara itu, *Variational Autoencoders* (VAE) mengembangkan konsep dasar *autoencoder* dengan memperkenalkan aspek probabilistik pada representasi laten. VAE adalah model generatif yang berusaha memahami bagaimana data terbentuk melalui distribusi probabilistik. Dalam prosesnya, *encoder* menghasilkan dua parameter—*mean* dan deviasi standar—untuk setiap *input*, yang mewakili distribusi probabilistik dari ruang laten. Kemudian, vektor laten diambil sampelnya dari distribusi ini dan dikirimkan ke *decoder* untuk merekonstruksi *input* asalnya.



Gambar 2. 6 Implementasi VAE dengan *Feedforward Neural Network*

Seperti yang ditunjukkan pada gambar 2. 6 VAE beroperasi di bawah prinsip *Variational Bayes Inference*, yang mengasumsikan bahwa variabel laten yang tidak teramati dapat diestimasi melalui distribusi posterior yang dihitung dari *input* yang diamati (Mak, Han dan Yin, 2023).

$$z = \mu(r) + \epsilon \cdot \sigma(r), \quad \epsilon \sim N(0, I) \tag{7}$$

$\mu(r)$ adalah *mean* (rata-rata) dari distribusi *Gaussian* yang diprediksi oleh *encoder* untuk data masukan *r*.

$\sigma(r)$ adalah *standard deviation* (simpangan baku) atau scale dari distribusi *Gaussian* tersebut.

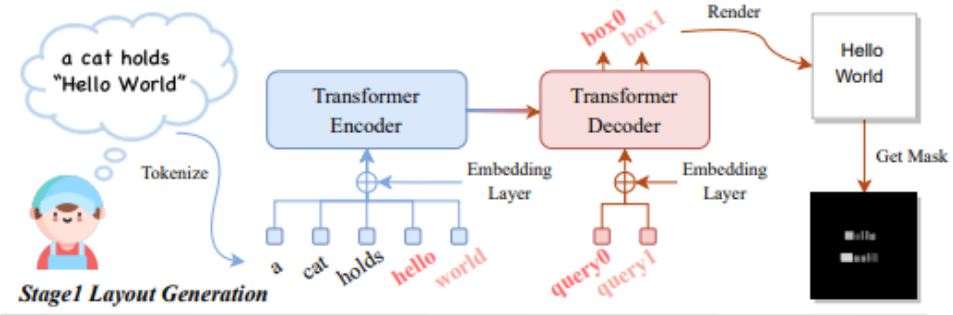
ϵ adalah variabel acak yang diambil dari distribusi Normal

Selama pelatihan, VAE bukan hanya dilatih untuk merekonstruksi gambar, tetapi juga untuk menghasilkan vektor laten yang terdistribusi secara normal. *Loss Function* yang digunakan dalam VAE tidak hanya mencakup rekonstruksi error, tetapi juga KL divergence, yang mengukur seberapa besar

perbedaan antara distribusi ruang laten yang dipelajari dan distribusi *Gaussian* standar. Dengan demikian, VAE tidak hanya berfokus pada kemampuan untuk merekonstruksi gambar, tetapi juga pada kemampuan untuk menghasilkan representasi laten yang terdistribusi dengan baik dari data *input*.

2.2.5 *Transformer*

Transformers merupakan arsitektur inti yang digunakan dalam model-model difusi, termasuk model dengan *kapabilitas open-vocabulary* seperti *Stable Diffusion*. Dalam generasi gambar berbasis teks (*Text-to-Image Generation*), transformers memainkan peran penting dalam memahami representasi teks *input* dan mentransformasikannya menjadi kondisi awal untuk proses difusi (Bolya dan Hoffman, 2023). Proses ini melibatkan dekoding teks menjadi representasi vektor, yang kemudian digunakan untuk memandu langkah-langkah iteratif dalam menghasilkan gambar, seperti yang ditunjukkan pada gambar 2. 7.



Gambar 2. 7 Transformer dengan *Encoder* dan *Decoder* (Chen *dkk.*, 2024)

Keunggulan transformer dalam menangani data sekuensial memungkinkan model untuk:

- Memahami Modalitas Berbeda: Misalnya, teks sebagai data sekuensial linguistik dan gambar sebagai representasi spasial.
- Menyelaraskan Modalitas: Transformer memadukan informasi lintas modalitas, seperti memetakan teks deskriptif ke elemen visual spesifik dalam gambar.

- Fleksibilitas Waktu Perturbasi: Setiap modalitas dapat memiliki tingkat perturbasi (*timesteps*) yang berbeda, dan transformer beradaptasi untuk memproses masukan dari berbagai tingkatan ini.

  (Bao *dkk.*, 2023)

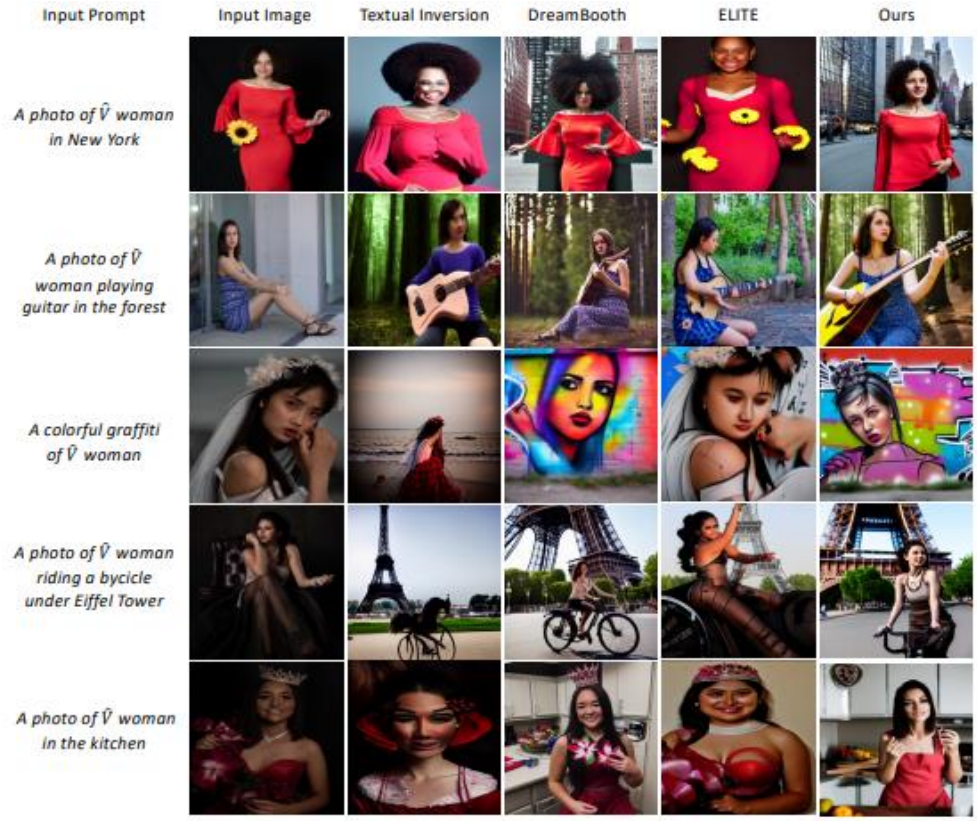
2.2.6 *Dreambooth Fine Tuning*

Fine-tuning adalah proses penyesuaian model yang telah dilatih sebelumnya agar lebih cocok untuk tugas spesifik dengan menggunakan *Dataset* yang lebih relevan. . Salah satu metode yang digunakan adalah *fine-tuning*, di mana model generatif diadaptasi secara khusus menggunakan set data kecil yang terkait erat dengan subjek tertentu (Ruiz *dkk.*, 2023). Misalnya, dalam model *Dreambooth, model diffusion fine-tuned* dengan data gambar dari subjek tertentu, memungkinkan pembuatan gambar karakter spesifik dari deskripsi teks dengan akurasi tinggi seperti yang ditunjukkan pada Gambar 2. 8.

Gambar 2. 8 Gambar *Input* dengan Berbagai *Promt* (Ruiz *dkk.*, 2023)

Dreambooth adalah teknik *fine-tuning* yang dikembangkan oleh *Google Research* untuk menghasilkan gambar-gambar yang spesifik menggunakan model berbasis difusi, seperti *Stable Diffusion*. Pengguna dapat membuat model yang menghasilkan gambar berdasarkan karakter, objek, atau gaya tertentu yang dipersonalisasi, dengan melatih model agar mampu "mempelajari" atribut visual unik dari beberapa contoh gambar *input*. Ini menjadikan model dapat menghasilkan gambar baru yang memiliki kemiripan atau fitur khusus dari objek atau karakter tersebut, yang sangat

berguna untuk aplikasi seperti avatar, karakter virtual, atau bahkan pembuatan konsep desain (Park, Ko dan Jang, 2023).



Gambar 2. 9 Perbandingan Bebeberapa Metode *Fine-tune* (Shi, 2024)

Seperti yang ditunjukkan oleh Gambar 2. 9 proses pelatihan dalam *Dreambooth* dimulai dengan mengumpulkan gambar-gambar dari objek yang diinginkan. Model dasar, seperti *Stable Diffusion*, kemudian di-*fine-tune* dengan gambar-gambar ini, untuk menjaga agar model tetap mempertahankan kemampuan umum yang ada dalam model awalnya sambil belajar fitur spesifik dari objek yang ingin direpresentasikan. *Dreambooth* mempertahankan kemampuan untuk menghasilkan gambar dengan resolusi tinggi dan detail yang tajam, namun tetap menjaga konteks visual yang mungkin berbeda dari gambar asli. Keunggulan *Dreambooth* terletak pada fleksibilitas dan kemampuannya untuk mempelajari detail yang presisi dari

data pelatihan terbatas, memungkinkan kreasi visual yang tak terbatas dan unik untuk setiap pengguna.

Dreambooth lebih cocok untuk personalisasi objek atau karakter baru, sementara LoRA lebih sering digunakan untuk modifikasi gaya atau perubahan kecil dalam model. LoRA lebih cocok untuk skenario multi-adaptasi, di mana ingin melatih beberapa konsep secara modular, tetapi jika hanya satu konsep spesifik yang ingin dipelajari secara mendalam, *Dreambooth* memberikan hasil yang lebih (Luo, 2023).

2.2.7 Regularisasi *Drop-out*

Regularisasi adalah teknik yang digunakan untuk mencegah *overfitting* yang sering terjadi pada model *deep learning*, yaitu kondisi di mana model mempelajari data pelatihan terlalu baik hingga kehilangan kemampuan untuk melakukan generalisasi pada data baru (Alkhairi *dkk.*, 2024). Model terlalu menyesuaikan dengan data pelatihan sehingga kehilangan kemampuan untuk melakukan generalisasi pada data uji. *Overfitting* dapat menyebabkan penurunan performa model terhadap data yang belum pernah dilatih sebelumnya. Salah satu metode regularisasi yang umum digunakan adalah *Drop-out*.

Dropout adalah teknik regularisasi yang menonaktifkan sejumlah unit (*neuron*) secara acak selama pelatihan, sehingga model tidak terlalu bergantung pada fitur spesifik yang mungkin hanya relevan pada data pelatihan (Murtopo *dkk.*, 2024). Dalam *Stable Diffusion*, *Dropout* dapat membantu mengurangi ketergantungan model pada fitur-fitur spesifik dalam *dataset*, sehingga menghasilkan *output* yang lebih beragam dan mampu menangkap konteks yang lebih luas dari *prompt*.

2.2.8 Parameter

Dalam pengembangan sistem *Text-to-Image Generation* berbasis *Diffusion Models*, melibatkan serangkaian elemen teknis termasuk dalam pengaturan *hyperparameter* untuk menghasilkan gambar yang sesuai dengan

deskripsi teks yang diberikan. Berikut parameter yang digunakan dalam penelitian ini :

- *Dropout* Rate : Digunakan untuk mengurangi *overfitting* dengan secara acak menonaktifkan sejumlah unit dalam jaringan selama pelatihan, yang memaksa model untuk belajar representasi yang lebih general (D'Angelo *dkk.*, 2023).

  $\hat{y} = W \,.\, (x \odot m) + b$ (8)

  $\hat{y}$: *Output* prediksi model setelah penerapan *Dropout*.

  W : Matriks bobot dari *layer* model.

  x : *Input* data.

  m : Mask *Dropout* (nilai biner: 1 untuk mempertahankan, 0 untuk drop).

