# PEMBUATAN LANSKAP FANTASI BERBASIS *DIFFUSION MODEL* DENGAN *FINE-TUNING LOW-RANK ADAPTATION* (LoRA)

**LAPORAN TUGAS AKHIR**

Laporan Ini Disusun Untuk Memenuhi Salah Satu Syarat Memperoleh Gelar Sarjana Strata 1 (S1) Pada Program Studi S1 Teknik Informatika Fakultas Teknologi Industri Universitas Islam Sultan Agung


**DISUSUN OLEH :**

**TSABIT GHOLIB**

**NIM 32602100120**

**PROGRAM STUDI TEKNIK INFORMATIKA**

**FAKULTAS TEKNOLOGI INDUSTRI**

**UNIVERSITAS ISLAM SULTAN AGUNG**

**SEMARANG**


**2025**

# *GENERATION* FANTASY LANDSCAPE BASED ON *DIFFUSION MODELS* WITH *LOW-RANK ADAPTATION* (LoRA) *FINE-TUNING*

**FINAL PROJECT REPORT**

This report is prepared as part of the requirements for obtaining a Bachelor's Degree (S1) in the Informatics Engineering Program, Faculty of Industrial Technology, Sultan Agung Islamic University.


**Arranged By :**

**TSABIT GHOLIB**
**NIM 32602100120**

**INFORMATICS ENGINEERING DEPARTEMENT**
**FACULTY OF INDUSTRIAL TECHNOLOGY**
**SULTAN AGUNG ISLAMIC UNIVERSITY**
**SEMARANG**

**2025**

**LEMBAR PENGESAHAN**
**TUGAS AKHIR**
**PEMBUATAN LANSKAP FANTASI BERBASIS *DIFFUSION MODEL***
**DENGAN *FINE-TUNING LOW-RANK ADAPTATION* (LORA)**

**TSABIT GHOLIB**
**NIM 32602100120**

Telah dipertahankan di depan tim ujian sarjana tugas akhir
Program Studi Teknik Informatika
Universitas Islam Sultan Agung
Pada tanggal : 25-3-2025

**TIM PENGUJI TUGAS AKHIR:**

| | | |
|---|---|---|
| **Ghufron ST, M.Kom**<br>**NIDN. 062079005**<br>(Ketua Penguji) | ..... | 5-3-2025 |
| **Badie'ah ST, M.Kom**<br>**NIDN. 0619018701**<br>(Anggota Penguji) | ..... | 4-3-2025 |
| **Ir. Sri Mulyono, M.Eng**<br>**NIDN. 0626066601**<br>(Pembimbing) | ..... | 6-3-2025 |

Semarang, .....
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 | : Tsabit Gholib |
| NIM | : 32602100120 |
| Judul Tugas Akhir | :**PEMBUATAN LANSKAP FANTASI BERBASIS DIFFUSION MODEL DENGAN *FINE-TUNING LOW-RANK ADAPTATION (LoRA)*** |

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 dikemudian 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

Yang Menyatakan


Tsabit Gholib

**PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH**

Saya yang bertanda tangan dibawah ini :

Nama :Tsabit Gholib

NIM : 32602100120

Program Studi : Teknik Informatika

Fakultas : Teknologi Industri

Dengan ini menyatakan Karya Ilmiah berupa Tugas akhir dengan Judul : **PEMBUATAN LANSKAP FANTASI BERBASIS *DIFFUSION MODEL* DENGAN *FINE-TUNING LOW-RANK ADAPTATION* (LoRA)**

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

Yang Menyatakan


Tsabit Gholib

## KATA PENGANTAR

Puji syukur peneliti panjatkan kepada Allah swt, alhamdulillah atas rahmat dan karuniaNya peneliti dapat menyelesaikan tugas akhir dengan judul "Pembuatan Lanskap Fantasi Berbasis *Diffusion model* Dengan *Fine-tuning Low-Rank Adaptation* (LoRA)". Tugas Akhir ini disusun sebagai salah satu syarat untuk menyelesaikan studi pada program sarjana di Universitas Islam Sultan Agung.

Dalam proses penyusunan Tugas Akhir ini, peneliti mendapat banyak sekali dukungan, bimbingan, serta motivasi dari berbagai pihak. Oleh karena itu, peneliti ingin mengucapkan rasa terima kasih yang sebesar-besarnya kepada :

1. Prof. Dr. H. Gunarto, SH., M.Hum, Selaku Rektor Universitas Islam Sultan Agung, yang telah memberikan kesempatan kepada peneliti untuk belajar di universitas ini.
2. Dr. Ir. Hj.Novi Marlyana, S.T., M.T, Selaku Dekan Fakultas Teknologi Industri yang telah memberikan kesempatan untuk belajar di fakultas ini
3. Moch Taufik, S.T., M.Kom, selaku Ketua Program Studi S1 Teknik Informatika yang telah memberikan dukungan akademik selama proses penelitian berlangsung.
4. Ir. Sri Mulyono, selaku dosen pembimbing yang telah memberikan bimbingan, arahan, serta masukan yang sangat berharga dalam penyusunan tugas akhir ini.
5. Seluruh Dosen dan Staf Akademik Fakultas Teknologi Industri, yang telah memberikan ilmu serta pengalaman berharga selama proses perkuliahan.

Peneliti menyadari bahwa tugas akhir ini masih jauh dari kesempurnaan, oleh karena itu, kritik dan saran yang membangun sangat diharapkan untuk perbaikan di masa mendatang. semoga tugas akhir ini dapat bermanfaat bagi para pembaca dan menjadi referensi bagi penelitian-penelitian selanjutnya.

Semarang, 20 Februari 2025
Penulis

Tsabit Gholib

## DAFTAR ISI

## DAFTAR TABEL

## DAFTAR GAMBAR

## ABSTRAK

*Adanya peningkatan jumlah pemain video game membuat industri game bersaing untuk menciptakan video game yang menarik, salah satunya dengan cara set-up lanskap secara imersif. Penelitian ini bertujuan untuk mengembangkan sistem yang mampu menciptakan lanskap fantasi dari deskripsi teks menggunakan Diffusion Models dan Low Rank Adaptation sebagai metode fine-tuning. Stable diffusion adalah jenis generative models yang dapat menghasilkan gambar dari teks. Low Rank Adaptation merupakan salah satu teknik fine-tuning yang hanya mengambil beberapa layer untuk dilatih. Dengan kolaborasi antara Stable diffusion 2.1 dan fine-tuning Low Rank Adaptation (LoRA) mampu menghasilkan gambar lanskap fantasi yang dapat digunakan untuk menjadi landasan dalam pembuatan konsep art pada sebuah video game. Hasil pengujian menggunakan CLIP-based Maximum Mean Discrepancy (CMMD) adalah 0.756 yang menunjukkan bahwa skor model yang dihasilkan lebih rendah daripada model awal yang belum melalui proses fine-tuning yang memiliki skor 0.867. Serta nilai skor Fréchet Inception Distance (FID) yaitu 29.96 yang menandakan model sudah bagus dan mirip dengan konteks dataset asli.*

*Kata kunci : diffusion model, stable diffusion, Low Rank Adaptation, lanskap fantasi.*

*The increasing number of video game players makes the game industry compete to create interesting video games, one of which is by setting up immersive landscapes. This research aims to develop a system capable of creating fantasy landscapes from text descriptions using Diffusion Models and Low Rank Adaptation as fine-tuning methods. Stable diffusion is a type of generative models that can generate images from text. Low Rank Adaptation is a fine-tuning technique that only takes a few layers to train. With the collaboration between Stable diffusion 2.1 and fine-tuning Low Rank Adaptation (LoRA) is able to produce fantasy landscape images that can be used as a basis for making concept art in a video game. The test result using CLIP-based Maximum Mean Discrepancy (CMMD) is 0.725 which shows that the resulting model Score is lower than the initial model that has not gone through the fine-tuning Process which has a Score of 0.867. And the value of the Fréchet Inception Distance (FID) Score is 29.96 which indicates that the model is good and is similar to the original dataset context.*

*Keyword : diffusion model, stable diffusion, Low Rank Adaptation, fantasy landscape.*

# BAB I
# PENDAHULUAN

## 1.1 Latar Belakang

Dalam industri video game, desain lingkungan dunia imajiner memiliki peran krusial dalam membangun atmosfer dan memperkuat narasi. Pada tahun 2024, jumlah pemain video game diperkirakan mencapai 3,32 miliar di seluruh dunia (Howarth, 2024), mendorong pengembang untuk menciptakan pengalaman bermain yang lebih imersif. Salah satu aspek penting dalam pengembangan dunia game adalah concept art, yang memberikan panduan visual untuk menciptakan lanskap, karakter, dan elemen estetika lainnya sebelum tahap produksi. Dalam genre fantasi, lanskap tidak hanya berfungsi sebagai latar tetapi juga memperkaya cerita dan gameplay, sebagaimana terlihat dalam nominasi *Game of the Year* 2024 seperti "*Astro Bot*", "*Black Myth: Wukong*", dan "*Final Fantasy VII Rebirth*" (Awards, 2024).

Pembuatan lanskap fantasi dalam *concept art* menuntut keseimbangan antara kreativitas dan kebutuhan teknis. Tantangan utama dalam proses ini meliputi kejelasan visual, keseimbangan detail artistik dengan performa teknis, serta keterbatasan waktu dan anggaran produksi. Desain dunia yang buruk dapat mengurangi keterlibatan pemain, sementara desain yang terlalu kompleks bisa menurunkan performa teknis (Mehrafrooz, 2024). Oleh karena itu, seniman dan desainer harus mengoptimalkan proses kreatif agar menghasilkan lingkungan game yang menarik dan efisien.

Teknologi kecerdasan buatan, khususnya *diffusion* model seperti *Stable diffusion*, menawarkan solusi inovatif untuk mengotomatisasi proses penciptaan visual. Model ini mampu menghasilkan gambar berkualitas tinggi dari deskripsi teks dalam waktu singkat, sehingga mempercepat produksi *concept art*. Dalam industri game, teknologi ini telah digunakan dalam simulasi game seperti DOOM untuk menciptakan visual yang dinamis secara real-time (Valevski dkk., 2024). Selain itu, model generatif ini mempermudah pembuatan lanskap dan elemen grafis lainnya tanpa memerlukan usaha manual yang besar (Rantanen, 2023).

Salah satu inovasi dalam teknologi generatifadalah penggunaan Low-Rank Adaptation (LoRA) sebagai teknik fine-tuning. LoRA meningkatkan kemampuan *diffusion* model dalam menghasilkan representasi visual yang lebih kaya dan sesuai dengan deskripsi imajinatif. Teknologi ini telah digunakan dalam industri game dan film untuk mempercepat proses kreatif dan meningkatkan kolaborasi antara manusia dan AI. Model seperti Runway memungkinkan eksplorasi desain yang lebih dinamis dengan menciptakan concept art dan storyboard secara otomatis (Totlani, 2023).

Penelitian ini bertujuan untuk mengembangkan teknik penciptaan lanskap fantasi yang lebih efisien menggunakan *diffusion* model dan LoRA. Dengan mengeksplorasi kolaborasi antara manusia dan AI, penelitian ini akan meninjau tantangan serta dampak implementasi teknologi ini terhadap industri kreatif. Diharapkan, penelitian ini dapat berkontribusi dalam memperluas pemahaman tentang pemanfaatan AI dalam pengembangan lanskap fantasi dan memperkaya batasan kreativitas di berbagai bidang.

## 1.2 Perumusan Masalah

Bagaimana cara menciptakan sistem yang dapat membuat lanskap fantasi dari deskripsi teks dengan menggunakan *diffusion model* khususnya *stable diffusion* v2.1 dan *Low-Rank Adaptation* (LoRA) sebagai metode *fine-tuning*?

## 1.3 Pembatasan Masalah

Adapun batasan masalah pada penelitian ini yaitu :

1. Penelitian ini hanya membuat visualisasi lanskap berdasarkan deskripsi teks, tanpa mencakup karakter atau elemen lain yang berhubungan dengan cerita.
2. Penelitian ini dibatasi pada pembuatan gambar lanskap fantasi, tidak mencakup lanskap dunia nyata atau genre lain selain fantasi.
3. Gambar yang dihasilkan terbatas pada resolusi tertentu untuk memastikan konsistensi dalam kualitas visual.

## 1.4 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah mengembangkan sistem yang mampu menciptakan lanskap fantasi dari deskripsi teks menggunakan *diffusion model* dan LoRA sebagai metode *fine-tuning*.

## 1.5 Manfaat

Adapun manfaat dalam penelitian ini yaitu :

1. Menyediakan alternatif dalam pembuatan lanskap fantasi untuk kebutuhan *concept art* dalam video *game*.
2. Meningkatkan variasi dan opsi dalam desain lanskap fantasi
3. Mendukung pengembangan studio *game* dengan sumber daya yang terbatas

## 1.6 Sistematika Penulisan

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

BAB I : PENDAHULUAN

Bab ini menjelaskan mengenai latar belakang pemilihan judul tugas akhir "Pembuatan Lanskap Fantasi Berbasis *Diffusion Model* Dengan *Fine-tuning Low-Rank Adaptation* (LoRA)." Rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian dan sistematika penulisan.

BAB II : Tinjauan Pustaka Dan Dasar Teori

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

BAB III : METODE PENELITIAN

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

BAB IV : HASIL PENELITIAN DAN IMPLEMENTASI SISTEM

Bab ini menjelaskan hasil dari penelitian yang telah dilakukan.

BAB V: 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

Teknologi *text-to-image* telah menjadi topik utama dalam perkembangan kecerdasan buatan, terutama dalam upaya menghasilkan gambar dari deskripsi teks dengan presisi dan estetika yang tinggi (Li dkk., 2023.). Menurut Chenshuang Zhang dkk (2024), kemampuan ini memungkinkan model kecerdasan buatan menerjemahkan ide-ide konseptual menjadi visualisasi yang nyata, sebuah lompatan besar dalam aplikasi *Computer Vision* dan NLP. Pada awalnya, teknik seperti *Generative Adversarial Networks* (GANs) digunakan untuk tugas ini, namun GANs memiliki kelemahan, seperti ketidakstabilan selama pelatihan dan akumulasi kesalahan pada gambar yang kompleks.

Dalam beberapa tahun terakhir, *Diffusion model* muncul sebagai alternatif yang lebih stabil dan efektif dibanding kan GANs untuk menghasilkan gambar berbasis teks. Penelitian sebelumnya menunjukkan bahwa *diffusion model* lebih unggul dibandingkan model generatif lainnya, seperti *Generative Adversarial Networks* (GANs), dalam hal stabilitas dan kualitas gambar. GANs sering menghadapi masalah ketidakstabilan pelatihan dan *mode collapse*, sedangkan *diffusion model* mampu menghasilkan gambar yang lebih realistis dan kaya akan detail, terutama ketika menangani latar belakang dan tekstur yang kompleks (Peng, 2024). Selain itu, *diffusion model* dapat dikondisikan dengan teks atau parameter tambahan untuk memberikan kontrol yang lebih tinggi terhadap hasil menghasilkan, seperti mengontrol elemen geometris atau tekstur tertentu. Hal ini relevan dengan pembuatan lanskap fantasi, di mana setiap detail seperti pegunungan, sungai, atau suasana magis harus diterjemahkan secara tepat sesuai deskripsi pengguna.

Meskipun *diffusion model* kuat dalam menghasilkan gambar, diperlukan teknik tambahan untuk mencapai personalisasi yang lebih dalam yaitu berupa teknik *fine-tuning*. Salah satu pendekatan yang menonjol adalah *Low-Rank Adaptation* (LoRA), "pilihan kami untuk *fine-tuning Stable diffusion* karena LoRA menawarkan keseimbangan yang baik antara ukuran berkas dan kekuatan

pelatihan"(Jin dan Song, 2023). LoRA memungkinkan penambahan fitur atau gaya baru dengan cara yang hemat sumber daya, menjadikannya ideal untuk kebutuhan spesifik seperti pembuatan komik dengan gaya artistik tertentu. Dalam konteks tugas akhir ini, LoRA akan digunakan untuk menyesuaikan *diffusion model* dengan tema dan gaya artistik tertentu yang sesuai, sehingga meningkatkan akurasi dan estetika hasil menghasilkan dalam pembuatan lanskap fantasi.Penelitian terkait juga menunjukkan bahwa teknik *fine-tuning* seperti LoRA mampu meningkatkan hasil model dalam skenario spesifik. Sebagai contoh, penerapan LoRA pada *stable diffusion* memungkinkan model menghasilkan gambar dengan detail gaya tertentu yang lebih tajam dan koheren setelah proses *fine-tuning*. Hasil pengujian menunjukkan bahwa DataInf dapat meningkatkan performa model dibandingkan model awal, terutama dalam metrik visualisasi FID yang lebih rendah, menunjukkan peningkatan kualitas gambar (Kwon dkk., 2024).