  $\odot$: Operasi elemen-*wise multiplication* (perkalian elemen satu per satu).

  b : Bias yang ditambahkan.

- *Loss* Function: Fungsi *loss* berperan sebagai indikator seberapa baik model melakukan tugasnya. Dalam konteks *Diffusion Models*, fungsi *loss* seperti *Mean Squared Error* (MSE) sering digunakan untuk menghitung perbedaan antara gambar yang dihasilkan dengan gambar referensi, memberikan umpan balik untuk memperbaiki model.

  $\mathcal{L}_{total} = \mathcal{L}_{content} + \lambda \mathcal{L}_{regularization}$ (9)

  $\mathcal{L}_{total}$: Total *loss* atau kerugian yang dihitung selama pelatihan.

  $\mathcal{L}_{content}$: *Loss* untuk menjaga kesesuaian konten atau kualitas generasi.

  $\mathcal{L}_{regularization}$: Regularisasi untuk mencegah *overfitting* .

  λ : Koefisien regulasi untuk mengontrol kontribusi dari regularisasi.

- *Noise* Schedule: *Noise* schedule digunakan untuk mengontrol pengurangan *noise* bertahap selama proses *denoising*. Penjadwalan *noise*

yang tepat memungkinkan model untuk mengubah gambar yang penuh dengan *noise* menjadi gambar yang jelas dengan detail yang tepat.

2.2.9 *Art Nouveau*

Art Nouveau adalah aliran seni dan desain yang berkembang pada akhir abad ke-19 hingga awal abad ke-20 (sekitar 1890 hingga 1910). Gaya ini dikenal dengan bentuk-bentuk melengkung yang elegan, garis-garis yang organik dan alami, serta penggunaan motif dekoratif yang terinspirasi oleh alam, seperti bunga, daun, dan bentuk hewan (Radityasari *dkk.*, 2023). *Art Nouveau* muncul sebagai reaksi terhadap desain industri yang lebih kaku dan formal dari abad ke-19 dan merupakan upaya untuk mengembalikan unsur seni dan keindahan dalam kehidupan sehari-hari. Selain itu, penggunaan warna hangat yang terdesaturasi menjadi salah satu ciri khas gaya ini, yang dirancang untuk memberikan nuansa fantasi dan menciptakan suasana yang melankolis serta penuh imajinasi (Kushwaha dan Srivastava, 2021). Elemen-elemen ini menjadikan *Art Nouveau* sangat cocok digunakan untuk menggambarkan karakter-karakter fantasi, karena dapat menciptakan nuansa yang memadukan keindahan alam dan imajinasi kreatif.

(a) (b) (c)

Gambar 2. 10 (a) Zodiac oleh Alphonse Mucha, (b) Daydream oleh Alphonse Mucha, dan (c) Princess Hyacinth oleh Alphonse Mucha

- Garis melengkung dan organik: Salah satu ciri utama *Art Nouveau* adalah penggunaan garis melengkung yang mengalir dan alami, yang terinspirasi

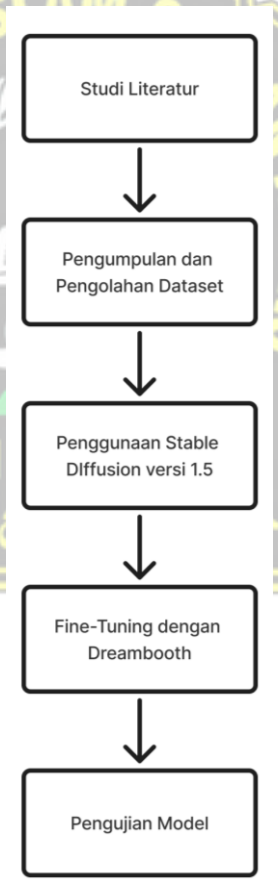
dari bentuk-bentuk alami seperti tumbuhan, bunga, dan aliran air (Saraswati, Utami dan Pemayun, 2024).

- Motif Alam: Penggunaan elemen-elemen dari alam sangat dominan, termasuk flora dan fauna. Desain ini sering mencerminkan keseimbangan dan harmoni alam.
- Dekorasi yang rumit dan detail: Gaya ini menekankan pada keindahan ornamen dan detail yang dibuat dengan presisi, menggabungkan estetika dengan fungsionalitas.
- Pengaruh dari simbolisme: *Art Nouveau* juga dipengaruhi oleh gerakan simbolisme dalam seni, yang lebih fokus pada ekspresi perasaan dan imajinasi daripada realitas objektif.

BAB III
METODE PENELITIAN

3.1 Metode Penelitian

Pada tahap pelatihan, penulis akan membuat generator gambar dengan menggunakan *Stable Diffusion* v1.5 dan menerapkan *fine-tuning* melalui metode *Dreambooth*. Kemudian, pada tahap pengembangan aplikasi berbasis *web*, penulis akan menggunakan *Streamlit* untuk membangun antarmuka yang memungkinkan pengguna memasukkan deskripsi karakter secara langsung dan mendapatkan hasil visualisasi gambar sesuai dengan parameter yang telah ditetapkan.



Gambar 3. 1 Tahapan Penelitian

Keluaran dari tugas akhir ini adalah sistem yang dapat menghasilkan gambar karakter dari novel fantasi berdasarkan deskripsi teks yang diberikan

oleh pihak penulis atau penerbit. Dengan menggunakan pendekatan Diffusion model, mana memanfaatkan proses transformasi iteratif dari *noise* acak menjadi gambar yang lebih realistis berdasarkan pola pelatihan model (*training*).

3.1.1 Studi Literatur

Meninjau dan menganalisis sumber-sumber yang relevan (jurnal atau artikel) dengan topik seputar *Stable Diffusion*, *Fine-tuning*, *Dreambooth*, dan *Regularization Drop-out.*

3.1.2 Pengumpulan dan Pengolahan *Dataset*

Pada tahap pengumpulan *dataset*, langkah-langkah yang dilakukan adalah sebagai berikut:

- Identifikasi Sumber Data : Menggunakan Pinterest untuk memperoleh *Dataset* yang mencakup gambar-gambar karakter fantasi.
- Seleksi Gambar : Memilih gambar yang sesuai dengan gaya *Art Nouveau*, dengan memperhatikan ekspresi wajah, gaya, dan atribut visual

Setelah *dataset* terkumpul, dilakukan proses pengolahan dengan langkah-langkah berikut :

- Pengelompokan Gambar : Mengelompokkan gambar berdasarkan pose, gaya, dan atribut visual yang dimiliki karakter untuk memudahkan proses penyesuaian.
- Validasi Keberagaman dan Penyesuaian Data : Memastikan *Dataset* mencakup variasi pose, atribut, dan gaya visual untuk membantu model menghasilkan gambar.
- Penyimpanan *Dataset* : *Dataset* yang mencakup gambar-gambar karakter fantasi dengan gaya tema *Art Nouveau* dalam berbagai pose, gaya, dan atribut visual akan dikelompokkan ke menjadi dua (*Data Training* dan Data Uji). Data pelatihan harus mencakup beragam contoh karakter anime, baik dari segi pose maupun atribut visual, agar model dapat mempelajari variasi karakteristik gaya anime secara lebih mendalam.

3.1.3 Penggunaan *Stable Diffusion* 1.5

Stable Diffusion 1.5 dipilih dalam penelitian ini karena kemampuannya untuk menghasilkan gambar berkualitas. Namun, meskipun memiliki banyak keunggulan, model ini tetap menghadapi tantangan dalam dua aspek utama:

- Isu Fidelitas

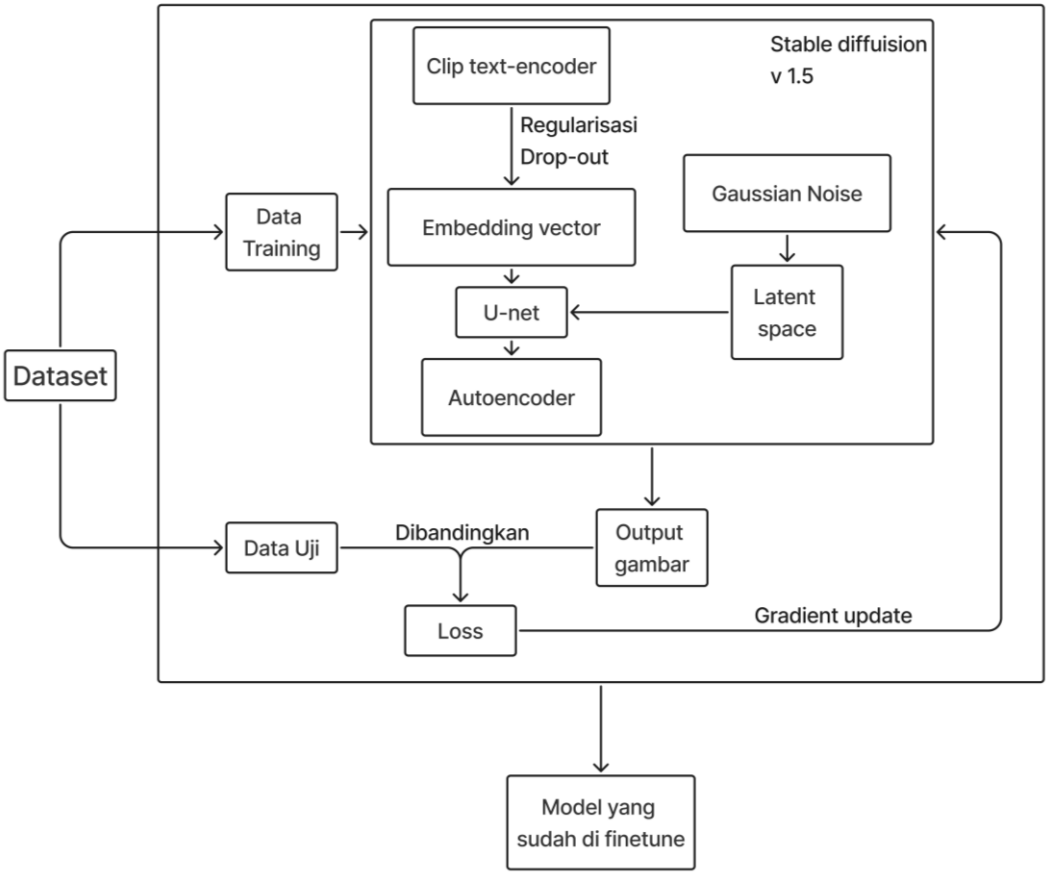
  Misalnya, saat menggambarkan karakter dengan atribut visual yang unik, seperti pakaian rumit, ekspresi wajah tertentu, atau pola gaya tertentu (contohnya *Art Nouveau*), hasil generasi model ini bisa menjadi terlalu generik atau tidak sepenuhnya konsisten dengan *prompt*. Masalah ini terutama terjadi karena model hanya memiliki pelatihan generik pada *dataset* besar tanpa adaptasi spesifik untuk elemen-elemen unik.

- Isu Personalisasi

  Model generatif seperti *Stable Diffusion* 1.5 secara *default* dirancang untuk generalisasi, menghasilkan gambar yang mencakup berbagai gaya dan tema. Namun, ketika diminta untuk menghasilkan visualisasi subjek tertentu berdasarkan referensi atau deskripsi spesifik, model ini dapat mengalami kesulitan dalam menjaga konsistensi atribut unik dari subjek yang digambarkan. Ini menimbulkan tantangan dalam menghasilkan personalisasi karakter yang sesuai dengan harapan, terutama dalam konteks dunia fantasi.

Dengan memadukan *Stable Diffusion* 1.5 dan *Dreambooth*, penelitian ini berfokus pada solusi untuk meningkatkan fidelitas visual dan memastikan personalisasi karakter yang lebih konsisten. Pendekatan ini relevan untuk menghasilkan karakter fantasi bergaya *Art Nouveau*, yang membutuhkan perhatian terhadap detail estetika dan keunikan elemen visual.

3.1.4 Fine Tuning dengan *Dreambooth*



Gambar 3. 2 Alur Kerja *Training* Sistem

Berikut adalah alur kerja *training sample* seperti yang ditunjukkan pada Gambar 3. 2 :

1. Tahap *Preprocessing Text* dengan *Encoder*
   - *Text/Keyword Encoding* : Deskripsi teks yang merepresentasikan karakter diproses menggunakan CLIP *Text Encoder*, menghasilkan representasi vektor teks (*embedding* vector). *Embedding teks* ini digunakan sebagai *prompt*, yang menjadi acuan bagi model untuk memahami elemen visual karakter. *Library Transformers* juga

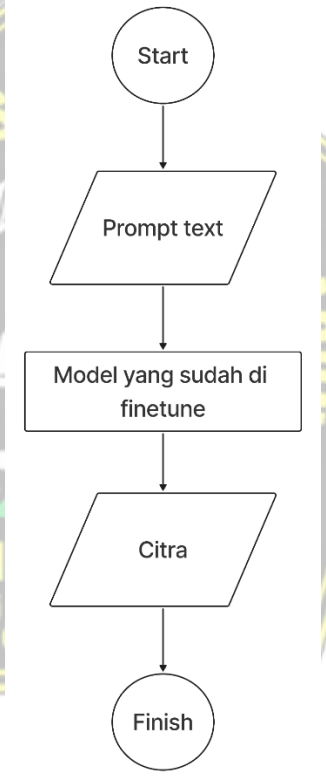
digunakan untuk memastikan kompatibilitas *embedding* dengan arsitektur *Stable Diffusion*.

2. Tahap Pelatihan dengan *Fine-tuning Dreambooth*
   - Pembelajaran Berbasis Gambar Referensi : Gambar-gambar referensi dari *dataset* digunakan untuk mengaitkan deskripsi teks dengan karakter visual tertentu.
   - *Forward Pass* : Model menerima text *embedding* dan gambar referensi untuk menghasilkan latent representation. Gambar awal ini mengandung *noise* yang secara bertahap akan dihilangkan selama pelatihan.
   - Perhitungan *Loss* : Menggunakan fungsi *loss*, seperti *mean squared error* (MSE), untuk menghitung kesalahan antara gambar yang dihasilkan dan gambar referensi.
   - *Backward Pass* : *Loss* yang dihitung digunakan untuk memperbarui bobot jaringan neural.
3. Tahapan di Dalam *Diffusion Models* (*Stable Diffusion* v1.5)
   - *Gaussian Noise* Addition : *Noise Gaussian* ditambahkan iteratif di latent space, memastikan gambar acak sepenuhnya sebelum di-de*noise*.
   - U-Net Architecture : Proses *denoising* terjadi di latent space, dengan U-Net bertugas menghapus *noise* secara bertahap sesuai informasi teks. Penerapan *Dropout* dan augmentation dalam U-Net meningkatkan generalisasi model.
   - Variational Auto*encoder Decoder* : Setelah proses *denoising* selesai, representasi laten diterjemahkan kembali menjadi gambar melalui *decoder* auto*encoder*, menghasilkan *output* gambar akhir.
4. Tahap Evaluasi dan Validasi Model
   - Validasi Gambar : *Dataset* validasi digunakan untuk menguji performa model dengan *prompt* yang berbeda.
   - Pengukuran Akurasi Visual : Hasil gambar dibandingkan dengan gambar referensi untuk memastikan konsistensi gaya dan kesesuaian deskripsi.

- Iterasi *Fine-tuning* : Jika hasil belum optimal, pelatihan dilanjutkan dengan menyesuaikan parameter seperti *learning rate* atau *dataset* tambahan.

3.2 Perancangan Alur Sistem

Perancangan alur sistem bertujuan untuk menggambarkan langkah-langkah utama dalam pengolahan deskripsi teks menjadi gambar karakter menggunakan model *Stable Diffusion* 1.5 yang telah di-*fine-tune* dengan *Dreambooth*. Alur sistem terdiri dari beberapa tahap utama :



Gambar 3. 3 Rancangan Alur Sistem

- *Input* Deskripsi Teks (*Prompt*)

  Pengguna memasukkan deskripsi teks yang menggambarkan atribut karakter, termasuk pakaian, ekspresi, pose, dan elemen gaya visual (seperti motif *Art Nouveau*).

- *Encoding* Teks

  Deskripsi teks diproses menggunakan CLIP Text *Encoder* untuk menghasilkan representasi vektor teks (text *embedding*). Vektor ini digunakan untuk memandu model generatif dalam menghasilkan gambar.

- Generasi Gambar
  a. Model pre-trained langsung digunakan untuk proses transformasi teks ke gambar.
  b. *Gaussian Noise* Addition: Proses dimulai dengan gambar acak (*noise*).
  c. *Denoising*: Model secara iteratif menghapus *noise* menggunakan *embedding teks* sebagai panduan untuk membentuk elemen visual.
  d. *Latent Space Transformation*: Representasi laten diterjemahkan kembali menjadi gambar resolusi tinggi menggunakan *variational* auto*encoder* (VAE).

- Evaluasi dan Validasi
  a. Keselarasan Visual: Gambar hasil dibandingkan dengan data *testing* untuk memastikan kesesuaian atribut visual.
  b. *Fidelity* dan Konsistensi: Analisis dilakukan untuk menilai sejauh mana gambar mencerminkan deskripsi teks secara detail.

3.3 Analisis Kebutuhan Sistem

Analisis kebutuhan sistem dilakukan untuk memastikan sistem yang dibangun dapat memenuhi tujuan penelitian, yaitu menghasilkan gambar karakter fantasi bergaya *Art Nouveau* berdasarkan deskripsi teks. Analisis ini mencakup kebutuhan perangkat keras, perangkat lunak, serta kebutuhan fungsional dan non-fungsional sistem.

- Perangkat keras
  a. Komputer atau Laptop:Spesifikasi minimum: RAM 8 GB, *prosesor dual-core* (contoh: Intel i5 atau AMD Ryzen 5). Sistem operasi apa pun (Windows, macOS, atau Linux) yang mendukung *browser web modern.*

b. Akses Internet Stabil: Dibutuhkan untuk mengakses Google Colab dan mengunggah/unduh data serta model.

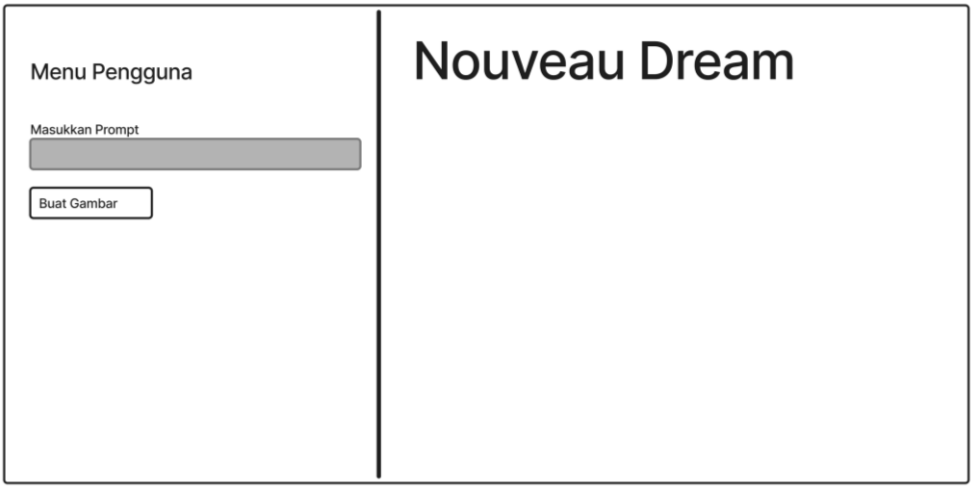
- Perangkat lunak
  a. Google Colab : Akses GPU atau TPU di Google Colab (misalnya, Tesla T4 atau K80) untuk mendukung pelatihan dan inferensi model.
  b. *Library*

Tabel 3. 1 Tabel *Library*

Library	**Deskripsi**	**Fungsi**
Pytorch	*Library deep learning* fleksibel untuk membangun, melatih, dan menguji model pembelajaran mendalam.	• Implementasi model *Stable Diffusion* 1.5. • Mendukung backpropagation untuk pelatihan model. • Memanfaatkan GPU untuk efisiensi.
Transformers	*Library* untuk pemrosesan teks dan *Encoding* representasi vektor.	• *Encoding* deskripsi teks menggunakan CLIP Text *Encoder*. • Integrasi dengan Diffusers untuk proses teks ke gambar.
Diffusers	*Library* khusus untuk *Diffusion Models*, termasuk *Stable Diffusion*.	• Pipeline untuk proses *denoising*. • Mendukung *fine-tuning Dreambooth*. • Akses ke pre-trained model *Stable Diffusion*.
Matplotlib	*Library* visualisasi data.	• Menampilkan grafik evaluasi *fidelity* dan personalisasi.

		• Visualisasi metrik pelatihan, seperti *loss* function.
NumPy	*Library* untuk komputasi numerik.	• Normalisasi data numerik. • Operasi pada array dan matriks untuk pemrosesan gambar dan analisis data.
Pandas	*Library* manipulasi data berbasis tabel.	• Pengorganisasian metadata *dataset* • Menyajikan hasil evaluasi fidelitas dalam tabel.

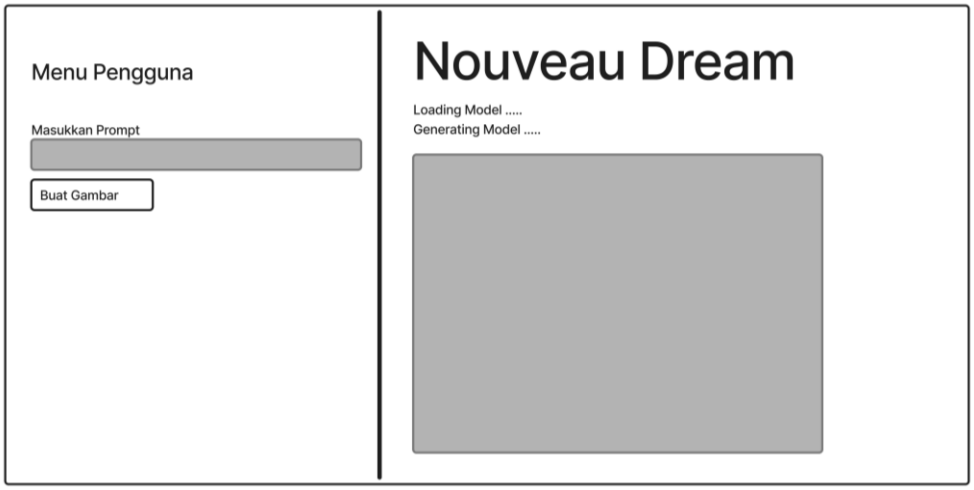
3.4 Perancangan User Interface



Gambar 3. 4 Tampilan Awal Sistem

Perancangan *User Interface* untuk *AI Image Generator* dirancang dengan pendekatan minimalis yang mengutamakan fungsionalitas utama dan

kemudahan penggunaan. Pada tampilan awal, halaman dimulai dengan judul yang terletak di bagian atas sebagai penanda utama. Di bawahnya, terdapat kolom *input* berbentuk persegi panjang dengan latar abu-abu muda, yang berfungsi sebagai tempat pengguna memasukkan *prompt* atau deskripsi gambar yang ingin dihasilkan. Untuk memberikan panduan, kolom ini dilengkapi dengan *placeholder* bertuliskan "*Please Enter your prompt here!"*. Tepat di bawah kolom *input*, terdapat tombol bertuliskan "*Generate Image*" dengan desain sederhana—latar putih dan teks hitam—yang mudah diidentifikasi dan diakses pengguna.



Gambar 3. 5 Tampilan Saat Generasi Gambar

Setelah tombol "*Generate Image*" ditekan, antarmuka menampilkan status proses seperti "*Loading Model..."* dan "*Generating Model..."* di bawah tombol untuk memberikan feedback kepada pengguna mengenai status generasi gambar. Di bagian bawah halaman, terdapat area besar berbentuk persegi panjang dengan latar abu-abu yang berfungsi sebagai tempat *output* gambar ditampilkan setelah proses selesai. Elemen ini dirancang secara sentral untuk memprioritaskan perhatian pengguna terhadap hasil yang dihasilkan.

BAB IV
HASIL DAN ANALISIS PENELITIAN

4.1 Inisialisasi Model dan Persiapan Sistem

4.1.1 Data Preprocessing

Dataset mencakup beberapa gambar karakter bergaya *Art Nouveau* yang diambil dari Pinterest. Gambar dalam *dataset* ini memiliki format .png, .jpg, dan .jpeg. Sebelum digunakan, *dataset* melalui proses penyesuaian nama file dan pemisahan data menjadi set pelatihan (*training set*) dan pengujian (*testing set*).