Selain itu, integrasi *diffusion model* dengan teknik *fine-tuning* seperti LoRA mendukung kebutuhan personalisasi dalam industri kreatif. Dalam konteks pembuatan lanskap fantasi, personalisasi ini memungkinkan model menyesuaikan *output* berdasarkan gaya seniman atau estetika tertentu yang diinginkan pengguna, seperti dunia magis ala cerita rakyat atau latar futuristik. Pengguna dapat menggabungkan elemen-elemen unik seperti pola warna tertentu, tata cahaya spesifik, hingga bentuk geometri yang tidak biasa, sehingga menghasilkan visualisasi yang lebih kaya dan imersif. Seperti yang dicatat oleh (Wang dkk, 2024), *diffusion model* telah digunakan secara luas dalam sektor hiburan untuk menciptakan efek visual yang realistis, lingkungan virtual, dan animasi dalam video *game*, mengurangi biaya dan waktu produksi secara signifikan. Hal ini menunjukkan bahwa model generatif tidak hanya mampu menghasilkan visual yang kompleks tetapi juga memungkinkan kolaborasi yang lebih kreatif antara manusia dan AI. Dengan demikian, seniman dapat menggunakan AI sebagai alat yang memperkaya proses desain mereka tanpa kehilangan aspek artistik atau identitas visual yang diinginkan

Secara keseluruhan, penggunaan *diffusion model* dengan *fine-tuning* melalui LoRA menawarkan pendekatan yang menjanjikan untuk menghasilkan lanskap

fantasi secara otomatis dari deskripsi teks. Teknik ini tidak hanya memungkinkan kontrol yang lebih besar atas elemen visual, tetapi juga dapat mengurangi kebutuhan komputasi dibandingkan dengan pelatihan ulang model besar dari awal. Oleh karena itu, penerapan LoRA pada *diffusion model* dapat menjadi solusi ideal dalam menciptakan visualisasi lanskap fantasi berkualitas tinggi, membuka peluang baru dalam industri kreatif, *game*, dan hiburan.

## 2.2 Dasar Teori

### 2.2.1 Lanskap Fantasi

Lanskap fantasi adalah representasi visual dari dunia imajinatif yang tidak terikat pada realitas fisik. Lanskap ini sering kali menggabungkan elemen-elemen fantastis, seperti lingkungan yang tidak biasa, makhluk mitologis, atau struktur yang melampaui logika dunia nyata. Seni lanskap fantasi menciptakan ruang yang mendukung eksplorasi ide-ide kreatif dan narasi visual yang mendalam, sering digunakan dalam berbagai media, seperti lukisan, film, animasi, dan permainan video.

Dalam seni lanskap fantasi, terdapat keunikan dalam menggabungkan unsur realisme dengan imajinasi untuk menciptakan suasana yang magis dan memikat. Sebagaimana lanskap dalam karya *Lugas Syllabus* yang umumnya menampilkan elemen seperti horison laut biru, perbukitan, rawa-rawa, serta gerumbul yang dihuni oleh berbagai figur fantasi dari beragam latar belakang (Wahyu, 2024). Lanskap ini berfungsi sebagai ruang antara realitas dan imajinasi, memberikan pengalaman visual yang mendalam sekaligus menyampaikan pesan simbolis atau naratif yang kaya.

### 2.2.2 *Text-to-image Generation*

*Text-to-image Generation* adalah teknologi kecerdasan buatan yang mengubah deskripsi teks menjadi gambar visual. Proses ini dimulai dengan tokenisasi, di mana teks dibagi menjadi unit-unit kecil seperti kata atau frasa, yang kemudian diubah menjadi representasi numerik. Setelah teks ter-tokenisasi, model *Natural Language Process*ing (NLP) digunakan untuk memahami makna dan hubungan antar kata. Representasi numerik ini kemudian diproses oleh model

generatif seperti *Diffusion Models* untuk menghasilkan gambar yang sesuai dengan deskripsi. Teknologi ini memanfaatkan model yang dilatih pada *dataset* pasangan teks-gambar untuk mempelajari hubungan antara deskripsi teks dan representasi visual, memungkinkan pembuatan gambar yang kaya dan beragam dalam berbagai genre dan tema (Sohn dkk., 2023).

### 2.2.3 *Diffusion model*

*Diffusion model* adalah salah satu jenis *generative models* yang digunakan untuk menghasilkan data, seperti gambar, dari distribusi yang rumit. Prosesnya terdiri dari dua langkah utama: *noise* dan *generative models*.

Pada *noise*, *noise* secara bertahap ditambahkan ke data (misalnya, gambar) melalui serangkaian langkah waktu hingga data tersebut berubah menjadi *noise* murni. Tujuannya adalah untuk mensimulasikan transformasi data ke distribusi Gaussian standar. Proses ini dapat dijelaskan dengan rumus sebagai berikut:

$$q(xt|xt-1) = N\left(xt; \sqrt{1-\beta t} \cdot xt-1, \beta t \cdot \mathrm{I}\right) \quad (1)$$

Di mana:

a. $x_t$ adalah data pada langkah $t$,
b. $x_t - 1$ adalah data pada langkah sebelumnya,
c. $\beta_t$ adalah skala *noise* kecil yang dikendalikan di setiap langkah,
d. $N$ adalah distribusi normal, dan
e. $I$ adalah matriks identitas.

Dengan mengaplikasikan proses ini selama $T$ langkah, data asli $x_0$ akan menjadi *noise* murni $x_T$.

Pada *generative models*, proses berlawanan dilakukan. Model dilatih untuk memprediksi dan menghilangkan *noise* secara bertahap, mengembalikan data dari *noise* menuju gambar yang bermakna. Proses ini bersifat probabilistik dan menggunakan parameter yang dipelajari selama pelatihan. Rumus *generative models* adalah:

$$p\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu\theta(x_t, t), \sum_\theta(x_t, t)) \quad (2)$$

Di mana:

1. $\mu\theta(x_t, t)$ adalah mean prediksi yang dipelajari oleh model,
2. $\sum\theta(x_t, t)$ adalah varians yang dipelajari, dan

3. $\theta$ merepresentasikan parameter model yang dioptimalkan.

Model dilatih untuk meminimalkan *loss* yang mengukur perbedaan antara distribusi forward dan *reverse*. Salah satu fungsi *loss* yang umum digunakan adalah:

$$L = \mathbb{E}_q[||\in - \in_\theta (x_t, t)||^2] \quad (3)$$

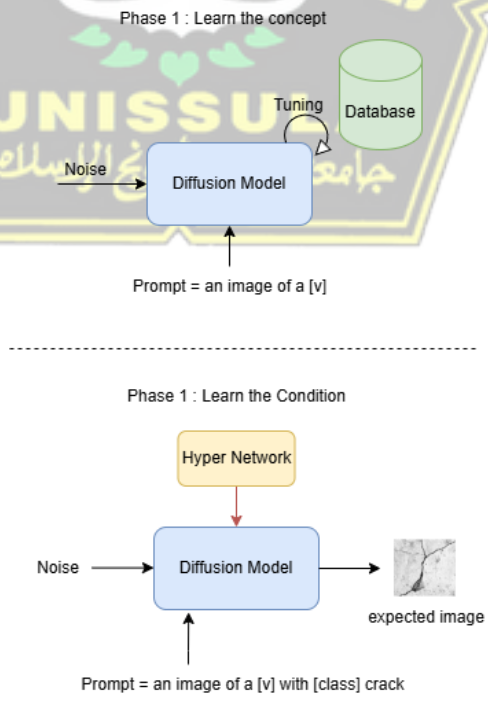
Di mana:

a. $\in$ adalah *noise* asli yang ditambahkan,

b. $\in_\theta (x_t, t)$ adalah prediksi *noise* oleh model.

Dalam kasus *Stable diffusion*, proses ini dilakukan dalam ruang laten (latent space), yang memungkinkan pengolahan data yang lebih efisien. Teks masukan dikodekan menggunakan *text* encoder (seperti CLIP), dan representasi tekstual ini digunakan sebagai kondisi dalam *generative models*. Dengan mekanisme ini, *Stable diffusion* mampu menghasilkan gambar-gambar unik, termasuk lanskap fantasi, berdasarkan deskripsi teks.

Proses ini tidak hanya menghasilkan gambar yang realistis, tetapi juga memberikan kontrol atas hasil akhir, menjadikannya alat yang sangat bermanfaat dalam berbagai aplikasi industri, seperti permainan, film, dan periklanan (Valvano dkk., 2024)

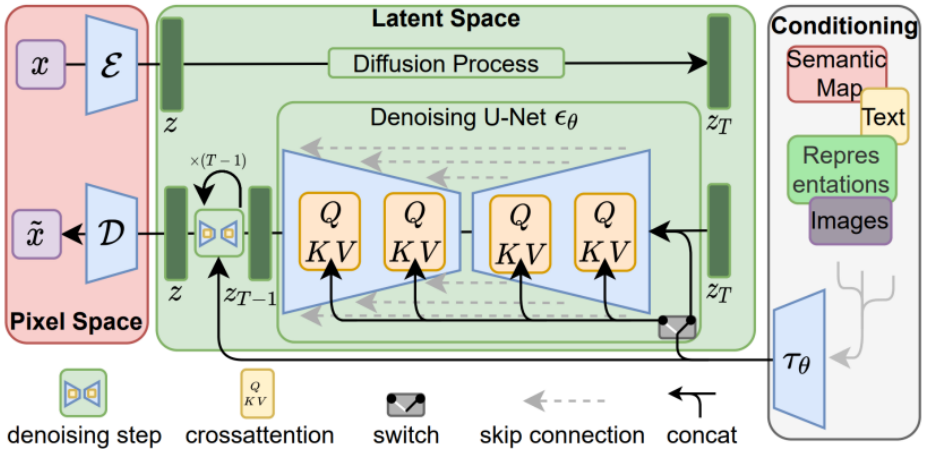



Gambar 2. 1 *Diffusion model Training: From Concept Learning to Condition Application*

### 2.2.4 *Stable diffusion*

*Stable diffusion* merupakan salah satu model generatif yang menggunakan arsitektur berbasis *diffusion* untuk menghasilkan gambar dari deskripsi teks. Model ini bekerja dengan membalikkan proses difusi, yang dimulai dengan gambar *noise* acak dan secara bertahap menyempurnakannya hingga menghasilkan gambar yang sesuai dengan *prompt* yang diberikan. Pada perilisian *stable diffusion* versi 2.0, gambar yang dapat dihasilkan secara default adalah 512x512 pixel dan 768x768 pixel (Stability.ai, 2022). Kemudian Pada versi 2.1, peningkatan akurasi dan fleksibilitas model dalam berbagai domain visual telah dilakukan. *Stable diffusion* v2.1 digunakan karena model ini terbuka (*open-source*) dan dapat diakses, berbeda dengan Midjourney yang bersifat tertutup dan hanya tersedia melalui layanan cloud. Hal ini memungkinkan peneliti untuk melakukan *fine-tuning* model sesuai kebutuhan(Jin dkk., 2023).

Dalam pengembangannya, model generatif *Stable diffusion* bekerja pada ruang laten, memungkinkan peningkatan efisiensi komputasi tanpa mengorbankan kualitas gambar. Model ini menggunakan VQ-GAN sebagai representasi laten yang pertama-tama memproses kode visual dalam tahap awal. Dengan cara ini, *Stable diffusion* tidak hanya mengurangi kompleksitas perhitungan tetapi juga meningkatkan detil visual pada gambar yang dihasilkan, dibandingkan dengan metode pemodelan pada ruang piksel(Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, 2024).




Gambar 2. 2 Arsitektur *Stable diffusion*

Gambar diatas menunjukkan bagaimana *Stable diffusion* bekerja untuk menghasilkan gambar dari teks atau kondisi lainnya. Berikut adalah penjelasan bagian-bagian utamanya :

1. Pixel Space (Ruang Piksel) - (Kiri, Merah Muda)
   a. $\chi$: *Input* berupa *text* maupun gambar.
   b. $\mathcal{E}$ (*encoder*): Mengubah *input* $\chi$ dari ruang piksel menjadi representasi laten $\mathcal{Z}$.
   c. $\tilde{\chi}$ : Gambar hasil rekonstruksi setelah proses difusi selesai.
   d. $\mathcal{D}$ (Decoder): Mengubah kembali representasi laten $\mathcal{Z}$ menjadi gambar dalam ruang piksel $\tilde{\chi}$.
2. Latent Space (Ruang Laten) - (Tengah, Hijau)
   Ruang laten memungkinkan model bekerja lebih efisien dibandingkan langsung di ruang piksel, karena bekerja dalam dimensi yang lebih kecil.
   $\mathcal{Z}_{\mathcal{T}}$ : Representasi laten dengan *noise* maksimal (pada iterasi awal).
   $\mathcal{Z}_{\mathcal{T}} - 1$ : Representasi laten yang secara bertahap mengalami proses *denoising* hingga kembali membentuk gambar yang jelas.
3. *Diffusion Process* (Proses Difusi)
   a. Proses *forward* (menambah *noise*):
   $$q(\mathcal{Z}_{\mathcal{T}} | \mathcal{Z}_{\mathcal{T}} - 1) \quad (4)$$
   Model menambahkan *noise* bertahap pada gambar laten $\mathcal{Z}$ sampai menjadi distribusi acak $\mathcal{Z}_{\mathcal{T}}$.
   b. Proses *reverse* (menghapus *noise*):
   $$\mathcal{P}\theta(\mathcal{Z}_{\mathcal{T}} - 1 | \mathcal{Z}_{\mathcal{T}}) \quad (5)$$
   Model memprediksi dan menghapus *noise* secara bertahap menggunakan *Denoising* U-Net.
4. *Denoising* U-Net (Tengah, Biru)
   a. $\epsilon_\theta$ Model yang digunakan untuk memperkirakan *noise* dalam data laten.
   b. *Cross-attention* (Kotak Kuning dengan $\mathcal{Q}, \mathcal{K}, \mathcal{V}$):

1) $\mathcal{Q}$(*Query*), $\mathcal{K}$(*Key*), $\mathcal{V}$(*Value*) adalah komponen dalam mekanisme atensi (mirip dengan *Transformer*).
   a. *Query*

      Merupakan representasi dari "pertanyaan" yang diajukan untuk mencari informasi relevan dalam data input, untuk mencari hubungan antara elemen-elemen dalam input.

   b. *Key*

      Merupakan representasi dari "kata kunci" atau fitur yang terkait dengan bagian input tertentu, untuk mencocokkan *query* dan menemukan bagian-bagian input yang relevan

   c. *Value*

      Merupakan informasi atau representasi yang dihasilkan dari elemen input yang relevan, yang akan dikirimkan sebagai *output* setelah perhatian (*attentian)* dihitung.

2) Digunakan untuk menghubungkan informasi dari berbagai bagian gambar serta memahami hubungan antara teks dan gambar.

c. *Switch* (Ikon panah berputar hitam-putih):
   Menunjukkan bahwa model dapat menghubungkan dan mengolah ulang informasi secara dinamis.

d. *Skip Connection* (Panah abu-abu putus-putus):
   Berguna untuk mempertahankan informasi dari lapisan awal agar tidak hilang selama proses *denoising*.

e. *Concat* (Panah ke kanan dengan garis lurus):
   Menunjukkan bahwa informasi dari berbagai tahap dipadukan sebelum diproses lebih lanjut.

5. *Conditioning* (Kanan, Berbagai Warna)

   Model dapat dipandu oleh beberapa jenis masukan:

   1. Teks (Kuning): Deskripsi teks yang menentukan isi gambar.
   2. Peta Semantik (Merah): Struktur dalam gambar yang ingin dipertahankan.
   3. Representasi lain (Hijau): Gaya atau ciri khas tertentu.

4. Gambar lain (Ungu): Gambar referensi yang dapat digunakan sebagai dasar.

   $\mathcal{T}_\theta$: Modul yang mengkodekan informasi dari conditioning sebelum dikombinasikan dengan proses difusi.