Gambar 4. 1 Contoh *Dataset*

Dataset yang digunakan terdiri dari 30 gambar karakter bergaya *Art Nouveau*, yang kemudian dibagi ke dalam *training* set dan testing set dengan skema 80:20. Dalam pembagian ini, sebanyak 24 gambar dimasukkan ke dalam *training* set untuk melatih model, sementara 6 gambar digunakan sebagai testing set untuk mengevaluasi performa model. Pembagian ini bertujuan untuk memastikan bahwa model dapat belajar pola dari mayoritas data, namun tetap memiliki data yang belum pernah dilihat sebelumnya untuk menguji kemampuannya dalam menghasilkan gambar yang sesuai dengan deskripsi *input*.

```
[5]  import shutil

     source_folder = "/content/drive/MyDrive/training/nou"
     destination_folder = "/content/resized_images"
     file_list = os.listdir(source_folder)
     file_list.sort()
     for index, file_name in enumerate(file_list):
         if file_name.lower().endswith(('.png', '.jpg', '.jpeg')):
             new_name = f"art{index + 1}.jpg"  # Format nama baru
             source_path = os.path.join(source_folder, file_name)
             destination_path = os.path.join(destination_folder, new_name)
             shutil.copy(source_path, destination_path)
             print(f"{file_name} -> {new_name}")
     print("Pengaturan dataset selesai!")
```

Gambar 4. 2 Kode untuk Menyesuaian Nama File

Gambar 4.1 menunjukkan proses otomatisasi dalam penyesuaian nama file *dataset* menggunakan kode Python. Setiap file di folder sumber diberi nama baru secara berurutan, seperti art1.jpg, art2.jpg, dan seterusnya, untuk menjaga keteraturan. Proses ini menggunakan pustaka shutil untuk memindahkan file ke folder tujuan yang telah ditentukan.

4.1.2 Evaluasi Kesamaan Gambar

Evaluasi kesamaan gambar dilakukan untuk menilai sejauh mana hasil keluaran model generasi gambar sesuai dengan *prompt* atau gambar referensi yang diberikan. Proses evaluasi ini berguna untuk mengukur performa model dalam menghasilkan gambar yang relevan dan sesuai dengan tujuan yang diinginkan.

```
[ ]  def get_image_features(image):
         img_tensor = preprocess(image).unsqueeze(0).to(device)
         with torch.no_grad():
             image_features = clip_model.encode_image(img_tensor)
         return image_features

     similarity_scores_gen_vs_ref = []
     reference_image_features = [get_image_features(ref_img) for ref_img in reference_images]

     for i, gen_img in enumerate(generated_images):
         gen_img_features = get_image_features(gen_img)
         ref_similarity_scores = [
             torch.nn.functional.cosine_similarity(gen_img_features, ref_features).item()
             for ref_features in reference_image_features
         ]
         avg_ref_similarity = sum(ref_similarity_scores) / len(ref_similarity_scores)
         similarity_scores_gen_vs_ref.append(avg_ref_similarity)

     print("\nPerbandingan Skor Kesamaan (Gambar Hasil Generasi vs Gambar Referensi):")
     for i, avg_ref_score in enumerate(similarity_scores_gen_vs_ref):
         print(f"Gambar {i+1} - Skor rata-rata kesamaan dengan gambar referensi: {avg_ref_score:.2f}")
```

Gambar 4. 3 Kode Untuk Perbandingan Skor Kesamaan (Gambar Hasil Generasi vs Gambar Referensi)

Evaluasi ini dilakukan untuk menilai kesesuaian antara gambar hasil generasi dengan gambar referensi yang digunakan sebagai target. Sama seperti pada evaluasi sebelumnya, model CLIP digunakan untuk mengekstraksi fitur dari gambar hasil generasi dan gambar referensi. Kedua gambar diubah menjadi vektor fitur menggunakan fungsi `clip_model.encode_image`, kemudian vektor fitur tersebut dinormalisasi untuk menjaga keadilan perbandingan. Skor kesamaan dihitung menggunakan metode dot product antara vektor fitur dari kedua gambar. Nilai skor ini mencerminkan seberapa dekat gambar hasil generasi dengan gambar referensi yang diharapkan. Evaluasi ini sangat penting terutama dalam kasus di mana hasil yang diinginkan harus menyerupai referensi visual tertentu.

4.1.3 Penggunaan *Stable Diffusion* 1.5

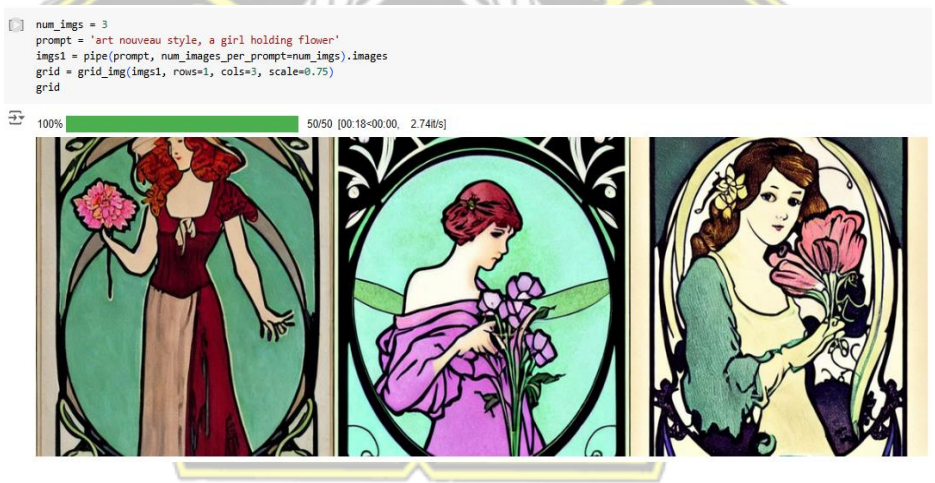
Model *Stable Diffusion* 1.5 dimuat menggunakan pipeline dari pustaka *diffusers*. Kode berikut menunjukkan bagaimana model diakses

menggunakan metode `from_pretrained`, di mana model dijalankan dengan tipe data `float16` untuk mengoptimalkan penggunaan memori.

```
[ ]  pipe = StableDiffusionPipeline.from_pretrained("runwayml/stable-diffusion-v1-5", torch_dtype=torch.float16)
```

Gambar 4. 4 Kode Untuk Memuatan Model *Stable Diffusion* 1.5

Model ini bertugas mengubah deskripsi teks (*prompt*) menjadi gambar dengan mempertimbangkan gaya atau karakteristik tertentu yang disebutkan dalam *prompt*. Setelah model berhasil dimuat, *prompt* berupa deskripsi teks digunakan untuk menghasilkan beberapa gambar. Contoh berikut menunjukkan penggunaan *prompt Art Nouveau style, a girl holding flower* untuk menghasilkan tiga gambar dengan gaya *Art Nouveau*.



Gambar 4. 5 Generasi Gambar Menggunakan *Prompt*

Penggunaan *prompt* yang sama digunakan secara berulang, model dapat menghasilkan variasi gambar dengan perbedaan pada komposisi, palet warna, atau detail tertentu, meskipun tetap mempertahankan elemen utama yang disebutkan dalam deskripsi. Variasi ini terjadi karena model bekerja dengan pendekatan stokastik, di mana setiap proses generasi bersifat probabilistik dan dapat menghasilkan interpretasi visual yang berbeda dalam setiap iterasi. Dengan melakukan pengujian menggunakan *prompt* yang sama secara

berulang, pengguna dapat menganalisis bagaimana model menyesuaikan elemen-elemen visual dalam berbagai hasil serta mengevaluasi konsistensi dan fleksibilitas model dalam menerapkan gaya yang diinginkan.

Kode ini menghasilkan tiga gambar dalam satu baris menggunakan fungsi grid_img. Gambar-gambar tersebut merepresentasikan hasil generasi model berdasarkan deskripsi teks.

```
CLIP Score Gambar 1: 0.365478515625
CLIP Score Gambar 2: 0.35205078125
CLIP Score Gambar 3: 0.369873046875
Rata-rata CLIP Score: 0.3624674479166667
```

Gambar 4. 6 Evaluasi Hasil dengan CLIP Score

Rata-rata CLIP Score untuk ketiga gambar adalah 0.3624674479166667, yang menunjukkan tingkat kesesuaian rata-rata antara *prompt* dan gambar hasil generasi. Evaluasi ini membantu memahami performa model dalam menghasilkan gambar yang relevan dengan deskripsi teks.

4.1.4 Pengaturan Parameter

```
[ ]  prompt = 'art nouveau style, a girl holding flower'
```

Gambar 4. 7 *Prompt* yang Digunakan

Dalam menghasilkan gambar dengan *prompt* "*Art Nouveau style, a girl holding flower*", terdapat beberapa parameter penting yang dapat diatur untuk memengaruhi hasil visual. Untuk setiap eksperimen, *prompt* yang sama digunakan secara konsisten agar perubahan efek dari setiap parameter dapat diamati dengan jelas.

a. Seed

```
seed = 777
generator = torch.Generator('cuda').manual_seed(seed)
img = pipe(prompt, generator=generator).images[0]
img
```

```
100%  50/50 [00:07<00:00,  7.30it/s]
```

Gambar 4. 8 Gambar dengan seed 777

Parameter seed digunakan untuk memastikan reprodusibilitas gambar. Dengan menggunakan nilai seed yang sama, generator akan menghasilkan gambar yang konsisten meskipun proses dilakukan di waktu berbeda.

b. *Inference steps*

```
[ ]  import matplotlib.pyplot as plt

     plt.figure(figsize=(18,8))
     for i in range(1, 6):
       n_steps = i * 10
       generator = torch.Generator('cuda').manual_seed(seed)
       imginference = pipe(prompt, num_inference_steps=n_steps, generator=generator).images[0]
       plt.subplot(1, 5, i)
       plt.title('num_inference_steps: {}'.format(n_steps))
       plt.imshow(img)
       plt.axis('off')
     plt.show()
```

Gambar 4. 9 Nilai inference steps (10, 20, 30, 40, 50)

Gambar 4. 10 Perubahan detail pada hasil.

Inference steps menentukan jumlah langkah dalam proses menghasilkan gambar. Semakin tinggi nilainya, hasil gambar biasanya lebih detail, tetapi memerlukan waktu lebih lama. Contoh pengaturan langkah adalah 10, 20, 30, 40, dan 50.

c. *Guidance scale* (CFG)

```
plt.figure(figsize=(18,8))
for i in range(1, 6):

  n_guidance = i + 4
  generator = torch.Generator("cuda").manual_seed(seed)
  img = pipe(prompt, guidance_scale=n_guidance, generator=generator).images[0]

  plt.subplot(1,5,i)
  plt.title('guidance_scale: {}'.format(n_guidance))
  plt.imshow(img)
  plt.axis('off')

plt.show()
```

Gambar 4. 11 Nilai guidance scale (CFG) dari 5 hingga 9

Gambar 4. 12 Mengevaluasi pengaruh kekuatan generator mengikuti *prompt*

Guidance scale mengatur seberapa kuat generator mengikuti deskripsi dalam *prompt*. Nilai kecil menghasilkan gambar yang lebih bebas, sedangkan nilai besar lebih sesuai dengan *prompt*. Contoh nilai CFG adalah 5, 6, 7, 8, dan 9.

d. Image size (dimensions)

Gambar 4. 13 Gambar dengan dimensi 512x512

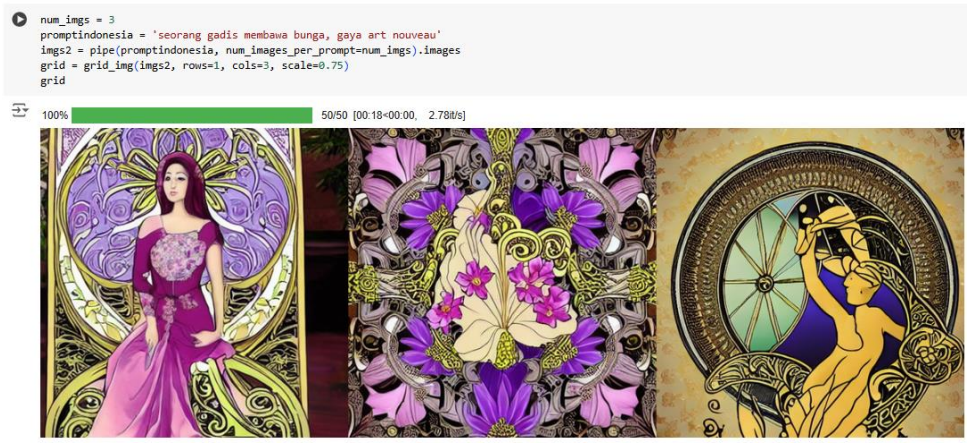
Dalam *Stable Diffusion* 1.5, ukuran gambar yang digunakan secara default adalah 512×512 piksel. Pemilihan ukuran ini didasarkan pada keseimbangan antara kualitas visual dan efisiensi komputasi. *Model Stable Diffusion* 1.5 dilatih dengan resolusi 512×512 piksel, sehingga ukuran ini menghasilkan gambar dengan tingkat akurasi dan detail optimal sesuai dengan data pelatihan. Selain itu, penggunaan dimensi persegi dipilih karena lebih seragam dalam distribusi piksel, memudahkan proses pemrosesan dan meminimalkan distorsi yang dapat terjadi jika menggunakan rasio aspek yang lebih ekstrem. Untuk menghasilkan gambar dengan resolusi berbeda, dapat menyesuaikan parameter *height* dan *width,* tetapi perubahan signifikan dari ukuran *default* dapat

mempengaruhi kualitas dan koherensi visual hasil yang dihasilkan oleh model.