   Proses Keseluruhan *Stable diffusion* Berdasarkan Gambar

   1. *Input* dalam bentuk teks, gambar, atau kondisi lain diberikan ke model.
   2. *Encoder* $\mathcal{E}$ mengonversi gambar ke ruang laten.
   3. Model menambahkan *noise* secara bertahap hingga gambar benar-benar acak.
   4. *Denoising* U-Net $\in_\theta$ membalikkan proses ini dengan menghapus *noise* bertahap menggunakan *cross-attention.*
   5. *Conditioning* membantu model memahami teks atau gambar referensi agar sesuai dengan keinginan pengguna.
   6. *Decoder* $\mathcal{D}$ mengonversi kembali data laten yang telah bersih menjadi gambar akhir.

Gambar ini memberikan gambaran mendalam tentang bagaimana *Stable Diffusion* mengubah deskripsi teks menjadi gambar berkualitas tinggi dengan pemrosesan yang lebih efisien di ruang laten dibandingkan langsung di ruang piksel. Dalam proses ini, *embedding* berperan penting dalam mengonversi berbagai jenis *input*, seperti teks atau gambar, menjadi representasi numerik yang dapat dipahami oleh model.

Ketika pengguna memberikan prompt teks, misalnya *"a fantasy landscape with mountains and rivers"*, teks tersebut terlebih dahulu diproses oleh *text encoder*, dalam hal ini menggunakan model CLIP (*Contrastive Language-Image Pretraining*). CLIP kemudian mengubah teks menjadi embedding vektor, sebuah representasi angka berdimensi tinggi yang menangkap makna serta konteks dari teks tersebut.

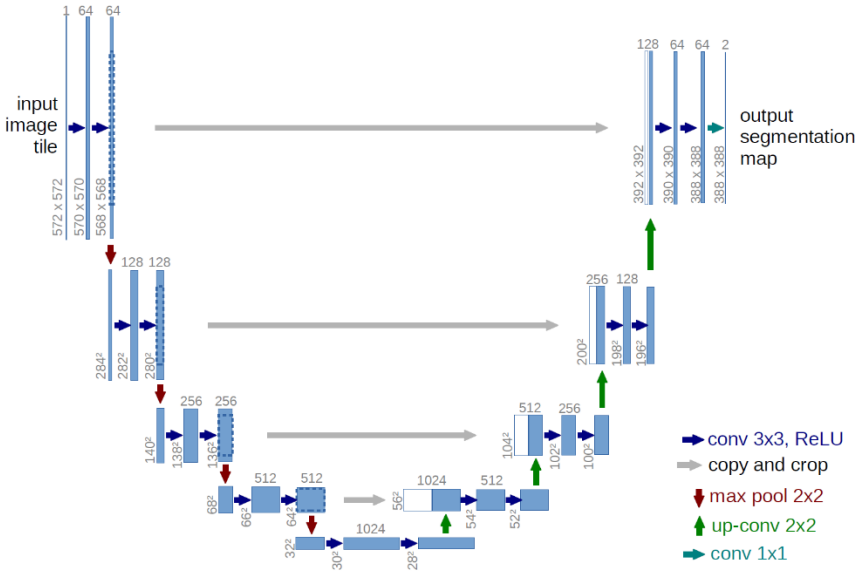
Sementara itu, jika model menerima gambar sebagai *conditioning* (misalnya untuk *image-to-image generation*), gambar tersebut akan diproses melalui *image encoder* yang menggunakan *Variational Autoencoder* (VAE).

*Encoder* ini mengubah gambar dari bentuk piksel menjadi *latent representation* format yang lebih ringkas tetapi tetap mempertahankan informasi esensial tentang bentuk, warna, dan tekstur. *Embedding* ini kemudian digunakan dalam proses *cross-attention* di dalam Denoising U-Net, di mana model menggunakan informasi dari teks atau gambar untuk membimbing proses *denoising*. Dengan kata lain, *embedding* memungkinkan model memahami dan menerjemahkan informasi input ke dalam bentuk yang dapat digunakan untuk menghasilkan gambar yang sesuai dengan kondisi yang diberikan.

### 2.2.5 U-Net

U-Net adalah arsitektur yang banyak digunakan dalam tugas segmentasi gambar karena desain *encoder-decoder*-nya yang memungkinkan ekstraksi dan penggabungan fitur secara efektif melalui *skip connections*. Struktur ini memungkinkan model mempertahankan detail resolusi tinggi yang penting untuk segmentasi yang presisi, terutama dalam aplikasi biomedis di mana akurasi sangat penting. Kerangka kerja *encoder-decoder* dari U-Net memfasilitasi penangkapan konteks pada fase *encoding*, sementara fase decoding menggabungkan konteks ini dengan informasi resolusi tinggi, menghasilkan keluaran segmentasi yang lebih detail (Yin, 2022).

Model ini juga diadaptasi dalam sistem generatif modern seperti *stable diffusion*, yang memanfaatkan proses difusi kondisional untuk menghasilkan gambar lanskap fantasi secara realistis sesuai dengan deskripsi teks, memungkinkan hasil yang mendekati imajinasi pengguna.




Gambar 2. 3 Arsitektur U-Net

Gambar diatas menunjukkan arsitektur U-Net, sebuah model kecerdasan buatan yang digunakan untuk tugas segmentasi gambar, seperti dalam bidang medis dan pemrosesan citra. Arsitektur ini terdiri dari dua bagian utama: *encoder* (bagian kiri) yang berfungsi untuk mengekstrak fitur penting dari gambar, dan *decoder* (bagian kanan) yang bertugas mengembalikan gambar ke bentuk semula dengan mempertahankan detail penting. Pada bagian *encoder*, gambar mengalami proses konvolusi dan penyusutan (*downsampling*) melalui *max pooling*, yang mengurangi ukuran gambar tetapi meningkatkan jumlah fitur yang dipelajari.

Setelah fitur utama diperoleh, bagian *decoder* mulai memperbesar kembali gambar (*upsampling*) menggunakan *up-convolution* sambil menggabungkan informasi dari bagian *encoder* melalui *skip connections*. Ini membantu mempertahankan detail yang mungkin hilang selama penyusutan. Di tahap akhir, lapisan konvolusi 1x1 digunakan untuk menghasilkan peta segmentasi, yang menunjukkan bagian-bagian gambar yang telah diklasifikasikan. Dengan struktur ini, U-Net dapat memberikan hasil segmentasi yang akurat dan sering digunakan dalam berbagai bidang seperti analisis citra medis, pemetaan satelit, dan pengolahan gambar lainnya.

### 2.2.6 *Transformers*

*Transformers* adalah arsitektur yang telah merevolusi bidang pemrosesan bahasa alami dan juga diterapkan dalam model generatif. Mekanisme perhatian (*attention mechanism*) memungkinkan model untuk fokus pada bagian tertentu dari *Input* saat menghasilkan *output*, sehingga menghasilkan representasi yang lebih baik. Dalam konteks *Text-to-image Generation*, *transformers* digunakan untuk memahami hubungan antara teks dan gambar, memungkinkan model untuk menghasilkan gambar yang lebih sesuai dengan deskripsi yang diberikan. *Diffusion Transformers* (DiTs), yang dirancang berdasarkan praktik terbaik *Vision Transformers* (ViTs), memberikan kemampuan skalabilitas yang lebih baik dalam pengenalan visual dibandingkan jaringan konvolusi tradisional, serta menawarkan efisiensi dan ketahanan dalam pengembangan model generatif (Peebles dkk., 2023).

### 2.2.7 *Fine-tuning*

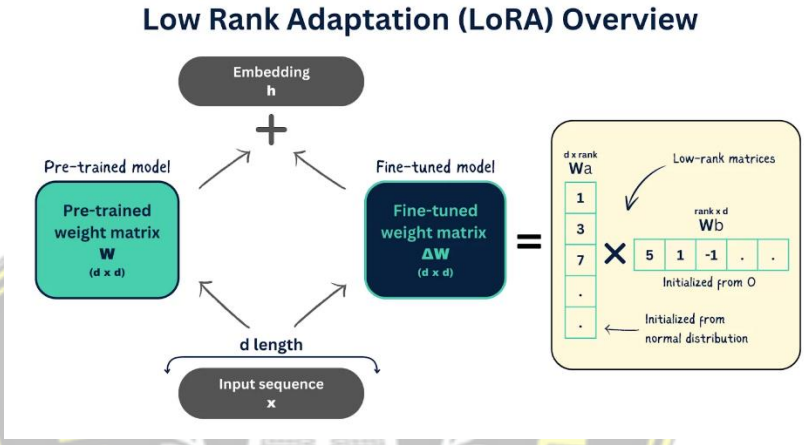
Dalam *Machine Learning, fine-tuning* adalah teknik yang meningkatkan kinerja model dengan menyesuaikan kembali bobot dan bias dari model yang telah dilatih sebelumnya pada tugas baru. Teknik ini biasanya digunakan pada model yang dilatih pada *dataset* besar sebelum digunakan pada tugas pengenalan citra atau teks yang lebih kecil(Pradana dkk, 2024).

Adaptasi model melalui pendekatan *Low-Rank Adaptation* (LoRA) memungkinkan proses *fine-tuning* yang efisien pada model *diffusion* yang telah terlatih, tanpa mengubah bobot utama model. LoRA bekerja dengan mempertahankan bobot asli sambil menambahkan matriks peringkat rendah yang dapat dilatih, sehingga adaptasi model dapat dilakukan dengan lebih hemat memori dan sumber daya komputasi (Luo dkk., 2024). Dalam studi ini, LoRA dipakai untuk menyesuaikan model *diffusion* dengan *dataset* gambar dan teks tertentu untuk menghasilkan konten spesifik yang relevan, tanpa perlu mengubah parameter inti model yang ada.

Pada model *diffusion*, proses *fine-tuning* dengan LoRA dilakukan melalui penambahan modul ke lapisan atensi pada model utama, kemudian dilatih pada *dataset* khusus. Dibandingkan dengan pelatihan penuh, LoRA dapat mempercepat

proses adaptasi dengan hanya mengoptimalkan modul tambahan tersebut, membuatnya lebih efisien dalam pemakaian sumber daya komputasi.

Di beberapa bidang seperti NLP dan *Computer Vision*, pendekatan bertahap digunakan untuk memungkinkan adaptasi model secara perlahan terhadap domain target, memanfaatkan beberapa tahap *fine-tuning* yang disesuaikan dengan data yang relevan. Ini bertujuan untuk membantu model beradaptasi lebih halus dengan distribusi data target.




Gambar 2. 4 *Low-Rank Adaptation* (LoRA)

Gambar ini memberikan gambaran umum tentang *Low-Rank Adaptation* (LoRA) dalam konteks pelatihan model. LoRA adalah teknik inovatif yang digunakan untuk *fine-tuning* model besar tanpa memodifikasi bobot utama model. Pendekatan ini dilakukan dengan menambahkan matriks adaptasi berperingkat rendah (*low-rank matrices*) yang memungkinkan model menangkap informasi baru tanpa mengubah keseluruhan parameter yang besar.

Secara matematis, proses ini dijelaskan melalui rumus:

$$\Delta W = A.B^T \quad (6)$$

Di mana:

a. $\Delta W$: Matriks perubahan kecil yang ditambahkan ke bobot utama. Matriks ini mencerminkan adaptasi baru yang diterapkan pada bagian tertentu dari model.

b. $A$: Matriks dengan dimensi $d \times r$, di mana:

   1) $d$: Dimensi asli dari bobot model (jumlah *Input* atau output).

2) $r$,: Peringkat rendah yang ditentukan (nilai rrr biasanya jauh lebih kecil dari ddd, sehingga mengurangi kompleksitas).

c. $B$: Matriks dengan dimensi $r \times d$. Ketika ditransposisikan ($B^T$), menghasilkan dimensi $d \times r$, yang memungkinkan perkalian matriks dengan $A$ menghasilkan dimensi yang sama dengan $\Delta W$.

Bobot model yang diperbarui setelah penambahan matriks adaptasi ini dituliskan sebagai:

$$W^1 = W + \Delta W \quad (7)$$

Keterangan:

a. $W$**:** Bobot utama dari model besar yang di-*freeze* selama proses *fine-tuning*. Ini menjaga parameter utama tetap tidak berubah untuk efisiensi komputasi.

b. $\Delta W$**:** Matriks perubahan kecil yang ditambahkan untuk memodifikasi kemampuan model pada tugas baru.

c. $W^1$**:** Bobot hasil (*updated weights*) setelah menambahkan perubahan adaptasi ($\Delta W$).

### 2.2.8 CLIP-*based Maximum Mean Discrepancy* (CMMD)

Untuk memvalidasi kualitas gambar yang dihasilkan secara kuantitatif, digunakan metrik CLIP-*based Maximum Mean Discrepancy* (CMMD). Metrik ini lebih andal dibandingkan *Fréchet Inception Distance* (FID) karena tidak mengasumsikan distribusi normal pada *embedding* gambar. CMMD mengukur jarak antara distribusi *embedding* gambar yang dihasilkan oleh model dengan distribusi *embedding* gambar dari *dataset* referensi (Jayasumana dkk., 2023).

Berikut merupakan konsep dari penggunaan CMMD :

1) CLIP *Embedding*: *Embedding* dari gambar diperoleh menggunakan model CLIP (*Contrastive Language-Image Pretraining*), yang telah dilatih pada jutaan pasangan gambar-teks. *Embedding* CLIP lebih kaya dalam merepresentasikan konten visual yang kompleks dan beragam dibandingkan *embedding* dari model Inception-v3 yang digunakan dalam FID.

2) MMD (*Maximum Mean Discrepancy*): CMMD mengukur jarak antara dua distribusi menggunakan kernel *Gaussian Radial Basis Function* (RBF). MMD

memungkinkan pengukuran jarak antara dua distribusi tanpa membuat asumsi distribusi tertentu.

### 2.2.9 *Fréchet Inception Distance* (FID)

FID adalah metrik standar yang digunakan untuk menilai kualitas visual gambar yang dibuat oleh model generatif ini mengukur kesamaan distribusi statistik antara gambar dari data pengujian dan gambar hasil model baru. FID menghitung jarak *Frechet* antara vektor fitur gambar.

FID bergantung pada perbandingan distribusi statistik dari fitur gambar yang diekstraksi dengan menggunakan jaringan *inception v3*. Metrik ini menganggap bahwa fitur gambar mengikuti distribusi *Gaussian multivariat* dan menghitung jarak *Frechet* antara dua distribusi tersebut. Jarak Fréchet dihitung dengan menggunakan mean dan covariance dari vektor fitur yang diperoleh dari kumpulan gambar nyata dan gambar yang dihasilkan model. Nilai FID yang dihasilkan berkorelasi negatif dengan kemiripan distribusi kedua kumpulan gambar.

# BAB III
# METODE PENELITIAN

## 3.1 Metode Penelitian

Penelitian ini menggunakan *stable diffusion* dengan menambahkan *fine-tuning low rank adaptation* dalam prosesnya. *Fine tuning* digunakan untuk melatih data dengan tujuan *output* yang dihasilkan oleh *stable diffusion* sesuai dengan keinginan *user*.

### 3.1.1 Studi Literatur

Studi literatur dilakukan dengan mencari referensi dari jurnal, buku, dan sumber terpercaya mengenai *diffusion models*, khususnya dalam konteks *generative models* dan aplikasi *text-to-image*. Selanjutnya, dilakukan pengkajian terhadap studi-studi literatur yang membahas pemetaan deskripsi teks ke gambar serta model yang digunakan dalam menciptakan lanskap fantasi. Selain itu, teknik *fine-tuning* yang relevan untuk diffusion model dalam lanskap fantasi diidentifikasi, termasuk eksplorasi bagaimana pendekatan tersebut dapat meningkatkan kualitas hasil generatif. Terakhir, penelitian mengenai *Low-Rank Adaptation* (LoRA) pada model *Machine Learning* ditinjau, terutama dalam konteks visualisasi lanskap untuk memahami perannya dalam optimasi model generatif.

### 3.1.2 Analisa Kebutuhan

**A. Kebutuhan Fungsional**

1. Input Deskripsi Teks

   Sistem harus menyediakan antarmuka untuk pengguna memasukkan deskripsi teks sebagai *input.*

2. Proses Fine-Tuning dengan LoRA

   Sistem harus dapat melakukan fine-tuning pada model *Stable Diffusion v2.1* menggunakan dataset yang telah disiapkan.

3. Generasi Gambar Lanskap Fantasi

   Sistem harus menghasilkan gambar lanskap fantasi berdasarkan input teks yang diberikan pengguna.

4. Parameter Kustomisasi

   Sistem harus memungkinkan pengguna menyesuaikan parameter seperti jumlah langkah *sampling, guidance scale*, dan resolusi gambar.

5. Penyimpanan dan Manajemen Hasil

   Sistem harus menyediakan fitur penyimpanan dan pengelolaan hasil gambar yang telah dihasilkan.

6. Antarmuka Pengguna Berbasis Web

   Sistem harus memiliki antarmuka berbasis web menggunakan Streamlit untuk interaksi pengguna.

7. Ekspor dan Unduh Hasil Gambar

   Pengguna harus dapat mengunduh hasil gambar dalam format PNG atau JPG.

8. Logging dan Monitoring Kinerja Model

   Sistem harus mencatat log penggunaan dan kinerja model untuk analisis dan *debugging*.

**B. Kebutuhan Non-Fungsional**

1. Performa

   Sistem harus mampu menghasilkan gambar dalam waktu yang wajar (kurang dari 1 menit per gambar dengan T4 GPU).

2. Keamanan

   Sistem harus membatasi akses model dan menyaring input untuk mencegah eksploitasi.

3. Scalability

   Sistem harus dapat dikembangkan lebih lanjut untuk mendukung model lebih besar seperti SDXL.

4. Ketersediaan

   Sistem harus memiliki uptime tinggi dengan infrastruktur berbasis cloud jika diperlukan.

5. Kemudahan Penggunaan

   Antarmuka harus intuitif dan mudah digunakan oleh pengguna non-teknis.

**C. Kebutuhan Perangkat**

1. Perangkat Lunak
   1) Python 3.10+
   2) Google Colab dengan GPU T4
   3) Stable Diffusion v2.1
   4) LoRA fine-tuning library (PEFT, Hugging Face Diffusers)
   5) Streamlit untuk antarmuka web
   6) CUDA dan PyTorch untuk akselerasi GPU
2. Perangkat Keras
   1) GPU NVIDIA T4 atau lebih tinggi
   2) RAM minimal 16GB
   3) Penyimpanan SSD minimal 20GB untuk model dan dataset

### 3.1.3 Pengumpulan dan *PreProcessing Dataset*

Pada tahap pengumpulan data, *dataset* lanskap fantasi diperoleh dari berbagai sumber, seperti repositori *online* dengan lisensi terbuka, hasil karya komunitas seni digital, dan *dataset* yang dihasilkan menggunakan generator berbasis AI. Gambar-gambar tersebut beragam dalam hal format dan ukuran. Selanjutnya, dilakukan *captioning* untuk memberikan deskripsi teks pada masing-masing gambar, yang menghasilkan pasangan gambar dan teks secara berurutan. Proses ini penting dalam menyediakan data yang dapat digunakan untuk pelatihan model yang mempelajari hubungan antara gambar dan teks.

Setelah gambar dan *caption* disiapkan, *dataset* dibagi menjadi dua bagian utama yaitu 80% untuk data pelatihan dan 20% untuk data pengujian. *Dataset* pelatihan digunakan untuk melatih model dalam mengenali pola antara gambar dan

*caption*, sementara *dataset* pengujian digunakan untuk mengukur kinerja model dalam memprediksi *caption* berdasarkan gambar baru yang belum dilihat sebelumnya

Data yang terkumpul kemudian melalui tahap *preProcessing* untuk memastikan kualitas dan konsistensi gambar sebelum digunakan dalam pelatihan model. Pada tahap ini dilakukan *preProcessing* data pelatihan yang mencakup data *augmentation* dan *standarization* untuk meningkatkan variasi dan memastikan skala data seragam. Langkah-langkahnya adalah *Resize* : mengubah ukuran gambar ke resolusi tertentu agar seragam

1) *Random Resized Crop* : memotong gambar secara acak untuk variasi *framing* objek
2) *Random Horizontal Flip* : membalik gambar secara horizontal dengan probabilitas 50%
3) *Color Jitter* : mengubah kecerahan, kontras, dan saturasi gambar secara acak

### 3.1.4 Pelatihan Model

Pelatihan model dalam penelitian ini dilakukan melalui beberapa tahapan penting untuk memastikan hasil yang optimal. Tahap pertama adalah pemilihan model, di mana *Stable Diffusion v2.1* dipilih sebagai *diffusion model* yang akan digunakan. Model ini dipilih karena memiliki kemampuan yang unggul dalam menghasilkan gambar dengan kualitas tinggi serta fleksibilitas dalam proses *fine-tuning*. Dalam penelitian ini, pendekatan yang digunakan untuk *fine-tuning* adalah *Low-Rank Adaptation* (LoRA), yang memungkinkan penyesuaian model dengan lebih efisien tanpa memerlukan pelatihan ulang seluruh *parameter* model.

Setelah model dipilih, tahap berikutnya adalah melakukan *fine-tuning* pada *Stable Diffusion v2.1* menggunakan *dataset* yang telah melalui tahap *preprocessing*. *Fine-tuning* dilakukan dengan teknik LoRA agar model dapat menyesuaikan diri dengan karakteristik data yang lebih spesifik, sehingga mampu menghasilkan gambar fantasi yang lebih akurat dan sesuai dengan kebutuhan penelitian. Proses ini melibatkan pelatihan model dengan dataset yang telah disiapkan, memastikan

bahwa model belajar pola-pola yang relevan untuk menghasilkan gambar dengan kualitas yang diharapkan.

Selama proses pelatihan berlangsung, digunakan *framework TensorBoard* untuk memantau kinerja model. *TensorBoard* memungkinkan visualisasi berbagai metrik pelatihan, seperti nilai *loss* dan *accuracy*, sehingga perkembangan model dapat diamati secara real-time. Dengan pemantauan ini, dapat dipastikan bahwa model mengalami peningkatan kualitas selama proses *fine-tuning* serta menghindari permasalahan seperti *overfitting* atau *underfitting*. Pemantauan ini juga membantu dalam menentukan kapan pelatihan model dapat dihentikan secara optimal untuk mendapatkan hasil yang terbaik.

### 3.1.5 Pengujian dan Evaluasi Model

1. Pengujian

   Pada tahap ini dilakukan pengujian untuk membuat gambar dari model yang sudah dilatih, pengujian ini menggunakan kumpulan teks dari gambar pada data pengujian yang belum pernah dilihat model selama proses pelatihan, sehinga menghasilkan jumlah gambar yang sama dengan data pengujian menggunakan teks yang sama.

2. Evaluasi

   Metode Evaluasi yang digunakan pada tahap ini menggunakan kombinasi metrik kuantitatif dan penilaian kualitatif.

   1) Metrik Kuantitatif

      a) CLIP-MMD (*CLIP Mean Maximum Discrepancy*)

         Memanfaatkan model CLIP (*Contrastive Language-Image Pretraining*), CLIP-MMD adalah metrik yang dirancang untuk mengevaluasi kesesuaian semantik antara teks *Input* dan gambar yang dihasilkan. CMMD mengukur jarak antara distribusi *embedding* gambar yang dihasilkan oleh model dengan distribusi *embedding* gambar dari *dataset* referensi (Jayasumana dkk., 2023).

         Misalkan $x = \{x_1, x_2, \ldots\ldots, x_m\}$ adalah *embedding* CLIP dari gambar yang dihasilkan model, dan $Y = \{y_1, y_2, \ldots\ldots, y_n\}$ adalah *embedding* CLIP dari gambars referensi. Jarak CMMD dihitung sebagai:

$$CMMD(X,Y) = \frac{1}{m(m-1)} \sum_{i \neq j} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k\,(y_i, y_j) - \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} k(y_i, y_j) \quad (8)$$

Dimana :

b) $k(x, y)$ adalah *Gaussian RBF kernel* yang didefinisikan sebagai :

$$k\,(x, y) = \exp\left(-\frac{||x-y||^2}{2\sigma^2}\right) \quad (9)$$

c) m dan n adalah jumlah *embedding* dalam masing-masing himpunan gambar dari model dan *dataset* referensi.

d) $\sigma$ adalah parameter bandwith kerner RBF.

c. Interpretasi Nilai CMMD

1) Nilai lebih rendah menunjukkan bahwa distribusi *embedding* dari gambar yang dihasilkan model lebih dekat ke distribusi *embedding* dari *dataset* referensi, yang berarti kualitas gambar lebih baik.
2) CMMD lebih stabil dan lebih konsisten dengan persepsi manusia dibandingkan FID.

Kemudian melakukan perbandingan Sebelum dan Sesudah *Fine-tuning* .Proses ini bertujuan untuk mengevaluasi efek dari *fine-tuning* LoRA (*Low-Rank Adaptation*) pada kualitas gambar yang dihasilkan. Perbandingan ini dilakukan dengan langkah berikut:

a. Penggunaan CMMD: Nilai CMMD dihitung sebelum dan sesudah proses *fine-tuning* LoRA. Nilai CMMD sebelum *fine-tuning* digunakan sebagai baseline, sedangkan nilai setelah *fine-tuning* digunakan untuk mengevaluasi perbaikan kualitas.
b. Visualisasi Perbandingan: Gambar dari model sebelum dan sesudah *fine-tuning* ditampilkan secara berdampingan agar perbedaannya dapat diamati secara visual. Perbaikan dapat diamati dari detail visual, kompleksitas lanskap, dan kekayaan elemen fantasi.

b) FID (*Fréchet Inception Distance*)

FID adalah metrik standar yang digunakan untuk menilai kualitas visual gambar yang dibuat oleh model generatif ini mengukur kesamaan distribusi statistik antara gambar dari data pengujian dan gambar hasil model baru. FID menghitung jarak *Frechet* antara vektor fitur gambar. Ini diperoleh dengan menggunakan model yang diatur, biasanya *inception-v3*. Nilai FID yang lebih rendah menunjukan bahwa distribusi gambar yang dihasilkan dari model lebih mirip dengan distribusi gambar aslinya. Berikut merupakan rumus dari FID :

$$FID = \| u_r - u_g \|^2 + \mathrm{Tr}(\sum_r + \sum_g - 2\,(\sum_r \sum_g)\,\tfrac{1}{2} \qquad (10)$$

1. Penjelasan rumus :

   $u_r$: Vektor rata-rata fitur gambar nyata (real).

   $u_g$: Vektor rata-rata fitur gambar yang dihasilkan (generated).

   $\sum_r$: Kovarians matriks dari distribusi fitur gambar nyata.

   $\sum_g$: Kovarians matriks dari distribusi fitur gambar yang dihasilkan.

   $\| u_r - u_g \|^2$: Jarak Euclidean kuadrat antara rata-rata fitur gambar nyata dan gambar yang dihasilkan.

   Tr(·): Operator jejak matriks, yang menjumlahkan elemen diagonal dari matriks.

   $(\sum_r \sum_g)\,\tfrac{1}{2}$: Akar matriks dari hasil perkalian kovarians gambar nyata dan gambar yang dihasilkan.

2. Cara Kerja Evaluasi FID dalam *Stable diffusion*:

   a. Ekstraksi Fitur dengan *Inception Network*

      1) Gambar nyata dan gambar hasil generasi dianalisis menggunakan model InceptionV3, terutama dari layer sebelum klasifikasi.

      2) Fitur yang diekstrak berupa vektor dari distribusi *Gaussian multivariat*.

   b. Perhitungan Statistik Distribusi

      1) Hitung mean $\upsilon$ dan kovarians $\sum$ dari fitur gambar nyata dan gambar yang dihasilkan.

   c. Menghitung Jarak *Fréchet*

1) Hitung jarak *Euclidean* kuadrat antara rata-rata distribusi fitur.
2) Hitung perbedaan kovarians menggunakan jejak matriks dan akar matriks.

d. Interpretasi Nilai FID

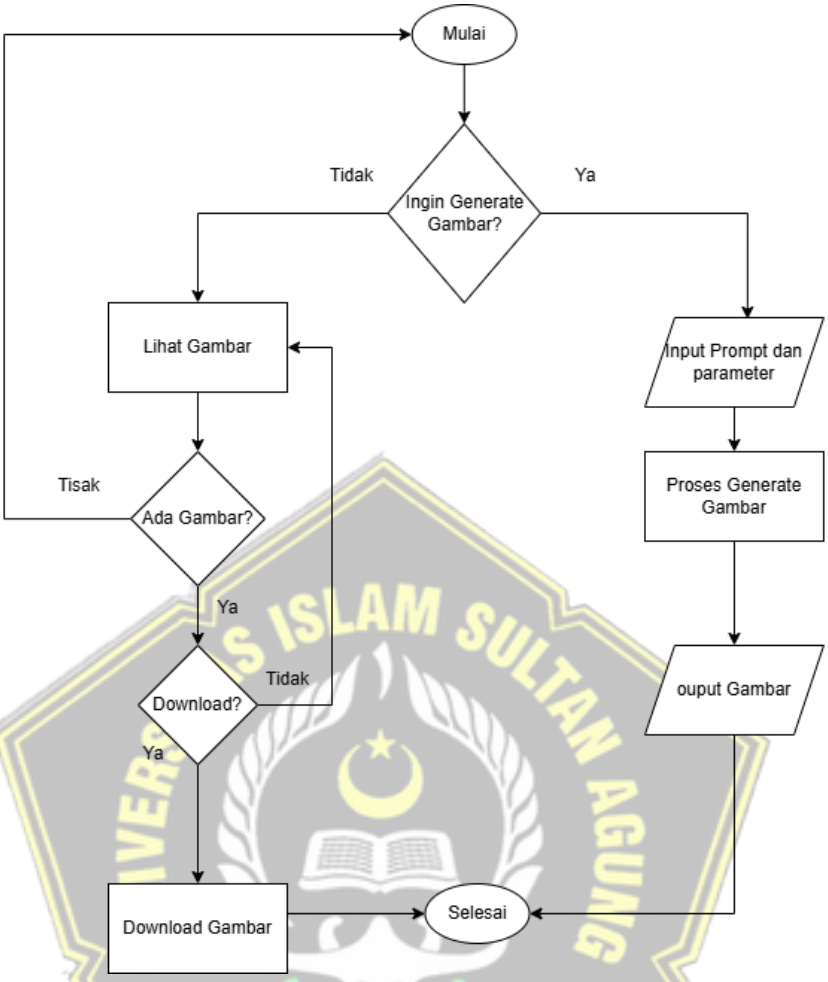
1) Semakin kecil nilai FID, semakin mirip distribusi gambar hasil generasi dengan gambar nyata, yang berarti kualitas gambar lebih baik.
2) Semakin besar nilai FID, semakin buruk kualitas gambar yang dihasilkan.

2) Metrik Kualitatif

Hasil gambar dari data pengujian dan hasil gambar dari model yang sudah dilatih akan dibandingkan menggunakan visual mata telanjang untuk membandingkan kualitas visual secara langsung. Perbandingan ini melibatkan analisis kualitas gambar dari segi detail, ketajaman, tekstur, dan keakuratan representasi visual yang sesuai dengan deskripsi teks yang diberikan.

### 3.1.6 Pembuatan Sistem

Gambar dibawah ini merupakan visualisasi *flowchart* sistem yang akan dibuat :




Gambar 3. 1 *Flowchart* Penggunaan Sistem

3. Pemilihan Opsi Generate Gambar

   Pengguna diberikan pilihan untuk melakukan proses pembuatan gambar baru atau melihat hasil gambar yang sudah tersedia.

   a. Jika pengguna memilih untuk menghasilkan gambar baru, sistem akan melanjutkan ke tahap *Input Prompt* dan *Parameter*.
   b. Jika pengguna memilih untuk melihat gambar yang sudah ada, sistem akan menampilkan gambar yang tersedia (jika ada).

2. *Input Prompt dan Parameter*

   Pada tahap ini, pengguna memasukkan deskripsi teks (prompt) yang menggambarkan lanskap fantasi yang ingin dibuat.

   a. Prompt harus berisi informasi yang cukup jelas agar sistem dapat menghasilkan gambar sesuai dengan harapan pengguna.

b. Contoh: "A golden forest with red mushroom."
c. Selain prompt, pengguna juga dapat menentukan parameter tambahan seperti *seed*, jumlah *inference step*, jumlah gambar, ukuran gambar, dan *guidance scale*.

3. Proses Menghasilkan Gambar

   Sistem akan memproses input pengguna dengan menggunakan model yang telah dilatih untuk menghasilkan gambar berdasarkan prompt yang diberikan.

   a. Model yang digunakan telah diadaptasi dengan LoRA (Low-Rank Adaptation) dan dijalankan menggunakan *Stable diffusion*.
   b. Setelah proses selesai, sistem akan menghasilkan gambar yang sesuai dengan deskripsi yang diberikan oleh pengguna.

4. *Output* Gambar

   Gambar yang telah dihasilkan akan disimpan dalam sistem dan siap untuk ditampilkan kepada pengguna.