4.2 *Prompt*

4.2.1 *Prompt* Indonesia

Penggunaan *prompt* dalam bahasa Indonesia bertujuan untuk menganalisis bagaimana sistem memahami dan menghasilkan gambar berdasarkan deskripsi dalam bahasa lokal. Hal ini penting untuk mengevaluasi sejauh mana model dapat mengenali dan menginterpretasikan konteks budaya serta struktur bahasa yang berbeda dari bahasa Inggris, yang umumnya menjadi bahasa utama dalam pelatihan model.



Gambar 4. 14 Penggunaan *Prompt* dengan Bahasa Indonesia

```
CLIP Score Gambar 1: 0.27490234375
CLIP Score Gambar 2: 0.2471923828125
CLIP Score Gambar 3: 0.2481689453125
Rata-rata CLIP Score: 0.2567545572916667
```

Gambar 4. 15 Hasil Clip Score Gambar yang Dihasilkan

Akurasi gambar yang dihasilkan dari *prompt* berbahasa Indonesia relatif lebih rendah dibandingkan dengan *prompt* berbahasa Inggris, yang umumnya lebih dioptimalkan untuk model ini. Hasil evaluasi menggunakan *CLIP Score* menunjukkan bahwa rata-rata kesesuaian gambar yang dihasilkan terhadap gambar referensi untuk *prompt* berbahasa Indonesia adalah 0,25, sedangkan

untuk *prompt* berbahasa Inggris mencapai 0,36. Perbedaan ini menunjukkan bahwa model lebih akurat dalam memahami deskripsi dalam bahasa Inggris, kemungkinan karena sebagian besar data pelatihan menggunakan bahasa tersebut.

Berdasarkan hasil tersebut, penelitian selanjutnya akan menggunakan *prompt* berbahasa Inggris untuk meningkatkan akurasi dan kualitas gambar yang dihasilkan. Penggunaan bahasa Inggris diharapkan dapat mengurangi ambiguitas dalam interpretasi model serta memastikan hasil yang lebih sesuai dengan deskripsi yang diberikan. Selain itu, pendekatan ini juga dapat membantu dalam membandingkan performa model dengan studi sebelumnya yang mayoritas menggunakan bahasa Inggris sebagai standar dalam proses generasi gambar.

4.2.2 *Negative prompt*

Negative prompt digunakan untuk menghindari elemen yang tidak diinginkan dalam gambar dengan memberikan instruksi eksplisit kepada model mengenai aspek visual yang harus dihindari. Dalam proses generasi gambar, *negative prompt* membantu meningkatkan relevansi hasil dengan cara mengurangi kemungkinan munculnya objek, warna, atau gaya tertentu yang tidak sesuai dengan deskripsi yang diinginkan. Penggunaan *negative prompt* menjadi penting terutama dalam memastikan konsistensi hasil, terutama ketika model cenderung menghasilkan elemen yang tidak sesuai dengan ekspektasi pengguna.

```
[ ]  neg_prompt = 'bad anatomy, ugly face, low quality, NSFW content'
```

Gambar 4. 16 neg_*prompt* yang Digunakan

Parameter ini dirancang untuk meningkatkan kualitas gambar yang dihasilkan dengan menghindari berbagai elemen yang tidak diinginkan. Dalam penerapannya, parameter ini berfungsi untuk mencegah anatomi tubuh yang tidak realistis (*bad anatomy*), memastikan wajah karakter tetap estetis

dan tidak mengalami distorsi (*ugly face*), serta meningkatkan ketajaman visual dengan mengeliminasi detail yang rendah atau kabur (*low quality*). Selain itu, parameter ini juga digunakan untuk menjaga keamanan konten dengan menghindari elemen yang bersifat eksplisit atau tidak pantas (*NSFW content*), sehingga hasil akhir tetap sesuai dengan standar etika dan estetika yang diharapkan.

Gambar 4. 17 Hasil Gambar

Hasil *CLIP Score* yang diperoleh, yaitu 0,70, 0,73, dan 0,64, menunjukkan bahwa kualitas gambar yang dihasilkan relatif baik dalam hal kesesuaian dengan gambar referensi. Skor yang mendekati angka 1,0 ini menandakan bahwa model berhasil memahami deskripsi dan merepresentasikannya dengan cukup akurat, meskipun terdapat variasi kecil antara setiap iterasi. Nilai yang lebih tinggi, seperti 0,73, menunjukkan hasil yang paling sesuai, sementara skor 0,64 masih menunjukkan tingkat kesesuaian yang dapat diterima, meskipun ada ruang untuk perbaikan dalam hal detail atau keselarasan visual.

```
Perbandingan Skor Kesamaan (Gambar yang Dihasilkan vs Gambar Referensi):
Gambar 1 - Skor rata-rata kesamaan dengan gambar referensi: 0.70
Gambar 2 - Skor rata-rata kesamaan dengan gambar referensi: 0.73
Gambar 3 - Skor rata-rata kesamaan dengan gambar referensi: 0.64
```

Gambar 4. 18 Perbandingan Skor Kesamaan

Dengan menggunakan *negative prompt*, gambar yang dihasilkan menunjukkan peningkatan kesesuaian dengan gambar referensi (rata-rata skor 0.69) dibandingkan kesesuaian dengan *prompt* asli (rata-rata skor 0.36). Ini mengindikasikan bahwa *negative prompt* efektif dalam memfilter elemen yang tidak diinginkan.

4.2.3 Spesifikasi Elemen *Prompt*

Pendekatan ini dilakukan dengan menambahkan spesifikasi yang lebih mendetail dalam *prompt*, seperti warna, posisi, atau gaya artistik tertentu, untuk memberikan arahan yang lebih jelas kepada model dalam menghasilkan gambar. Dengan memasukkan elemen-elemen spesifik ini, diharapkan model dapat menghasilkan hasil yang lebih sesuai dengan keinginan pengguna, baik dari segi estetika, komposisi, maupun kesesuaian visual. Penambahan detail ini juga bertujuan untuk mengurangi ketidakpastian dalam interpretasi *prompt*. Spesifikasi yang lebih rinci memungkinkan model untuk fokus pada aspek-aspek tertentu dalam gambar dan meningkatkan kualitas hasil secara keseluruhan.

Spesifikasi elemen *prompt* yang lebih detail, seperti "*surrounded by an ornate and decorate circular frame, muted pastel tones, soft lighting, and detailed textures*," membantu mengatasi masalah generasi latar belakang yang sering menjadi konstanta dalam gaya *Art Nouveau*. Pada gaya ini, latar belakang sering kali memiliki pola yang serupa, seperti penggunaan bingkai dekoratif berbentuk melingkar dan tekstur yang halus, yang menciptakan kesan kesatuan dan keharmonisan visual. Dengan memberikan instruksi spesifik dalam *prompt*, model dapat lebih fokus pada elemen-elemen tersebut, memastikan bahwa latar belakang mengikuti pola yang khas dan konsisten, seperti bingkai yang terornamen dengan warna pastel lembut dan pencahayaan yang lembut, yang merupakan ciri khas *Art Nouveau*. Hal ini membantu menjaga keselarasan antara karakter utama dan latar belakang, menghasilkan gambar dengan nuansa yang lebih kohesif dan estetis.

Gambar 4. 19 Gambar Referensi

Gambar 4. 20 Hasil Gambar

Hasil *CLIP Score* yang diperoleh setelah menambahkan spesifikasi detail dalam *prompt*, yaitu 0,75, 0,77, dan 0,79, menunjukkan peningkatan yang signifikan dalam kualitas gambar yang dihasilkan, terutama dalam hal kesesuaian dengan gambar referensi. Skor yang lebih tinggi, mendekati angka 1,0, mengindikasikan bahwa model berhasil memahami dan merepresentasikan elemen-elemen spesifik dalam deskripsi, seperti bingkai dekoratif berbentuk melingkar, warna pastel lembut, dan pencahayaan yang halus, dengan lebih akurat. Peningkatan skor ini mencerminkan efektivitas spesifikasi dalam memberikan arahan yang lebih jelas kepada model, sehingga menghasilkan gambar yang lebih kohesif dan sesuai dengan gaya *Art Nouveau* yang diinginkan. Skor 0,79 menunjukkan hasil yang sangat baik, sementara skor 0,75 dan 0,77 juga masih berada dalam rentang kesesuaian yang sangat memada

```
Perbandingan Skor Kesamaan (Gambar yang Dihasilkan vs Gambar Referensi):
Gambar 1 - Skor rata-rata kesamaan dengan gambar referensi: 0.75
Gambar 2 - Skor rata-rata kesamaan dengan gambar referensi: 0.77
Gambar 3 - Skor rata-rata kesamaan dengan gambar referensi: 0.79
```

Gambar 4. 21 Perbandingan Skor Kesamaan

Meskipun semua gambar mengikuti spesifikasi *prompt* yang sama, perbedaan tersebut terletak pada aspek komposisi dan interpretasi gaya *Art Nouveau* oleh model. Misalnya, pada satu gambar, bingkai dekoratif mungkin lebih tegas dengan detail yang lebih kaya, sementara pada gambar lainnya, elemen-elemen tersebut tampak lebih ringan dan halus. Selain itu, palet warna pastel yang digunakan juga menunjukkan variasi, dengan beberapa gambar cenderung lebih dominan pada warna lembut tertentu, sementara lainnya lebih seimbang.

4.3 Pengujian Model

4.3.1 Proses *Fine-tuning* Menggunakan *Dreambooth*

Pada bagian ini, dilakukan pengujian model menggunakan metode *fine-tuning* berbasis *Dreambooth*. Langkah-langkah ini mencakup pelatihan model dengan parameter yang telah ditentukan, pemuatan model hasil pelatihan, dan evaluasi hasil berupa visualisasi gambar sebelum dan sesudah pelatihan.

Pelatihan model dilakukan dengan menggunakan skrip berikut, yang dirancang khusus untuk menghasilkan gambar dalam gaya seni "*Art Nouveau*". Parameter penting seperti resolusi gambar, *learning rate*, dan jumlah langkah pelatihan telah disesuaikan dengan cermat untuk memastikan kualitas hasil yang optimal. Resolusi yang dipilih memungkinkan model untuk menangkap detail halus yang menjadi ciri khas gaya *Art Nouveau*, sementara *learning rate* disesuaikan agar model dapat belajar dengan efisien tanpa terjebak dalam solusi lokal yang kurang optimal. Selain itu, jumlah langkah pelatihan dipilih untuk memberikan waktu yang cukup bagi model untuk mengenali pola-pola halus dan kompleks dalam elemen visual *Art Nouveau*, seperti penggunaan ornamen, tekstur, dan komposisi yang khas.

Dengan penyesuaian parameter ini, diharapkan model dapat menghasilkan gambar yang tidak hanya sesuai dengan gaya yang diinginkan, tetapi juga memiliki kualitas visual yang tinggi dan akurat.

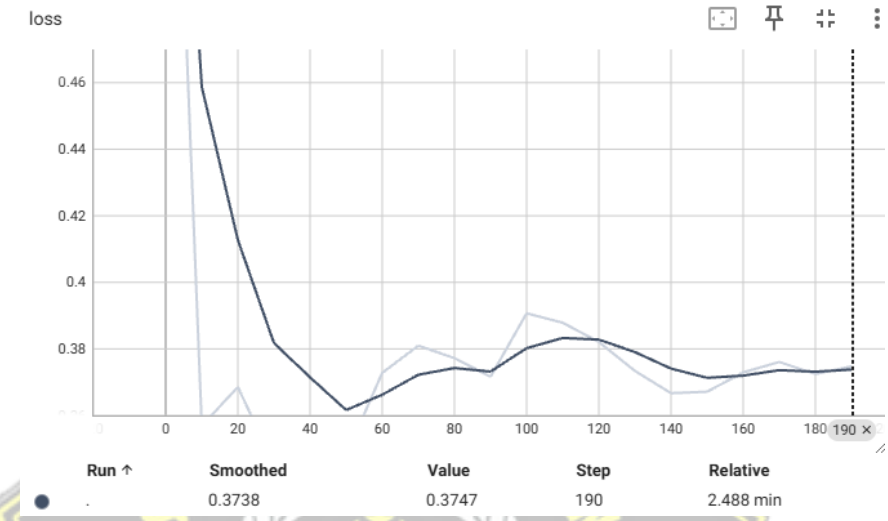
```
!python3 train_dreambooth.py \
  --pretrained_model_name_or_path="runwayml/stable-diffusion-v1-5" \
  --output_dir="/content/output" \
  --revision="main" \
  --with_prior_preservation --prior_loss_weight=1.0 \
  --seed=777 \
  --resolution=512 \
  --train_batch_size=1 \
  --train_text_encoder \
  --mixed_precision="fp16" \
  --use_8bit_adam \
  --gradient_accumulation_steps=1 \
  --learning_rate=1e-6 \
  --lr_scheduler="constant" \
  --lr_warmup_steps=80 \
  --num_class_images=20 \
  --sample_batch_size=4 \
  --max_train_steps=200 \
  --instance_prompt="art nouveau style" \
  --instance_data_dir="/content/file/MyDrive/training/nou" \
  --class_data_dir="/content/file/MyDrive/training/artn" \
  --class_prompt="art nouveau style"
```

Gambar 4. 22 Pengujian Model dengan *Fine-tune Dreambooth*

- `--with_prior_preservation`: Mengaktifkan penggunaan mekanisme prior preservation, yang bertujuan untuk menjaga kualitas fitur-fitur penting dari data latih sebelum diubah.
- `--prior_loss_weight=1.0`: Menentukan bobot dari prior *loss* yang digunakan untuk memastikan pelatihan tetap mempertahankan informasi prior dari data.
- `--seed=777`: Menetapkan nilai seed untuk memastikan reproduktibilitas hasil pelatihan, dengan menggunakan nilai seed tertentu agar hasil eksperimen dapat diulang.
- `--resolution=512`: Menentukan resolusi gambar yang digunakan dalam pelatihan, dalam hal ini 512x512 piksel, yang merupakan ukuran standar untuk model yang bekerja dengan gambar.

- --train_batch_size=1: Menetapkan ukuran *batch* pelatihan, yaitu jumlah sampel yang diproses dalam satu iterasi, diatur menjadi 1 untuk meminimalkan penggunaan memori.
- --train_text_*encoder*: Mengaktifkan pelatihan *encoder* teks, memungkinkan model untuk memproses deskripsi teks sebagai *input* untuk menghasilkan gambar.
- --mixed_precision="fp16": Menggunakan presisi campuran dengan tipe data 16-bit floating point (*fp16*), yang dapat mempercepat pelatihan dan mengurangi penggunaan memori tanpa mengurangi akurasi secara signifikan.
- --use_8bit_adam: Menggunakan optimizer Adam dengan representasi 8-bit untuk mengurangi penggunaan memori dan meningkatkan efisiensi komputasi.
- --gradient_accumulation_steps=1: Menentukan jumlah langkah akumulasi gradien yang dilakukan sebelum pembaruan parameter. Dengan nilai 1, ini berarti pembaruan dilakukan setiap langkah pelatihan.
- --learning_rate=1e-6: Menetapkan tingkat pembelajaran (*learning rate*) yang sangat kecil, yaitu 1e-6, untuk memastikan pembelajaran yang sangat hati-hati dan stabil.
- --lr_scheduler="constant": Menggunakan jadwal pembelajaran konstan, di mana *learning rate* tidak berubah selama pelatihan.
- --lr_warmup_steps=80: Menetapkan jumlah langkah pelatihan awal (80 langkah) untuk memanaskan *learning rate* secara perlahan sebelum mencapai nilai konstan.
- --num_class_images=20: Menentukan jumlah gambar kelas yang digunakan untuk melatih model pada setiap iterasi, yang dalam hal ini adalah 20 gambar.
- --sample_batch_size=4: Menentukan jumlah gambar yang dihasilkan dalam setiap *batch sampling*, yang diatur menjadi 4 untuk mengevaluasi hasil gambar lebih cepat.

- `--max_train_steps=200`: Menetapkan jumlah langkah pelatihan maksimum, yaitu 200 langkah, untuk mengontrol durasi pelatihan dan mencegah *overfitting* .

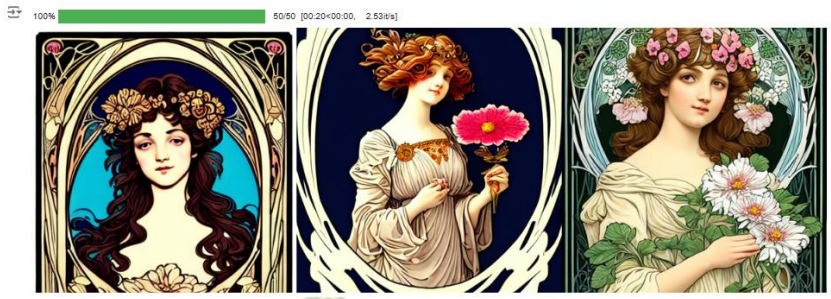


Gambar 4. 23 Grafik *Loss* Selama Proses Fine-tuning

Selama pelatihan, grafik *loss* menunjukkan nilai yang cukup stabil dengan hasil smoothed *loss* sebesar 0.3738 dan nilai *loss* aktual sebesar 0.3747. Perbedaan antara nilai smoothed dan nilai aktual *loss* terhitung relatif kecil, hanya sekitar 0.0009, yang menunjukkan bahwa fluktuasi *loss* selama pelatihan sangat minim. Dengan selisih yang sangat kecil ini, model menunjukkan konvergensi yang baik dan pelatihan berjalan dengan stabil. Selain itu, perubahan nilai *loss* yang relatif rendah, yaitu sekitar 2.488 menit, menunjukkan bahwa proses pelatihan berlangsung dengan efisien dan tidak terjadi perubahan besar dalam setiap iterasi pelatihan, menandakan kestabilan model dalam memahami data yang dilatih.

4.3.2 Pengujian Kualitas Gambar

1. Gambar yang dihasilkan dengan *prompt* biasa tanpa *negative prompt*.



Gambar 4. 24 Generasi Gambar dengan Model Baru

Gambar 4.24 menunjukkan generasi gambar yang dihasilkan dengan model baru hasil pelatihan. Elemen khas *Art Nouveau*, seperti pola garis melengkung dan motif bunga, terlihat dengan jelas dalam gambar ini, mencerminkan ciri khas gaya tersebut. Meskipun demikian, beberapa elemen yang tidak diinginkan, seperti anatomi yang kurang sempurna atau detail yang tidak terlalu halus, masih dapat terlihat dalam hasil gambar ini. Hal ini menunjukkan bahwa meskipun *fine-tuning* telah dilakukan, beberapa ketidaksempurnaan masih mungkin muncul, terutama tanpa penggunaan *negative prompt* yang dapat lebih efektif menghindari elemen-elemen yang tidak diinginkan dalam gambar.

2. Gambar yang dihasilkan dengan *prompt* biasa tetapi ditambah dengan *negative prompt*.

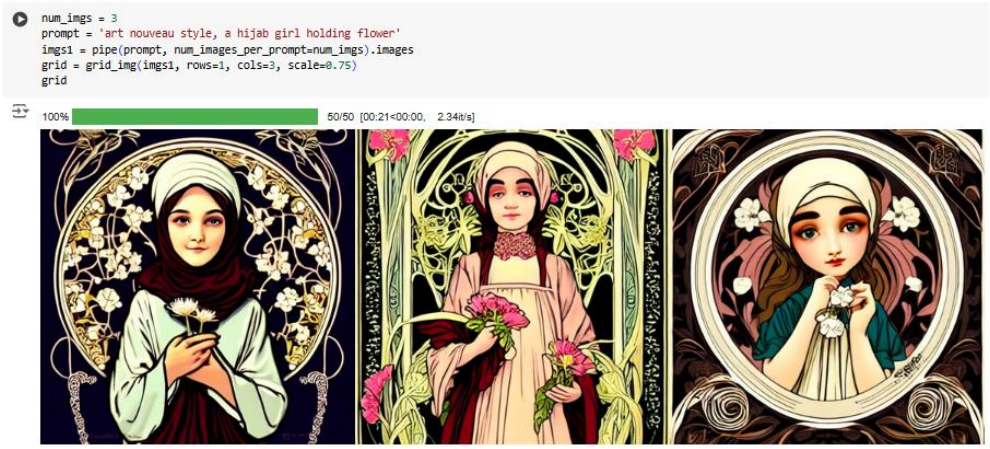
Gambar 4. 25 Generasi Gambar dengan Model Baru dan *Negative Prompt*

Gambar 4.26 menunjukkan generasi gambar yang dihasilkan dengan model baru setelah ditambahkan *negative prompt*. Dengan penggunaan *negative prompt*, elemen-elemen yang tidak diinginkan, seperti anatomi yang tidak proporsional atau ekspresi wajah yang kurang estetis, berhasil diminimalkan. Hasilnya adalah gambar yang lebih fokus pada gaya *Art Nouveau*, dengan elemen-elemen seperti pola garis melengkung dan motif bunga yang lebih jelas dan terdefinisi dengan baik. Selain itu, detail gambar menjadi lebih bersih dan konsisten, memberikan peningkatan kualitas visual secara keseluruhan. Penggunaan *negative prompt* terbukti efektif dalam meningkatkan hasil gambar, memastikan bahwa elemen-elemen yang tidak sesuai dengan estetika yang diinginkan dapat dihindari.

Penerapan *fine-tuning* pada model *Dreambooth* dengan tambahan *negative prompt* memberikan dampak terhadap kualitas gambar yang dihasilkan. Tanpa *negative prompt,* meskipun elemen *Art Nouveau* seperti garis melengkung dan motif bunga terlihat jelas, masih terdapat ketidaksempurnaan dalam aspek anatomi dan detail. Namun, dengan menambahkan *negative prompt*, elemen-elemen yang tidak diinginkan seperti anatomi yang tidak proporsional dan ekspresi wajah yang kurang estetis dapat diminimalkan, menghasilkan gambar yang lebih bersih, konsisten, dan lebih fokus pada gaya *Art Nouveau.*

4.3.3 Pengujian *Fidelity* dan Personalisasi Menggunakan *Prompt* Lain

Mengevaluasi tingkat akurasi (*fidelity*) antara deskripsi teks dan hasil gambar, serta kemampuan model untuk mempersonalisasi karakter dalam situasi yang lebih spesifik. Pengujian dilakukan dengan menggunakan variasi *prompt*, seperti menggambarkan perempuan berhijab,, dan mengganti subjek menjadi laki-laki.



Gambar 4. 26 Generasi Gambar dengan *Prompt* Lain

Hasil gambar menunjukkan representasi perempuan berhijab dalam gaya *Art Nouveau*, dengan tiga gambar yang memiliki perbedaan dalam gaya hijab. Dari ketiga gambar tersebut, hanya satu yang menggambarkan hijab dengan desain yang sepenuhnya menutupi rambut, sesuai dengan representasi hijab yang lebih tradisional. Sementara itu, dua gambar lainnya menampilkan hijab yang lebih ringan, hanya berfungsi sebagai penutup kepala tanpa menutupi rambut secara penuh. Meskipun elemen seperti motif bunga dan pola dekoratif *Art Nouveau* tetap dipertahankan, perbedaan gaya hijab ini mencerminkan variasi dalam interpretasi visual terhadap representasi perempuan berhijab dalam gaya seni tersebut. Penggunaan atribut hijab ini dapat ditingkatkan lebih lanjut untuk menciptakan keselarasan antara desain hijab dan elemen visual lainnya, agar lebih konsisten dalam menciptakan gambaran yang diinginkan.