5. Pemeriksaan Ketersediaan Gambar

   Jika pengguna memilih untuk melihat gambar yang telah tersedia, sistem akan melakukan pengecekan:

   a. Jika gambar tersedia, maka gambar akan ditampilkan kepada pengguna.
   b. Jika tidak ada gambar yang tersedia, pengguna akan diberikan opsi untuk kembali ke tahap awal dan melakukan proses generate gambar baru.

6. Unduhan Gambar

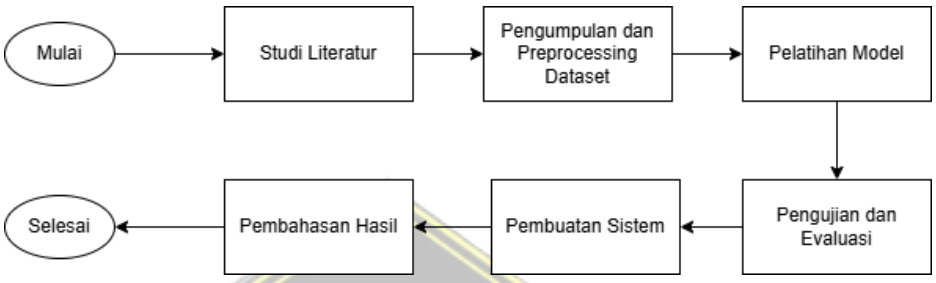
   Setelah melihat hasil gambar, pengguna diberikan opsi untuk mengunduh gambar tersebut.

   a. Jika pengguna memilih mengunduh gambar, sistem akan menyiapkan file gambar untuk diunduh.
   b. Jika pengguna memilih tidak mengunduh gambar, proses akan langsung berakhir.

7. Penyelesaian Proses

   Setelah pengguna mengunduh gambar atau memilih untuk tidak melanjutkan, proses selesai.

   Berikut merupakan visualisasi *Flowchart* metode penelitian yang penulis lakukan :




Gambar 3. 2 *Flowchart* Metode Penelitian

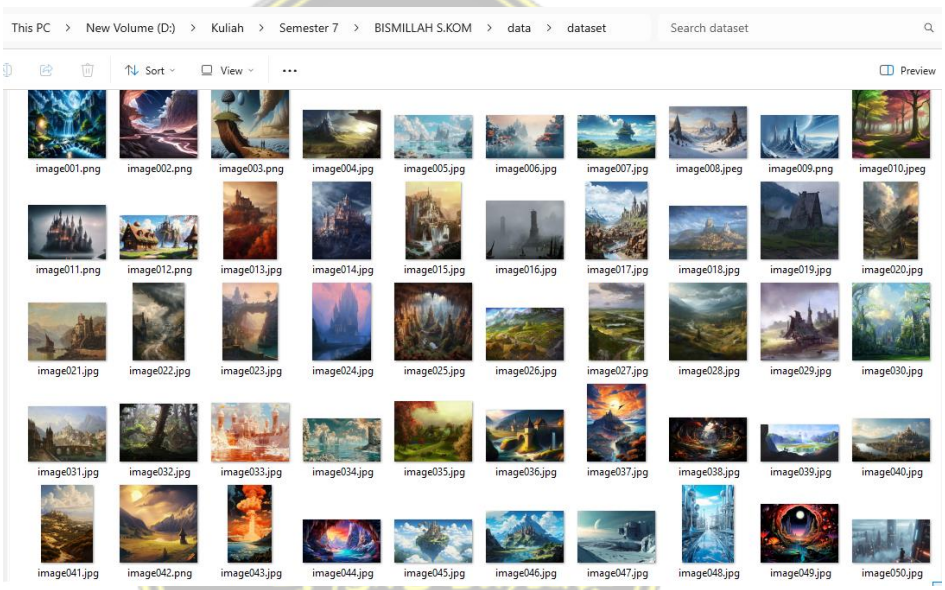
# BAB IV
# HASIL DAN ANALISIS PENELITIAN

## 4.1 Hasil *PreProcessing Dataset*

*Dataset* yang digunakan dalam penelitian ini telah melalui beberapa tahapan penting selama tahap *preProcessing*, seperti :

1. Pengumpulan *Dataset*

   *Dataset* ini dikumpulkan dari berbagai sumber yang relevan untuk mendukung penelitian ini. *Dataset* ini berupa gambar-gambar lanskap fantasi yang akan digunakan untuk pelatihan *stable diffusion* 2.1




Gambar 4. 1 *Dataset*

2. *Captioning* Gambar

   *Caption*ing dilakukan secara manual dan otomatis menggunakan model *Natural Language Processing (NLP)* untuk memastikan bahwa hubungan semantik antara teks dan gambar konsisten.

Tabel 4. 1 *Captioning Dataset*

| No | *image* | *Captions* |
|---|---|---|
| 1 | /content/drive/MyDrive/*dataset*_result/image01.png | fantasy landscape with waterfall and moon The landscape features lush green forests, tall mountains, and a vibrant sky. Towering mountains rise high, their peaks covered in snow and surrounded by mist. A mystical mist blankets the valley, adding an air of mystery to the enchanted scene |
| 2 | /content/drive/MyDrive/*dataset*_result/image002.png | a landscape with a mountain and a river The landscape features lush green forests, tall mountains, and a vibrant sky. Towering mountains rise high, their peaks covered in snow and surrounded by mist. A mystical mist blankets the valley, adding an air of mystery to the enchanted scene. |
| 3 | /content/drive/MyDrive/*dataset*_result/image03.png | a surreal painting of a man standing on a cliff with a tree and a cloud filled with planets |
| …. | ….. | ……. |
| 50 | /content/drive/MyDrive/*dataset*_result/image050.png | a man standing on a ledge looking at a city |

Captioning dalam penelitian ini dilakukan dengan kombinasi metode otomatis dan manual untuk memastikan hubungan semantik antara teks dan gambar tetap konsisten. Model BLIP (*Bootstrapping Language-Image Pretraining*) dari *Salesforce* digunakan untuk menghasilkan deskripsi awal berdasarkan konten visual gambar, di mana *BlipProcessor* memproses gambar menjadi representasi tensor, lalu *BlipForConditionalGeneration* menghasilkan

teks deskriptif. Setelah caption awal dibuat, dilakukan pascapemrosesan dengan menambahkan detail tambahan berdasarkan kata kunci dalam deskripsi, misalnya jika terdapat kata "*landscape*," maka teks diperpanjang dengan informasi seperti hutan hijau yang lebat, pegunungan tinggi, atau langit yang cerah, begitu pula untuk elemen seperti kastil, laut, pantai, desa, gunung, matahari terbenam, dan kabut. Dengan pendekatan ini, *caption* yang dihasilkan tidak hanya lebih akurat tetapi juga lebih kaya deskripsi, memperkuat hubungan semantik antara teks dan gambar sehingga memberikan anotasi yang lebih informatif dan kontekstual untuk berbagai keperluan, termasuk pelatihan model lebih lanjut atau menghasilkan konten berbasis AI

3. Pembagian *Dataset*

   *Dataset* yang sudah dilakukan *captioning* selanjutnya dibagi menjadi dua bagian utama, yaitu :

   a. Data pelatihan (*Training Dataset*)

      Data ini digunakan untuk melatih model agar dapat menghasilkan gambar yang sesuai dengan teks *Input*. *Dataset* ini berjumlah 80% dari pembagian *dataset* utama atau sejumlah 40 *Dataset*.

Tabel 4. 2 *Sample Dataset Training*

| No | *image* | *Captions* |
|---|---|---|
| 1 | /content/drive/MyDrive/*dataset*_result/image013.png | a castle in the middle of a forest with a river and mountains in the background A majestic castle stands tall amidst the mist, its towers reaching for the clouds. A mystical mist blankets the valley, adding an air of mystery to the enchanted scene. |
| 2 | /content/drive/MyDrive/*dataset*_result/image005.png | a beautiful view of a rocky beach with a few birds flying overhead The beach is bathed in golden sunlight, with gentle waves lapping at the sandy shore. |
| 3 | /content/drive/MyDrive/*dataset*_result/image038.png | a stream in a forest with trees and flowers |
| …. | ….. | ……. |
| 40 | /content/drive/MyDrive/*dataset*_result/image039.png | a man standing on a cliff looking out at a river |

Gambar diatas merupakan contoh dari *dataset* yang sudah dibagi menjadi data pelatihan sejumlah 40 data pasangan gambar dalam bentuk url dan deskripsinya pada kolom *image* dan *captions*.

b. Data pengujian (*Testing Dataset*)

Data ini digunakan untuk mengevaluasi performa model setelah proses pelatihan. *Dataset* ini berjumlah 20% dari pembagian *dataset* utama atau sejumlah 10 *dataset*.

Tabel 4. 3 *Sample Dataset testing*

| No | *image* | *captions* |
|---|---|---|
| 1 | /content/drive/My Drive/*dataset*_result/image014.png | a castle in the snow A majestic castle stands tall amidst the mist, its towers reaching for the clouds. A mystical mist blankets the valley, adding an air of mystery to the enchanted scene. |
| 2 | /content/drive/My Drive/*dataset*_result/image040.png | a castle on a mountain A majestic castle stands tall amidst the mist, its towers reaching for the clouds. A mystical mist blankets the valley, adding an air of mystery to the enchanted scene. |
| 3 | /content/drive/My Drive/*dataset*_result/image031.png | a view of a town with a bridge and a river |
| … | …. | … |
| 10 | /content/drive/My Drive/*dataset*_result/image020.png | a mountain scene with a lake and a mountain range in the background Towering mountains rise high, their peaks covered in snow and surrounded by mist. A mystical mist blankets the valley, adding an air of mystery to the enchanted scene. |

Gambar diatas merupakan contoh dari *dataset* yang sudah dibagi menjadi data pelatihan sejumlah 10 data pasangan gambar dalam bentuk url dan deskripsinya pada kolom *image* dan *captions*.

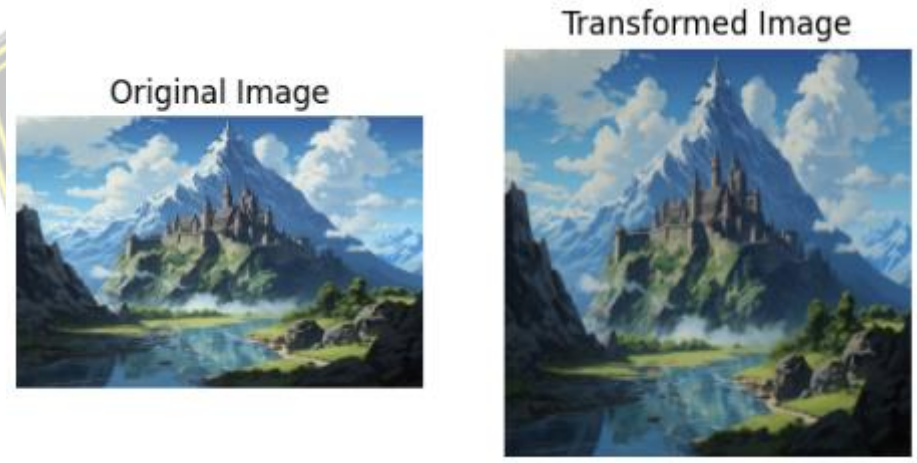
Tabel dibawah ini menunjukkan jumlah data yang digunakan pada penelitian ini. Terdiri dari *dataset* pelatihan dan *dataset* pengujian

Tabel 4. 4 Pembagian *Dataset*

| No. | ***Dataset*** | **Jumlah gambar** | **Jumlah *caption*** |
|---|---|---|---|
| 1. | *Dataset* Pelatihan | 40 | 40 |
| 2. | *Dataset* Pengujian | 10 | 10 |
| | **Total** | **50** | **50** |

c. *PreProcessing Dataset*

Pada proses *preProcessing*, gambar di augmentasi untuk menambah variasi *dataset* dengan berbagai proses seperti *random flip*, *resize*, dan *color jitter*. Seperti yang ditunjukkan pada gambar dibawah ini, gambar asli yang sudah di *preProcessing* dirubah menjadi gambar dengan ukuran 768 x 768, *flip horizontal*, dan sedikit menjadi lebih gelap




Gambar 4. 2 Hasil *PreProcessing*

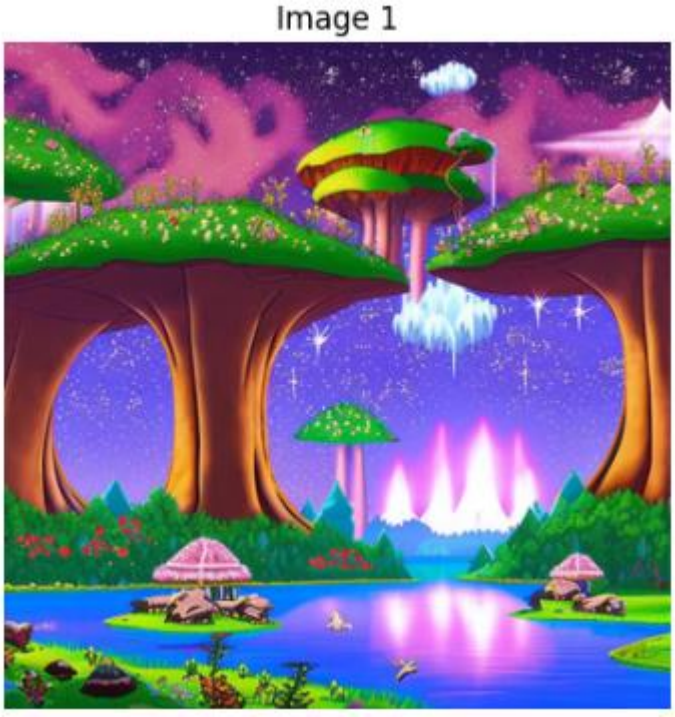
## 4.2 *Stable diffusion*

Pada *Stable diffusion* v2.1 dilakukan pengujian terkait pembuatan lanskap fantasi, dimana model ini merupakan model yang belum dilatih dan digunakan untuk menghasilkan beberapa contoh gambar dengan berbagai parameter yang berbeda seperti *prompt*, *seed*, dan *guidance scale.*

Tabel 4. 5 *Parameter Generate* Gambar 1 *Stable diffusion v2.1*

| | |
|---|---|
| *Prompt* | *Create a mystical fantasy landscape at the edge of an enchanted forest. Towering bioluminescent trees glow softly, while a crystal-clear river sparkles with magical energy. Giant mushrooms, glowing flowers, and* |

| | |
|---|---|
| | *luminous creatures populate the land. In the distance, towering mountains rise under a starry sky, with floating islands above. The air is misty, giving the scene a dreamy, otherworldly feel.* |
| Negative *prompt* | *cropped, cut-off, close-up, blurry, out of frame, bad composition, missing parts, deformed, distorted, unclear details, bad framing, oversaturated, unnatural lighting, unrealistic proportions* |
| *Seed* | 777 |
| *Guidance Scale* | 7.5 |
| *Image Size* | 768x512 |
| *Num Images* | 6 |

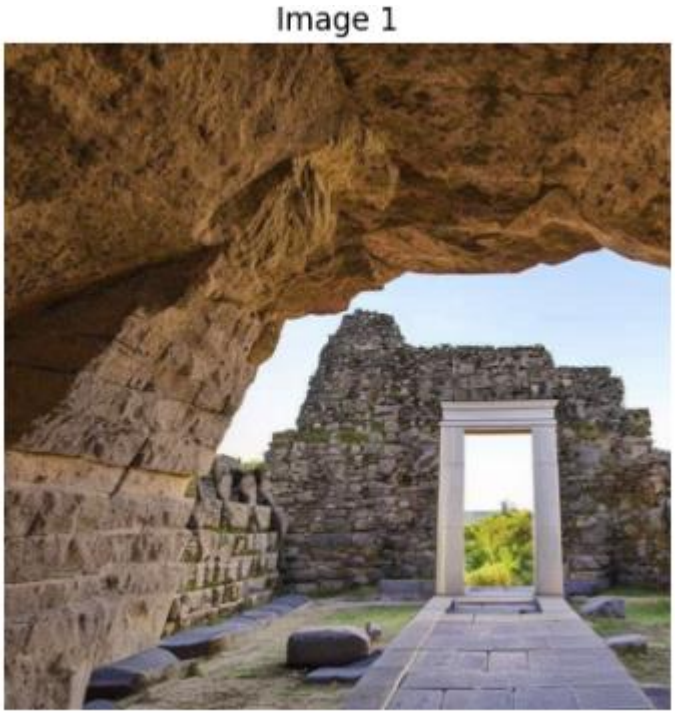



Gambar 4. 3 Hasil *Generate* Gambar 1 *Stable diffusion v2.1*

Tabel 4. 6 *Parameter Generate* Gambar 2 *Stable diffusion* v2.1

| | |
|---|---|
| *Prompt* | *Ancient ruins with hidden treasures* |
| Negative *prompt* | *cropped, cut-off, close-up, blurry, out of frame, bad composition, missing parts, deformed, distorted, unclear details, bad framing, oversaturated, unnatural lighting, unrealistic proportions* |