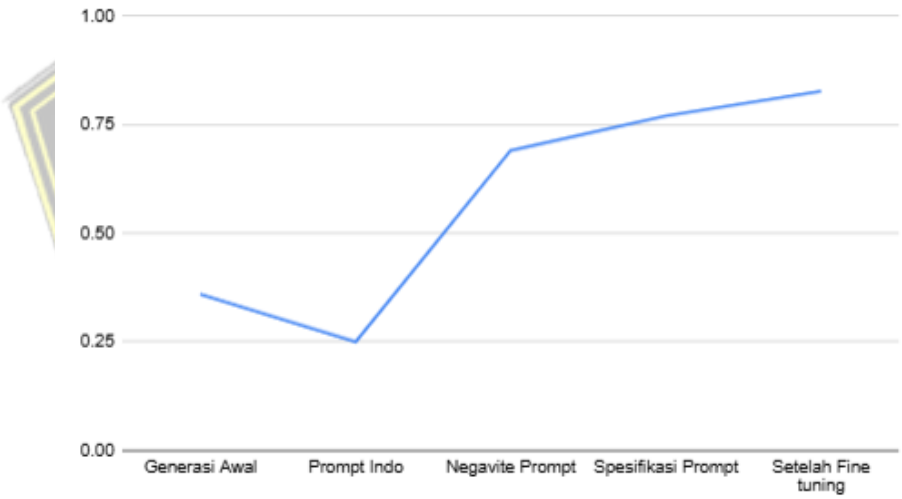
Gambar 4. 27 Penambahan *Negative Prompt*

Penambahan *negative prompt* membantu menghilangkan elemen yang tidak diinginkan, seperti anatomi yang kurang sempurna atau detail wajah yang tidak sesuai. Dalam gambar yang dihasilkan, dua gambar berhasil menampilkan perempuan dengan hijab yang menutupi aurat secara penuh, sementara satu gambar lainnya menggambarkan perempuan dengan hijab yang hanya berfungsi sebagai penutup kepala tanpa menutupi rambut sepenuhnya.

Pengujian selanjutnya mengganti subjek dari perempuan berhijab menjadi seorang laki-laki Yunani. Hasil gambar mencerminkan elemen gaya *Art Nouveau* dengan fokus pada karakter laki-laki dan elemen bunga yang diminta.

Gambar 4. 28 Generasi Gambar dengan *Prompt* Lain

Penyesuaian gaya dan atribut pada subjek laki-laki memberikan tantangan tersendiri, terutama dalam memastikan elemen fisik dan pakaian yang sesuai dengan ciri khas budaya Yunani. Pada gambar 4. 28, kesan Yunani tercermin melalui pakaian yang dililitkan dengan cara khas, serta aksesori tambahan seperti mahkota daun zaitun yang sering ditemukan dalam seni Yunani klasik. Meskipun fokusnya pada setengah badan, gambar ini tetap berhasil menampilkan keseimbangan antara elemen-elemen gaya *Art Nouveau* dan estetika Yunani, dengan penambahan motif bunga dan pola dekoratif yang memperkaya tampilan keseluruhan tanpa mengurangi kesan elegan dan halus pada karakter tersebut.



Gambar 4. 29 Gafik Nilai Pertumbuhan Clip Score

Kesimpulan dari pengujian menunjukkan bahwa model memiliki tingkat akurasi (*fidelity)* yang baik dalam mencocokkan deskripsi teks dengan hasil gambar, khususnya dalam merepresentasikan elemen gaya *Art Nouveau*. Penambahan *negative prompt* terbukti dalam meningkatkan kesesuaian gambar dengan deskripsi, dengan mengurangi elemen yang tidak diinginkan. Dalam hal personalisasi, model berhasil menghasilkan karakter perempuan berhijab dan laki-laki Yunani dengan elemen yang sesuai dengan deskripsi teks. Namun, atribut spesifik seperti pakaian tradisional, aksesori, dan detail

lainnya masih dapat ditingkatkan melalui pelatihan tambahan untuk mencapai tingkat akurasi yang lebih tinggi.

4.4 Hasil Implementasi Menggunakan Streamlit

Aplikasi AI Image Generator berhasil diimplementasikan menggunakan framework Streamlit, yang dirancang untuk membangun antarmuka pengguna berbasis web secara sederhana dan interaktif. Framework ini memungkinkan pengintegrasian model AI dengan antarmuka pengguna secara mulus, sehingga pengguna dapat menghasilkan gambar berdasarkan deskripsi teks yang dimasukkan.

Pada tampilan utama aplikasi, judul "*AI Image Generator*" ditampilkan di bagian atas sebagai identitas utama aplikasi. Di bawahnya, terdapat kolom *input* berbentuk persegi panjang yang memungkinkan pengguna memasukkan deskripsi atau *prompt*. Kolom ini dilengkapi dengan placeholder bertuliskan "*Please Enter your prompt here!*" untuk memandu pengguna. Tombol "*Generate Image*" ditempatkan di bawah kolom *input*, dengan desain minimalis berupa latar putih dan teks hitam, membuatnya mudah diakses.

Gambar 4. 30 Tampilan Awal Nouveau Dream

Gambar 4. 31 Generasi Gambar Menggunakan Streamlit

Setelah pengguna menekan tombol "*Generate Image*", aplikasi akan memberikan umpan balik berupa teks status seperti "*Loading Model...*" dan "*Generating Model...*" yang muncul tepat di bawah tombol. Feedback ini dirancang untuk memberi kejelasan bahwa proses sedang berlangsung. Setelah proses selesai, gambar hasil generasi ditampilkan pada area besar di bagian bawah halaman, dengan latar abu-abu yang dirancang agar hasil terlihat jelas.

Gambar 4. 32 Hasil Gambar Menggunakan Streamlit

4.4.1 Perbandingan Kinerja Berdasarkan Spesifikasi *Hardware*

Dalam pengujian sistem generasi gambar, waktu yang dibutuhkan untuk menghasilkan gambar bervariasi tergantung pada spesifikasi *Hardware* yang digunakan. Berikut adalah perbandingan kinerja pada beberapa konfigurasi perangkat keras yang diuji:

1. CPU i5 1135G7, RAM 16 GB

   Pada konfigurasi ini, sistem mampu menghasilkan gambar dengan rata-rata waktu sekitar 9 menit. Kinerja ini cukup optimal, memberikan hasil yang cepat tanpa adanya kendala yang signifikan selama proses generasi gambar.

2. CPU i3 6006U, RAM 12 GB

   Pada konfigurasi yang lebih rendah, yaitu menggunakan CPU i3 6006U dan RAM 12 GB, waktu yang dibutuhkan untuk menghasilkan gambar rata-rata adalah sekitar 25 menit. Proses ini lebih lama dibandingkan dengan sistem menggunakan CPU i5, yang menunjukkan bahwa kinerja CPU yang lebih rendah berdampak langsung pada waktu pemrosesan.

3. GPU NX450, VRAM 2 GB, float16

   Penggunaan GPU NX450 dengan VRAM 2 GB menunjukkan waktu rata-rata untuk menghasilkan gambar sekitar 12 menit. Meskipun lebih cepat dibandingkan dengan konfigurasi CPU i3, ada peningkatan kemunculan konten *NSFW* yang lebih tinggi. Hal ini mungkin terkait dengan cara GPU menangani proses generasi gambar yang dapat mempengaruhi hasil keluaran, terutama pada model-model yang lebih sensitif terhadap konten.

Pengujian terhadap berbagai spesifikasi *Hardware* menunjukkan bahwa spesifikasi perangkat keras memiliki dampak signifikan terhadap kinerja sistem dalam proses generasi gambar. Semakin tinggi spesifikasi perangkat keras yang digunakan, semakin cepat dan efisien proses pembuatan gambar. Kinerja CPU dan GPU secara signifikan mempengaruhi waktu pemrosesan serta kualitas gambar yang dihasilkan, penggunaan GPU dapat memberikan efisiensi waktu tetapi berisiko terhadap kualitas konten.

4.5 Regularisasi

4.5.1 *Dropout* Rate

Pada bagian ini, *Dropout* diterapkan pada CLIP *encoder* sebagai metode regularisasi untuk mencegah *overfitting* selama pelatihan model. *Dropout*

bekerja dengan menonaktifkan secara acak sejumlah unit dalam jaringan pada setiap iterasi, mendorong model untuk mempelajari representasi yang lebih robust dan tidak bergantung pada kombinasi fitur tertentu. Tingkat *Dropout* yang digunakan diatur secara hati-hati untuk memastikan keseimbangan antara generalisasi dan performa model, terutama pada tahap *Encoding* data teks dan gambar. Parameter *Dropout* rate yang dipilih dicantumkan untuk menunjukkan pengaruhnya terhadap stabilitas dan akurasi selama proses pelatihan.

```
class CLIPWithDropout(nn.Module):
    def __init__(self, original_model, dropout_rate):
        super(CLIPWithDropout, self).__init__()
        self.clip_model = original_model
        self.dropout = nn.Dropout(p=dropout_rate)

    def encode_image(self, image):
        image_features = self.clip_model.encode_image(image)
        image_features = self.dropout(image_features)
        return image_features

    def encode_text(self, text):
        text_features = self.clip_model.encode_text(text)
        text_features = self.dropout(text_features)
        return text_features
```

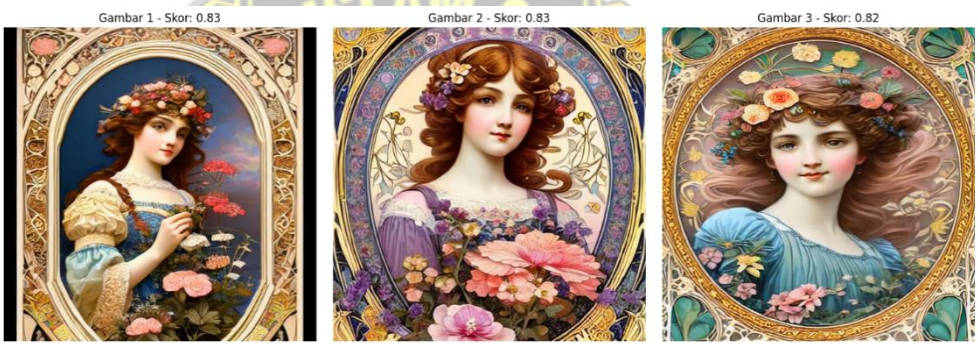
Gambar 4. 33 Penggunaan *Drop out* pada Clip *Encoder*

Dengan menonaktifkan secara acak sejumlah unit dalam jaringan, *Dropout* mendorong model untuk belajar representasi yang lebih fleksibel dan robust, yang tidak terikat pada pola-pola spesifik dari data pelatihan. Hal ini dapat menyebabkan penurunan dalam kesamaan antara gambar yang dihasilkan dengan *dataset* referensi, yang tercermin dalam pengurangan CLIP score. Penurunan CLIP score tersebut mungkin terjadi karena model tidak lagi mengandalkan pola atau fitur yang ditemukan dalam *dataset* dengan cara yang sama, namun lebih mengutamakan kemampuan untuk mengenali representasi yang lebih umum. Meskipun hal ini dapat mengurangi kesesuaian langsung dengan *dataset*, efek positifnya adalah model menjadi lebih mampu menghasilkan gambar yang beragam dan tidak terbatas pada

contoh yang telah ada dalam data pelatihan, sehingga meningkatkan fleksibilitas dan kemampuannya untuk menangani variasi dalam data baru.

4.5.2 Analisis *Overfitting* dan *Dropout*

Analisis ini membahas dampak penggunaan *Dropout* terhadap kemampuan model dalam mengatasi *overfitting* . Dengan membandingkan hasil pelatihan antara model yang menggunakan *Dropout* dan yang tidak, diperoleh gambaran tentang peran teknik ini dalam meningkatkan kemampuan generalisasi model. Grafik hasil pelatihan dan validasi disajikan untuk menunjukkan perbedaan tingkat akurasi serta pola *overfitting* yang terjadi pada kedua skenario tersebut.



Gambar 4. 34 Generasi Gambar "*a girl holding flower*" Sebelum Regularisasi *Drop out*

Gambar 4. 35 Hasil Generasi Gambar "*a girl holding flower*" Sesudah Regularisasi *Drop out*

Dalam kasus "a girl holding flower," perbandingan antara hasil sebelum dan sesudah penerapan *Dropout* menunjukkan perbedaan yang jelas dalam kinerja model. Sebelum *Dropout*, model menghasilkan CLIP score yang

tinggi, dengan skor 0.81, 0.83, dan 0.82, menunjukkan bahwa model sangat mengandalkan fitur-fitur tertentu yang ada dalam *dataset* yang mayoritas berisi gambar perempuan. Hasil ini mencerminkan ketergantungan model terhadap elemen-elemen yang sangat spesifik dari *dataset* pelatihan. Namun, setelah penerapan *Dropout*, meskipun skor CLIP sedikit menurun menjadi 0.78, 0.72, dan 0.81, terdapat peningkatan dalam eksplorasi elemen gaya artistik, seperti pewarnaan, garis, dan komposisi gambar. *Dropout* mendorong model untuk lebih kreatif dan fleksibel, menghasilkan gambar yang lebih bervariasi dan lebih baik dalam mengekspresikan elemen-elemen artistik yang sesuai dengan deskripsi *prompt*.

Gambar 4. 36 Generasi Gambar "*a man with pigeon*" Sebelum Regularisasi *Drop out*

Gambar 4. 37 Hasil Generasi Gambar "*a man with pigeon*" Sesudah Regularisasi *Drop out*

Pada kasus *"a man with pigeon,"* meskipun *dataset* yang digunakan mayoritas berisi gambar perempuan, eksperimen ini tetap menunjukkan perbedaan hasil yang signifikan antara sebelum dan sesudah penerapan *Dropout*. Sebelum *Dropout*, model menghasilkan CLIP score 0.73, 0.62, dan

0.70, yang menunjukkan bahwa model cenderung mengandalkan pola yang ada dalam *dataset* perempuan, seperti bentuk tubuh dan fitur wajah perempuan yang lebih dominan. Setelah penerapan *Dropout*, meskipun skor CLIP sedikit menurun menjadi 0.72, 0.71, dan 0.61, ada perbaikan dalam eksplorasi elemen-elemen gambar yang lebih kreatif, seperti gaya pakaian dan komposisi garis. Penurunan skor CLIP ini mengindikasikan bahwa *Dropout* berhasil mengurangi ketergantungan model terhadap *dataset* yang terbatas dan membantu model menghasilkan gambar yang lebih bervariasi meskipun subjek *prompt* berbeda dengan *dataset* pelatihan.

BAB V
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Penelitian ini telah berhasil menunjukkan bahwa deskripsi karakter berupa teks dapat diterjemahkan menjadi gambar bergaya *Art Nouveau* menggunakan *Diffusion Models*. Selain itu, dengan menerapkan *prompt* yang lebih spesifik dan negative *prompt*, model mampu meningkatkan fidelity hasil visualisasi. Teknik ini membantu memastikan gambar yang dihasilkan lebih akurat dan mendekati interpretasi yang diinginkan pengguna, sekaligus mengurangi elemen yang tidak diinginkan.

Hasil CLIP pada model sebelum penerapan Dropout menunjukkan rata-rata skor sekitar 0.30-an, namun setelah penerapan fine-tuning dan regulasi yang lebih ketat, nilai tersebut meningkat menjadi 0.80-an, menunjukkan perbaikan dalam pemahaman dan akurasi visualisasi. Meskipun demikian, penggunaan Dropout menghasilkan sedikit penurunan dalam nilai CLIP, sekitar 0.70-an. Meskipun demikian, ini tetap efektif dalam menghasilkan ilustrasi laki-laki, meskipun *dataset* yang digunakan tidak mencakup karakter laki-laki secara eksplisit.

5.2 Saran

Penelitian selanjutnya disarankan untuk mengeksplorasi penggabungan berbagai gaya seni guna menghasilkan visualisasi yang lebih beragam. Penggunaan *dataset* yang lebih luas dan mencakup elemen lingkungan atau latar juga dapat meningkatkan kemampuan personalisasi model. Selain itu, implementasi sistem dengan dukungan bahasa lokal, seperti Bahasa Indonesia, dapat menjadi langkah penting untuk meningkatkan relevansi dan aksesibilitas model dalam konteks lokal. Untuk meningkatkan efisiensi generasi gambar, penelitian berikutnya dapat memfokuskan pada optimasi arsitektur model atau penggunaan perangkat keras yang lebih canggih.

DAFTAR PUSTAKA

Alkhairi, P. *dkk.* (2024) "Optimasi LSTM Mengurangi Overfitting untuk Klasifikasi Teks Menggunakan Kumpulan Data Ulasan Film Kaggle IMDB," *Building of Informatics, Technology and Science (BITS)*, 6(2), hal. 1142–1150.

Anderson, J. dan Akram, N. (2024) "*Denoising* Diffusion Probabilistic Models (DDPM) Dynamics: Unraveling Change Detection in Evolving Environments," *Innovative Computer Sciences Journal*, 10(1), hal. 1–10.

Bao, F. *dkk.* (2023) "One transformer fits all distributions in multi-modal diffusion at scale," in *International Conference on Machine Learning*. PMLR, hal. 1692–1717.

Berahmand, K. *dkk.* (2024) *Autoencoders and their applications in machine learning: a survey, Artificial Intelligence Review*. Springer Netherlands. Tersedia pada: https://doi.org/10.1007/s10462-023-10662-6.

Bolya, D. dan Hoffman, J. (2023) "Token merging for fast *Stable Diffusion*," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, hal. 4599–4603.

Chávez, P. dan Ticona, W. (2024) "Implementation of Text-to-Image Generators in the Development of the Usability Interface for the Construction of a Web Page," in *2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. IEEE, hal. 926–930.

Chen, J. *dkk.* (2024) "Textdiffuser: *Diffusion Models* as text painters," *Advances in Neural Information Processing Systems*, 36.

Chen, S. dan Guo, W. (2023) "Auto-*encoder*s in deep learning—a review with new perspectives," *Mathematics*, 11(8), hal. 1777.

Croitoru, F.-A. *dkk.* (2023) "*Diffusion Models* in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9), hal. 10850–10869.

Croitoru, F.A. *dkk.* (2023) "*Diffusion Models* in Vision: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9), hal. 10850–10869. Tersedia pada:

https://doi.org/10.1109/TPAMI.2023.3261988.

D'Angelo, F. *dkk.* (2023) "Why do we need weight decay in modern deep learning?," *arXiv preprint arXiv:2310.04415* [Preprint].

Gu, Y., Fang, X. dan Deng, X. (2024) "A New Chinese Landscape Paintings Generation Model based on *Stable Diffusion* using *Dreambooth*," *arXiv preprint arXiv:2408.08561* [Preprint].

Gumulya, D. (2022) "Eksplorasi Material Inspirasi Gaya *Art Nouveau* Bertemu Dengan Ikon Indonesia Dengan Metode Atumics," *Jurnal Da Moda*, 3(2), hal. 69–78. Tersedia pada: https://doi.org/10.35886/damoda.v3i2.319.

Hua, M. *dkk.* (2023) "Dreamtuner: Single image is enough for subject-driven generation," *arXiv preprint arXiv:2312.13691* [Preprint].

Huberman-Spiegelglas, I., Kulikov, V. dan Michaeli, T. (2023) "An Edit Friendly DDPM *Noise* Space: Inversion and Manipulations," hal. 12469–12478. Tersedia pada: http://arxiv.org/abs/2304.06140.

Jung, C. *dkk.* (2024) "FlowAVSE: Efficient Audio-Visual Speech Enhancement with Conditional Flow Matching," hal. 2210–2214. Tersedia pada: https://doi.org/10.21437/interspeech.2024-701.

Krojer, B. *dkk.* (2023) "Are *Diffusion Models* Vision-And-Language Reasoners?," *Advances in Neural Information Processing Systems*, 36(NeurIPS), hal. 1–21.

Kushwaha, K. dan Srivastava, N.A. (2021) "Analytical study of Implication of *Art Nouveau* in designer's clothing," *International Journal of Innovation, Creativity and Change. www.ijicc.net*, 15(11), hal. 2021. Tersedia pada: www.ijicc.net.

Li, J.S. *dkk.* (2023) "Augmenters at SemEval-2023 Task 1: Enhancing CLIP in Handling Compositionality and Ambiguity for Zero-Shot Visual WSD through *Prompt* Augmentation and Text-To-Image Diffusion."

Lin, H. (2024) "DreamSalon: A Staged Diffusion Framework for Preserving Identity-Context in Editable Face Generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, hal. 8589–8598.

Luo, S. (2023) "Technical Report LCM-LORA: A UNIVERSAL STABLE-

DIFFUSION ACCELERATION MODULE," hal. 1–7.

Mak, H.W.L., Han, R. dan Yin, H.H.F. (2023) "Application of variational auto*Encoder* (VAE) model and image processing approaches in game design," *Sensors*, 23(7), hal. 3457.

Maulana, A.F. (2022) "Perancangan Sistem Informasi Perlombaan Berbasis Website untuk Kemudahan Penyampaian Informasi dan Pendaftaran Lomba," *OKTAL: Jurnal Ilmu Komputer dan Sains*, 1(03), hal. 263–270.

Murtopo, A.A. *dkk.* (2024) "PENERAPAN COMPUTER VISION UNTUK MENDETEKSI KELENGKAPAN ATRIBUT SISWA MENGGUNAKAN METODE CNN," *PROSISKO: Jurnal Pengembangan Riset dan Observasi Sistem Komputer*, 11(2), hal. 247–258.

Niu, Y. *dkk.* (2024) "Multi-model Style-aware Diffusion Learning for Semantic Image Synthesis," *ACM Transactions on Multimedia Computing, Communications, and Applications* [Preprint]. Tersedia pada: https://doi.org/10.1145/3686155.

Park, J., Ko, B. dan Jang, H. (2023) "StyleBoost: A Study of Personalizing *Text-to-Image Generation* in Any Style using *Dreambooth*," in *2023 14th International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, hal. 93–98.

Prasad, A. *dkk.* (2024) "*Stable Diffusion* Image Processing," *Library Progress International*, 44(3), hal. 5917–5925.

Radityasari, R.R.P. *dkk.* (2023) "Review Penggayaan Bangunan Casa Batllo: *Art Nouveau*," *JURNAL SYNTAX IMPERATIF: Jurnal Ilmu Sosial dan Pendidikan*, 4(4), hal. 453–459.

Rahmatulloh, A. (2024) "Custom Concept Text-to-Image Using *Stable Diffusion* Model in Generative Artificial Intelligence," *International Journal of Informatics and Computing*, 1(1), hal. 1–11.

Ruiz, N. *dkk.* (2023) "*Dreambooth*: Fine tuning text-to-image *Diffusion Models* for subject-driven generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, hal. 22500–22510.

Saraswati, I., Utami, N.M.P. dan Pemayun, T.U.N. (2024) "Newcomers Self-

Adjusment And Interpersonal Communication As An Idea For Digital Painting Art Creation," *Cita Kara : Jurnal Penciptaan Dan Pengkajian Seni Murni*, 4(2), hal. 247–256. Tersedia pada: https://doi.org/10.59997/ctkr.v4i2.3301.

Shi, J. (2024) "InstantBooth : Personalized *Text-to-Image Generation* without Test-Time Finetuning," hal. 8543–8552.

Siddique, N. *dkk.* (2021) "U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications," *IEEE Access*, 9, hal. 82031–82057. Tersedia pada: https://doi.org/10.1109/ACCESS.2021.3086020.

Tao, M. *dkk.* (2022) "Df-gan: A simple and effective baseline for text-to-image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, hal. 16515–16525.

Weng, W. dan Zhu, X. (2021) "UNet: Convolutional Networks for Biomedical Image Segmentation," *IEEE Access*, 9, hal. 16591–16603. Tersedia pada: https://doi.org/10.1109/ACCESS.2021.3053408.

Wu, Q. *dkk.* (2023) "Uncovering the Disentanglement Capability in Text-to-Image *Diffusion Models*," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2023-June, hal. 1900–1910. Tersedia pada: https://doi.org/10.1109/CVPR52729.2023.00189.

Yu, H. *dkk.* (2024) "Uncovering the Text *Embedding* in Text-to-Image *Diffusion Models*." Tersedia pada: http://arxiv.org/abs/2404.01154.

Zhang, S. (2023) "*Dreambooth*-based image generation methods for improving the performance of cnn," in *2023 IEEE 3rd International Conference on Electronic Technology, Communication and Information (ICETCI)*. IEEE, hal. 1181–1184.

Zhu, Y. *dkk.* (2023) "Conditional Text Image Generation with *Diffusion Models*," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2023-June, hal. 14235–14244. Tersedia pada: https://doi.org/10.1109/CVPR52729.2023.01368.