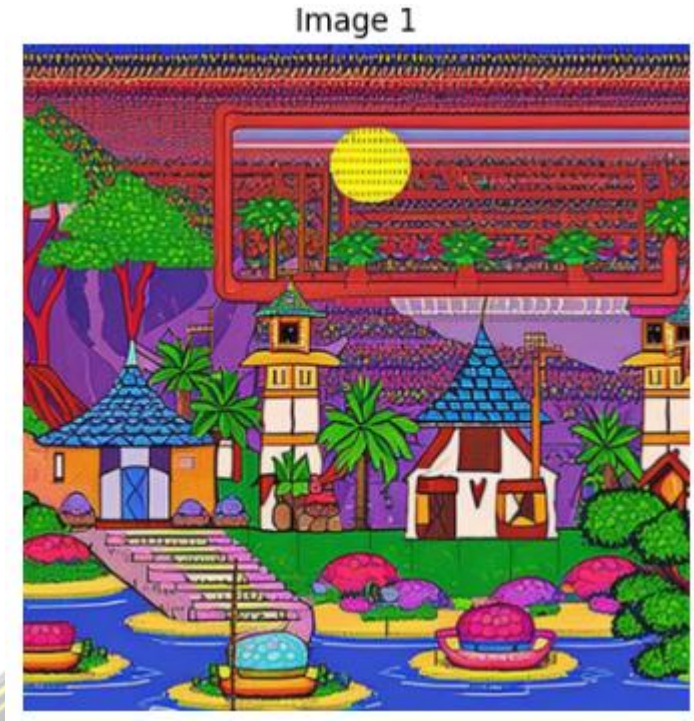
| | |
|---|---|
| *Seed* | 42 |
| *Guidance Scale* | 7.5 |
| *Image Size* | 768x512 |
| *Num Images* | 1 |




Gambar 4. 4 Hasil *Generate* Gambar 1 *Stable diffusion v2.1*

Tabel 4. 7 *Parameter Generate* Gambar 1 *Stable diffusion v2.1*

| | |
|---|---|
| *Prompt* | *a village with exotic goods, in the style of vibrant fantasy.* |
| Negative *prompt* | *cropped, cut-off, close-up, blurry, out of frame, bad composition, missing parts, deformed, distorted, unclear details, bad framing, oversaturated, unnatural lighting, unrealistic proportions* |
| *Seed* | 1332 |
| *Guidance Scale* | 7.5 |
| *Image Size* | 768x768 |
| *Num Images* | 1 |




Gambar 4. 5 Hasil *Generate* Gambar 1 *Stable diffusion v2.1*

## 4.3 Hasil Pelatihan Model

Pada tahap ini, model *Stable diffusion* 2.1 telah di*fine-tuning* dengan teknik LoRA. *Fine-tuning* ini dilakukan dengan tujuan meningkatkan kemampuan model untuk menghasilkan gambar yang lebih akurat yang sesuai dengan teks deskripsi yang diberikan. Beberapa komponen utama yang terlibat dalam proses pelatihan adalah:

1. *Fine-Tuning*
   a. Parameter *Fine-Tuning*

      Dalam penelitian ini, model *Stable diffusion v2.1* dari *StabilityAI* digunakan sebagai model dasar untuk menghasilkan lanskap fantasi. Model ini di*fine-tune* menggunakan *Low-Rank Adaptation* (LoRA). Proses pelatihan dilakukan dengan menggunakan skrip *training*, yang dieksekusi menggunakan *Accelerate* dari *Hugging Face* untuk distribusi efisien pada perangkat keras yang tersedia.

      Perintah berikut digunakan untuk menjalankan pelatihan:

```
!accelerate launch /content/drive/MyDrive/scripts/train_text_to_image_lora_csv.py \
  --pretrained_model_name_or_path="stabilityai/stable-diffusion-2-1" \
  --train_data_dir="/content/drive/MyDrive/train_dataset/train_dataset.csv" \
  --image_column="image" \
  --caption_column="captions" \
  --dataloader_num_workers=8 \
  --resolution=768 \
  --random_flip \
  --train_batch_size=1 \
  --gradient_accumulation_steps=4 \
  --max_train_steps=400 \
  --learning_rate=2e-5 \
  --max_grad_norm=1 \
  --lr_scheduler="cosine_with_restarts" \
  --lr_warmup_steps=100 \
  --output_dir="/content/drive/MyDrive/lora_outputs" \
  --checkpointing_steps=100 \
  --validation_prompt="Create a mystical fantasy landscape at the edge of an enchanted forest.
  --seed=42 \
  --rank=16 \
  --report_to tensorboard \
  --mixed_precision fp16 \
```

Gambar 4. 6 Parameter *Fine-Tuning*

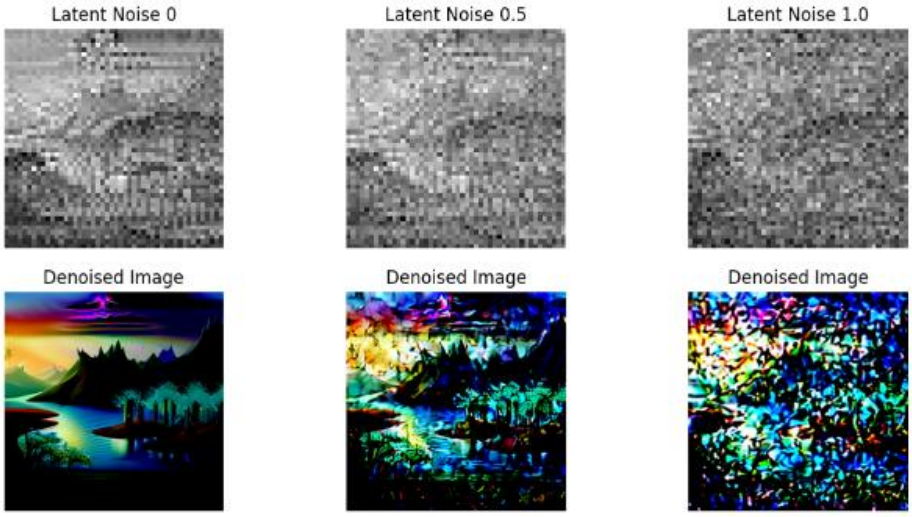
Beberapa parameter kunci dalam pelatihan ini adalah:

1) *Pretrained_model* : Model yang digunakan adalah *stabilityai/stable-diffusion-2-1*, yaitu model *pre-trained* yang telah dilatih pada *dataset* besar dan berfungsi sebagai dasar untuk proses *fine-tuning*, sehingga dapat disesuaikan dengan *dataset* spesifik yang digunakan dalam penelitian ini.
2) *Train_data_dir* : Data pelatihan disimpan dalam format .csv dengan dua kolom utama, yaitu *image* untuk *path* gambar dan *captions* untuk deskripsi gambar, yang memungkinkan model memahami hubungan antara teks dan gambar dengan lebih efektif.
3) *Dataloader_num_worker* : Menentukan jumlah *worker* untuk memproses data saat *training*, yang berpengaruh pada kecepatan pemrosesan data sebelum masuk ke model.
4) *Resolution* : Semua gambar diubah menjadi ukuran 768×768 piksel untuk konsistensi dalam pelatihan.
5) *Random_flip* : Teknik *augmentasi* berupa *random flip* diterapkan untuk meningkatkan variasi data dengan membalik gambar secara acak, yang membantu model belajar dari lebih banyak kemungkinan bentuk visual tanpa perlu menambah jumlah *dataset* secara signifikan.

6) *Batch Size*: Karena keterbatasan memori, *batch size* diatur menjadi 1, tetapi untuk tetap menjaga efisiensi pembaruan bobot model, digunakan *gradient_accumulation_steps* selama 4 langkah, yang memungkinkan akumulasi *gradien* sebelum dilakukan pembaruan parameter.
7) *Max_training_steps* : jumlah langlah pelatihan yang akan dijalankan
8) Optimasi: Proses optimasi menggunakan scheduler *cosine_with_restarts*, yang memungkinkan penurunan learning rate secara bertahap dengan mekanisme restart untuk menghindari *overfitting*, di mana *learning rate* awal ditetapkan pada 2e-5 dengan warmup selama 100 langkah guna menstabilkan pelatihan awal dan mencegah perubahan bobot yang terlalu drastis.
9) *Precision*: Pelatihan dilakukan dengan presisi campuran (fp16) untuk mengurangi konsumsi memori dengan tetap mempertahankan akurasi model, sehingga memungkinkan proses *fine-tuning* berjalan lebih efisien pada perangkat dengan keterbatasan sumber daya.
10) *Checkpoint*: Model disimpan setiap 100 langkah untuk evaluasi bertahap.
11) *Validation prompt*: Untuk memantau kualitas model, validasi dilakukan menggunakan *prompt* yang mendeskripsikan lanskap fantasi yang kompleks.
12) Rank LoRA: *Parameter rank*=16 digunakan untuk mengontrol jumlah matriks low-rank dalam adaptasi LoRA.
13) *TensorBoard*: Pelatihan dipantau secara real-time melalui *TensorBoard* untuk melacak metrik dan performa model.

b. Proses *Noise-Denoising*

Gambar di bawah ini menunjukkan bagaimana tingkat *noise* dalam ruang laten mempengaruhi hasil akhir dari model difusi.



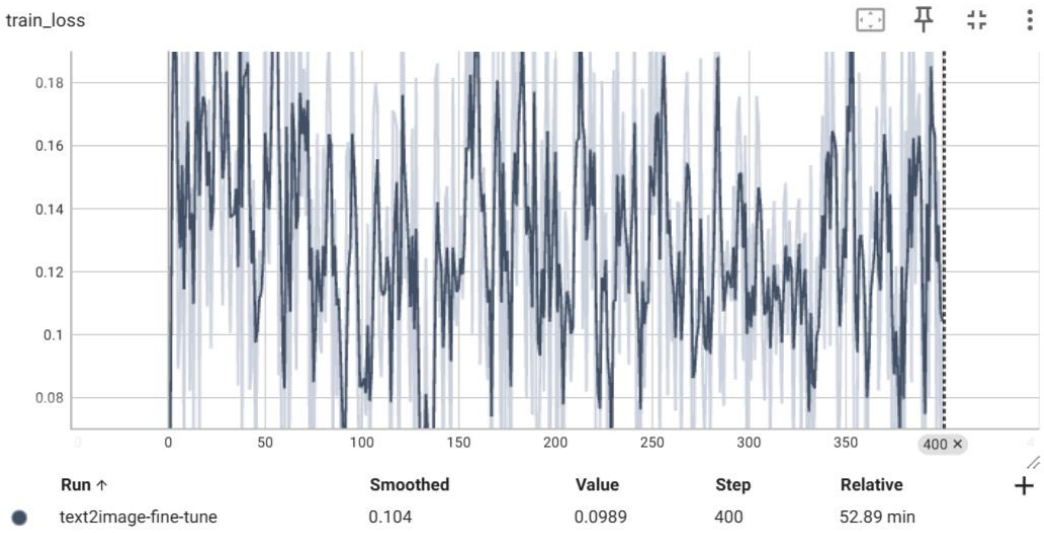

Gambar 4. 7 Visualisasi *Noising* dan *Denoising*

Gambar di atas menunjukkan proses penambahan *noise* dalam ruang *laten* dan hasil *denoising* menggunakan model difusi untuk menghasilkan gambar akhir. Pada baris pertama, tiga gambar grayscale menampilkan tingkat *noise* yang berbeda dalam ruang laten: tanpa *noise* (0), *noise* sedang (0.5), dan *noise* penuh (1.0). Semakin tinggi tingkat *noise*, semakin tidak terstruktur pola dalam ruang laten. Pada baris kedua, hasil denoising dari setiap tingkat *noise* ditampilkan sebagai gambar berwarna. Gambar pertama (tanpa *noise*) menghasilkan output yang jelas dan detail. Gambar kedua (*noise* 0.5) menunjukkan hasil yang lebih terdistorsi tetapi masih dapat dikenali, sementara gambar ketiga (*noise* 1.0) menghasilkan hasil yang sangat kacau dengan kehilangan detail yang signifikan. Ilustrasi ini menggambarkan bagaimana tingkat *noise* dalam ruang laten mempengaruhi kualitas akhir gambar yang dihasilkan melalui proses denoising dalam model difusi.

c. Grafik *fine-tuning*

Grafik *loss* dan jumlah langkah pelatihan digunakan untuk memantau proses *training* model. Grafik ini membantu mengevaluasi performa model seiring bertambahnya jumlah langkah pelatihan. Pada step ke-400, nilai *loss* mencapai 0.0989, dengan nilai *loss* yang telah dismoothing sebesar 0.104. Penurunan nilai *loss* ini menunjukkan bahwa model semakin baik dalam memprediksi gambar berdasarkan deskripsi teks. Meskipun terdapat fluktuasi

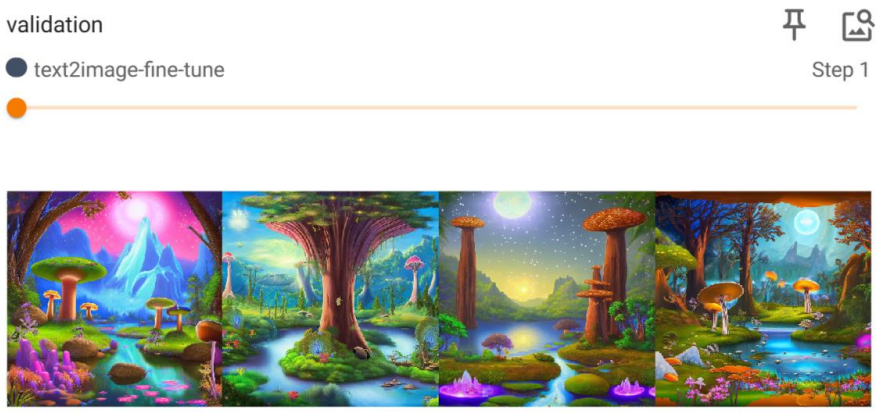
sepanjang training, tren penurunan secara keseluruhan mengindikasikan peningkatan kualitas model.




Gambar 4. 8 Grafik *Training*

2) Hasil *validation_prompt*

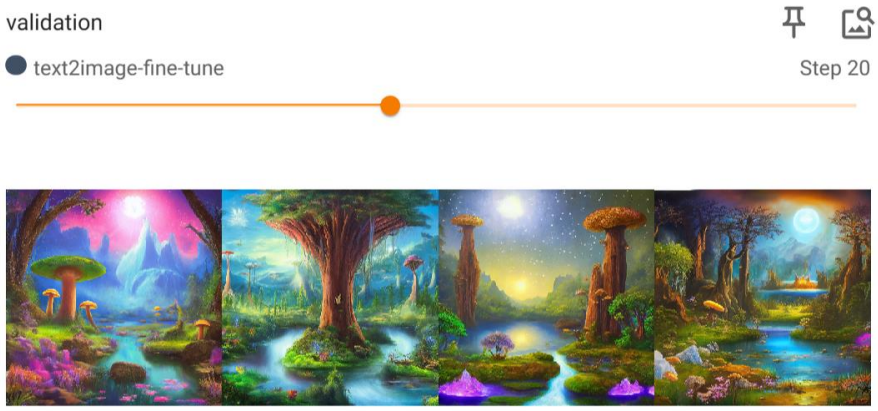
Gambar dibawah ini merupakan hasil dari *generate* gambar pada awal




Gambar 4. 9 Hasil *Generate Validation Prompt Step 1*

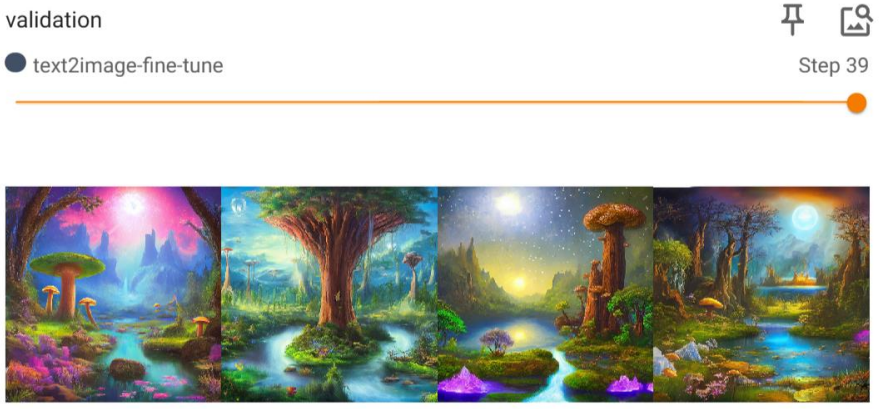
Pada Step 1, gambar yang dihasilkan masih dalam tahap awal pembentukan dengan struktur yang belum jelas. Elemen-elemen lanskap seperti pohon, jamur, dan pencahayaan masih terlihat kabur dan kurang terdefinisi. Warna yang muncul cenderung kasar dengan banyak *noise*,

menunjukkan bahwa model masih berada dalam proses awal transformasi dari *noise* acak menuju bentuk yang lebih bermakna.




Gambar 4. 10 Hasil *Generate Validation Prompt Step 20*

Pada Step 20, gambar mulai menunjukkan bentuk yang lebih jelas dibandingkan tahap sebelumnya. Lanskap fantasi yang dihasilkan memiliki elemen yang lebih terstruktur, dengan objek seperti pohon dan pencahayaan yang sudah tampak lebih nyata. Warna-warna juga mulai lebih seimbang, dan detail mulai muncul meskipun masih ada beberapa bagian yang terlihat belum sepenuhnya matang. Model telah mengurangi *noise* secara signifikan, membuat gambar semakin dekat dengan hasil akhir yang diinginkan.




Gambar 4. 11 Hasil *Generate Validation Prompt Step 39*

Pada Step 39, gambar mencapai tingkat kedetailan yang lebih tinggi, dengan tekstur yang lebih halus dan warna yang lebih kaya. Lanskap yang dihasilkan tampak lebih realistis dan konsisten, dengan pencahayaan yang

lebih alami serta struktur objek yang lebih stabil. Detail-detail kecil seperti refleksi cahaya dan gradasi warna sudah lebih baik, menunjukkan bahwa model telah menyelesaikan proses transformasi dari *noise* menjadi gambar berkualitas tinggi.

1. Hasil *Generate* Gambar Setelah *Fine-tuning*

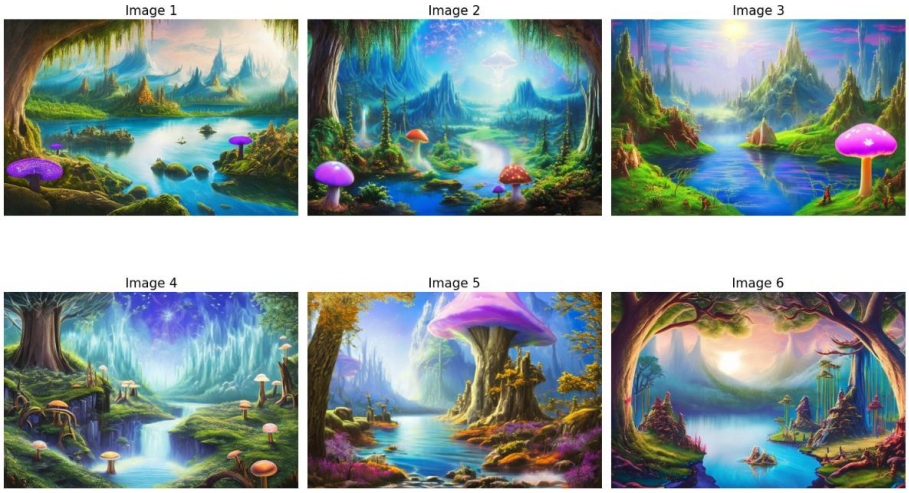
Setelah model selesai melalui proses pelatihan, beberapa contoh gambar dihasilkan untuk menilai peningkatan kualitas visual. Gambar-gambar ini menunjukkan perbaikan dalam hal:

1. Kesesuaian antara gambar dan teks deskripsi.
2. Ketajaman detail visual.
3. Konsistensi dalam gaya dan elemen gambar.

Gambar 6 menampilkan hasil menghasilkan dari model yang telah di-*fine-tuning* menggunakan LoRA. Gambar ini di*generate* dengan menggunakan parameter berikut ini :

Tabel 4. 8 Parameter *generate* gambar 1

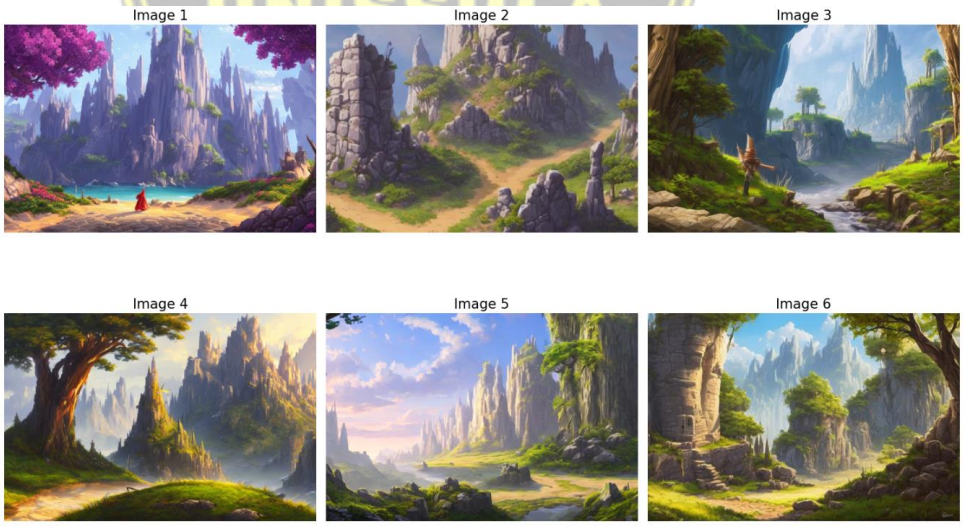
| | |
|---|---|
| *Prompt* | *Create a mystical fantasy landscape at the edge of an enchanted forest. Towering bioluminescent trees glow softly, while a crystal-clear river sparkles with magical energy. Giant mushrooms, glowing flowers, and luminous creatures populate the land. In the distance, towering mountains rise under a starry sky, with floating islands above. The air is misty, giving the scene a dreamy, otherworldly feel.* |
| Negative *prompt* | *cropped, cut-off, close-up, blurry, out of frame, bad composition, missing parts, deformed, distorted, unclear details, bad framing, oversaturated, unnatural lighting, unrealistic proportions* |
| *Seed* | 777 |
| *Guidance Scale* | 7.5 |
| *Image Size* | 768x512 |
| *Num Images* | 6 |




Gambar 4. 12 Hasil *Generate* Gambar setelah *Fine-tuning*

Tabel 4. 9 Parameter generate gambar 2

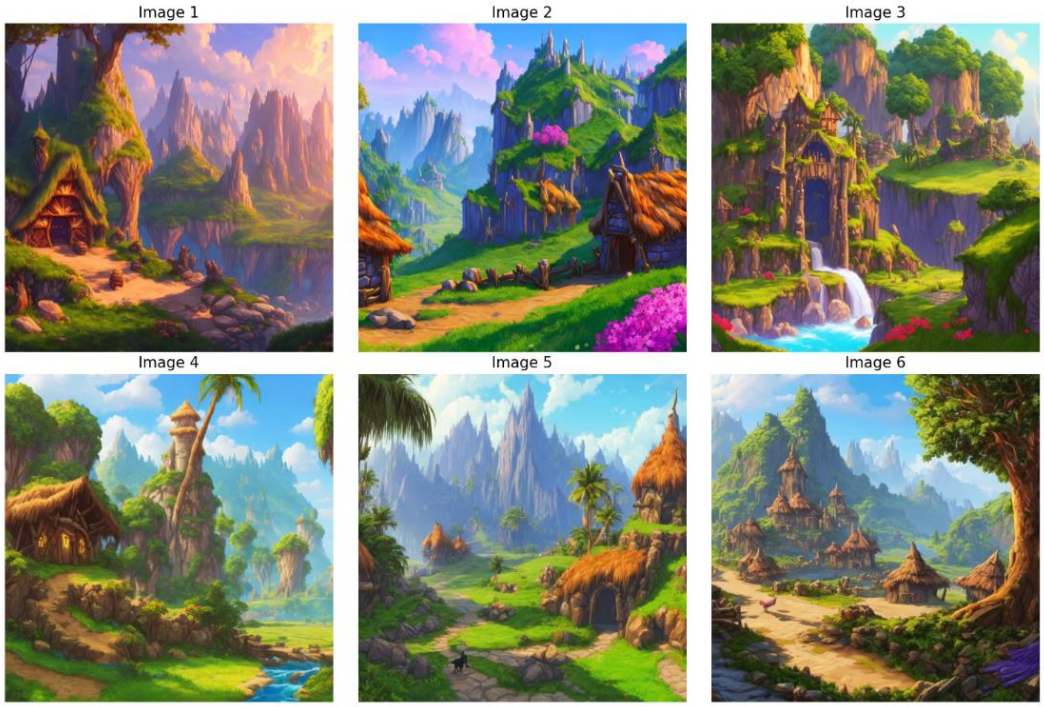
| | |
|---|---|
| *Prompt* | *Ancient ruins with hidden treasures* |
| Negative *prompt* | *cropped, cut-off, close-up, blurry, out of frame, bad composition, missing parts, deformed, distorted, unclear details, bad framing, oversaturated, unnatural lighting, unrealistic proportions* |
| *Seed* | 42 |
| *Guidance Scale* | 7.5 |
| *Image Size* | 768x512 |
| *Num Images* | 6 |




Gambar 4. 13 Hasil *Generate* Gambar Setelah *Fine-tuning*

Tabel 4. 10 Parameter generate gambar 3

| | |
|---|---|
| *Prompt* | *a village with exotic goods, in the style of vibrant fantasy.* |
| Negative *prompt* | *cropped, cut-off, close-up, blurry, out of frame, bad composition, missing parts, deformed, distorted, unclear details, bad framing, oversaturated, unnatural lighting, unrealistic proportions* |
| *Seed* | 1332 |
| *Guidance Scale* | 7.5 |
| *Image Size* | 768x768 |
| *Num Images* | 6 |




Gambar 4. 14 Hasil Generate Gambar Setelah Fine-tuning

## 4.4 Hasil Pengujian Model

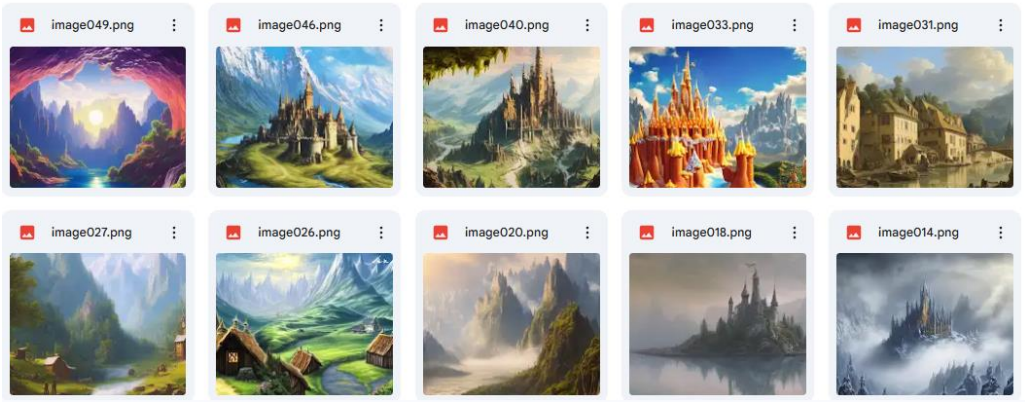
Untuk mengevaluasi performa model, pengujian dilakukan dengan menghasilkan gambar menggunakan teks dari *dataset* pengujian. Setiap teks diuji untuk menghasilkan satu gambar. Setiap teks menghasilkan satu gambar yang kemudian dianalisis untuk menilai kesesuaian semantik dan kualitas visual gambar tersebut.

Gambar 7 menunjukkan contoh gambar yang dihasilkan oleh model setelah pengujian menggunakan *prompt* yang sama dengan *dataset* pengujian, berikut merupakan contoh *prompt* yang digunakan dan gambar yang dihasilkan :

Tabel 4. 11 Contoh *prompt* pengujian

| | |
|---|---|
| *Image049* | *a fantasy landscape with a waterfall and a full moon The landscape features lush green forests, tall mountains, and a vibrant sky. Towering mountains rise high, their peaks covered in snow and surrounded by mist. A mystical mist blankets the valley, adding an air of mystery to the enchanted scene.* |
| *Image046* | *a castle in the mountains A majestic castle stands tall amidst the mist, its towers reaching for the clouds. A mystical mist blankets the valley, adding an air of mystery to the enchanted scene.* |
| *Image031* | *a view of a town with a bridge and a river* |
| *Image027* | *a painting of a landscape with a river and a small town The landscape features lush green forests, tall mountains, and a vibrant sky. Towering mountains rise high, their peaks covered in snow and surrounded by mist. A mystical mist* |

| | |
|---|---|
| | *blankets the valley, adding an air of mystery to the enchanted scene.* |
| *Image020* | *a mountain scene with a lake and a mountain range in the background Towering mountains rise high, their peaks covered in snow and surrounded by mist. A mystical mist blankets the valley, adding an air of mystery to the enchanted scene.* |
| *Immage018* | *a castle on a small island in the middle of a lake A majestic castle stands tall amidst the mist, its towers reaching for the clouds. A mystical mist blankets the valley, adding an air of mystery to the enchanted scene.* |




Gambar 4. 15 Hasil Pengujian

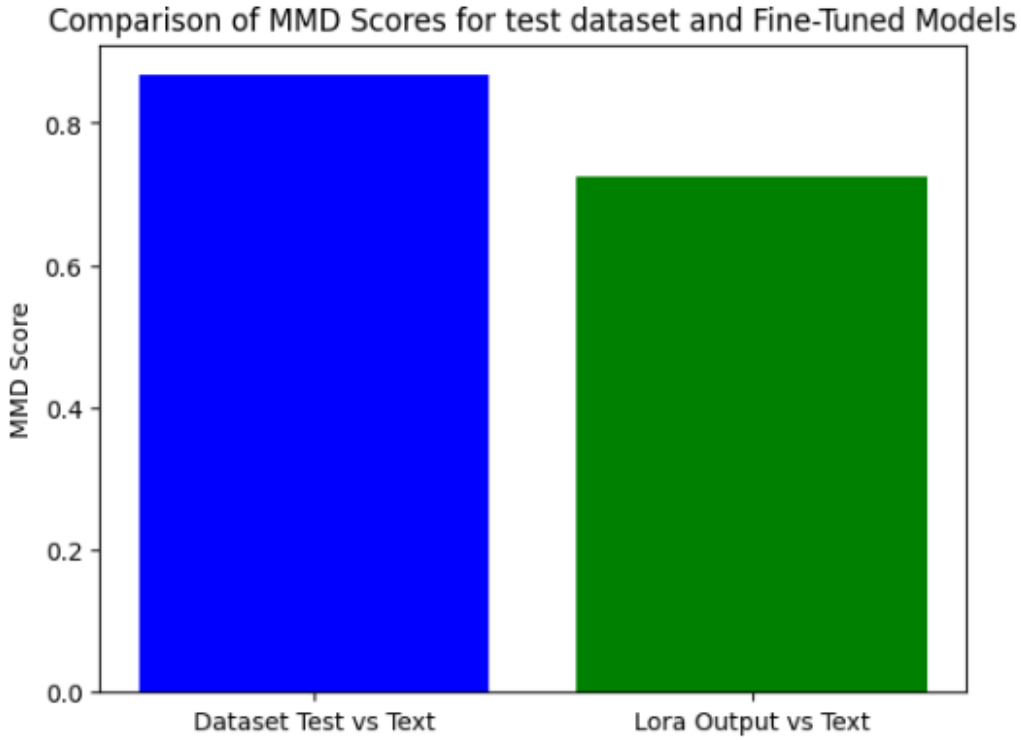
Gambar-gambar ini dianalisis menggunakan metrik kuantitatif berikut :

### 4.4.1 Metrik CLIP-MMD

CLIP-MMD digunakan untuk mengevaluasi kesesuaian semantik antara teks *Input* dan gambar yang dihasilkan. Nilai CLIP-MMD yang lebih rendah menunjukkan kesesuaian semantik yang lebih baik. Hasil pengujian menunjukkan:

a. *Dataset* pengujian dengan *caption* memiliki nilai CMMD *Score* sebesar 0.867, menunjukkan kesesuaian yang cukup tinggi antara gambar dan teks.
b. Setelah *fine-tuning* dengan LoRA, terjadi penurunan nilai CMMD *Score* menjadi 0.725, yang menandakan peningkatan kesesuaian semantik.

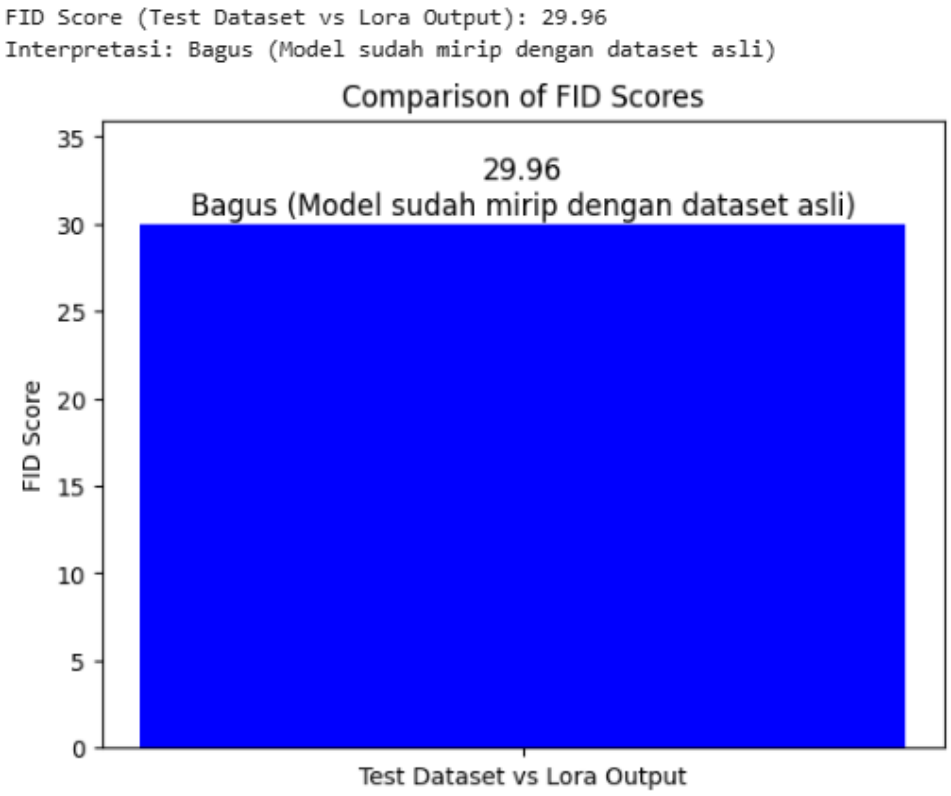
```
CMMD Score (Test Dataset vs Text): 0.8670487403869629
CMMD Score (Lora Output vs Text): 0.7252010703086853
Model fine-tuned menghasilkan gambar yang lebih sesuai dengan prompt!
```




Gambar 4. 16 Metrik CLIP-MMD

### 4.4.2 Metrik FID

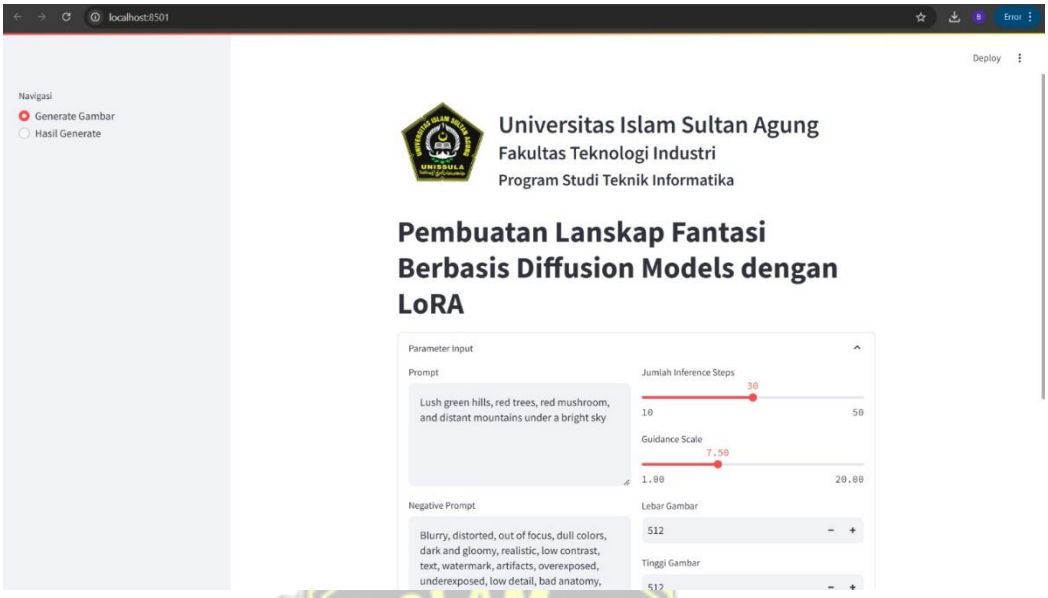
FID digunakan untuk mengukur kualitas visual gambar yang dihasilkan. Semakin rendah nilai FID, semakin baik kualitas visual gambar. Hasil evaluasi menunjukkan bahwa model *fine-tuned* menggunakan LoRA memiliki nilai FID 29.96, yang tergolong kategori nilai yang sudah bagus untuk evaluasi menggunakan FID, menunjukkan model sudah mirip dengan *dataset* asli. :




Gambar 4. 17 Metrik FID

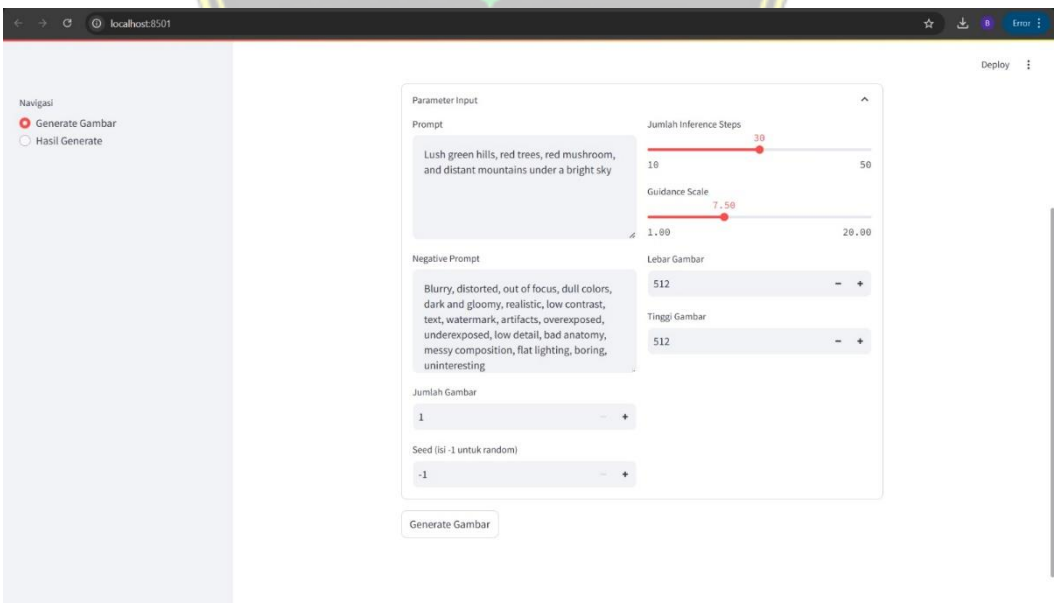
## 4.5 Tampilan Sistem

Website yang digunakan untuk membuat gambar lanskap fantasi dibuat menggunakan *streamlit*, pada *streamlit* ini memuat model *stable diffusion v2.1* yang selanjutnya ditambahkan berat LoRA dari hasil *fine-tuning* model awal.



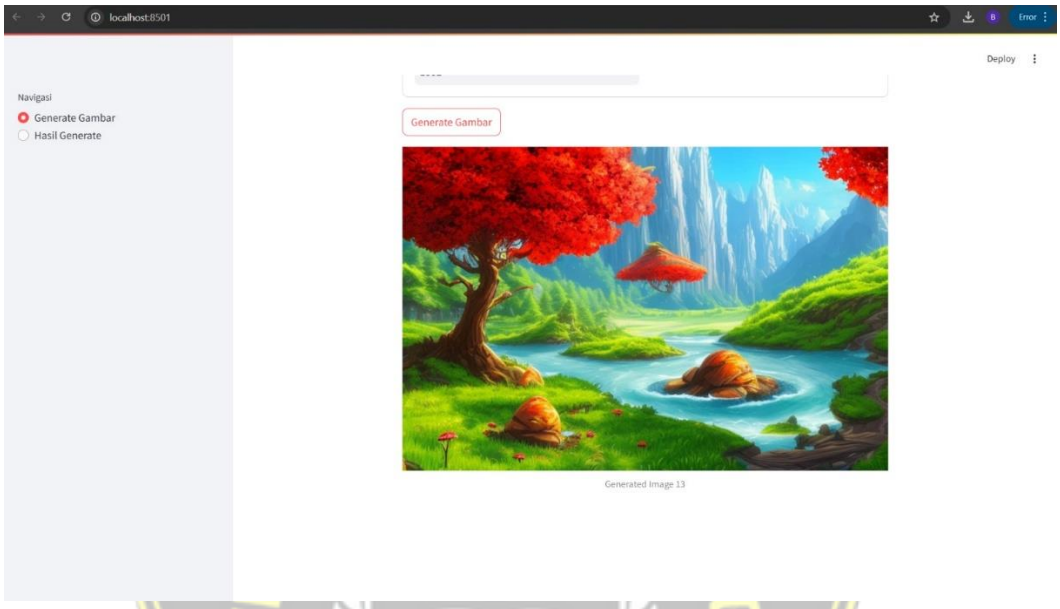

Gambar 4. 18 Tampilan Awal Website

Pada Gambar 4.7 merupakan tampilan awal ketika website dibuka, dimana pada tampilan tersebut terdapat *header* yang berisi Identitas Universitas, Fakultas, dan Program Studi. Selanjutnya terdapat judul tugas akhir yaitu "Pembuatan Lanskap Fantasi Berbasis *Diffusion Models* dengan *Fine-tuning Low-Rank Adaptation* (LoRA). Kemudian ada informasi model yang digunakan yaitu *stable diffusion v2.1* yang diambil dari *library* milik stabilityai, dan *Inputan dropdown* berat LoRA yang akan digunakan.



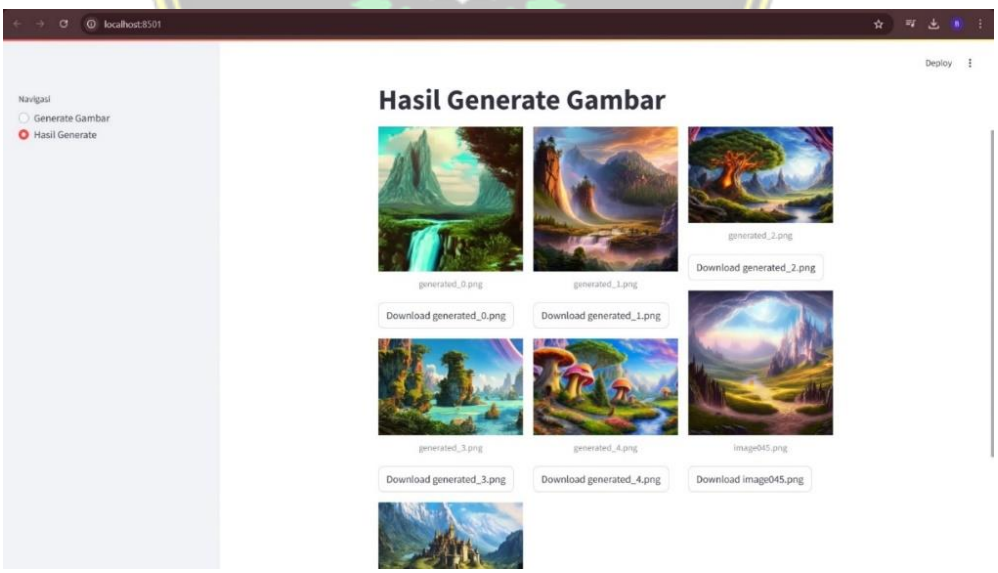

Gambar 4. 19 Parameter *Input*

Pada Gambar 4.8 merupakan tampilan parameter *Input* untuk menghasilkan gambar, berupa *prompt*, *negative prompt*, Jumlah Gambar, *Seed*, Jumlah *Inference steps*, *Guidance Scale*, Lebar Gambar, Tinggi Gambar, serta ada tombol untuk *generate* gambar.




Gambar 4. 20 Hasil *Generate* Gambar

Gambar 4.9 merupakan tampilan dari hasil gambar yang *digenerate* menggunakan *Inputan* yang sebelumnya digunakan.




Gambar 4. 21 Hasil *Generate* Gambar

Gambar 4.10 merupakan tampilan dari halaman yang berisi kumpulan hasil gambar yang sudah di*generate* melalui website, terdapat 2 ukuran gambar berbeda yaitu 512x512 dan 768x512.

## 4.6 Pembahasan Hasil

Berdasarkan hasil evaluasi menggunakan CLIP-MMD dan FID, dapat disimpulkan bahwa model *Stable diffusion* 2.1 yang telah di*fine-tune* menggunakan teknik LoRA berhasil meningkatkan kualitas gambar baik dari segi kesesuaian semantik maupun visual. Dibandingkan dengan model *Stable diffusion* tanpa *fine-tuning*, model ini menunjukkan keunggulan dalam hal:

1. Efisiensi komputasi, karena *fine-tuning* menggunakan LoRA lebih ringan dibandingkan *fine-tuning* penuh.
2. Peningkatan kualitas gambar, baik dari segi detail visual maupun kesesuaian dengan teks deskripsi.

# BAB V
# KESIMPULAN DAN SARAN

## 5.1 Kesimpulan

Berdasarkan hasil penelitian ini, dapat disimpulkan bahwa penerapan teknik *Low-Rank Adaptation* (LoRA) untuk *fine-tuning* model *Stable diffusion* 2.1 berhasil meningkatkan kualitas gambar, baik dari segi kesesuaian semantik dengan teks maupun kualitas visual. Teknik ini terbukti efektif dalam menurunkan nilai CLIP-MMD, yang mencerminkan peningkatan relevansi antara gambar dan teks, serta menurunkan nilai FID, yang menunjukkan peningkatan kualitas visual gambar.

Keunggulan dari penelitian ini terletak pada efisiensi penggunaan LoRA, yang memungkinkan peningkatan kualitas gambar tanpa membebani sumber daya komputasi secara berlebihan. Selain itu, penerapan metode ini dalam pembuatan gambar lanskap fantasi menunjukkan potensi besar dalam seni digital, desain grafis, dan industri kreatif lainnya. Dengan demikian, teknik LoRA dapat menjadi solusi yang efisien dan efektif bagi seniman digital serta pengembang model AI yang ingin meningkatkan hasil generatif mereka tanpa memerlukan perangkat keras berdaya tinggi.

## 5.2 Saran

Untuk penelitian selanjutnya, disarankan agar eksplorasi lebih lanjut dilakukan terhadap kombinasi LoRA dengan teknik lain, seperti kontrol kondisi tambahan atau *fine-tuning* multi-modal, guna meningkatkan kesesuaian gambar dengan teks secara lebih presisi. Selain itu, penggunaan *dataset* yang lebih besar dan beragam dapat membantu mengurangi bias model serta meningkatkan generalisasi hasil yang dihasilkan oleh *Stable diffusion* 2.1.

Selain aspek teknis, penelitian di masa depan juga dapat meneliti lebih lanjut dampak penggunaan LoRA dalam berbagai konteks aplikasi, seperti ilustrasi komersial, *game* development, dan *augmented reality*. Evaluasi subjektif dari pengguna akhir juga dapat menjadi langkah penting untuk memahami kualitas gambar secara lebih mendalam serta menyesuaikan hasil generatif dengan preferensi estetika yang lebih luas.

## DAFTAR PUSTAKA


Awards, T. G. (2024). *Game of the Year*. https://thegameawards.com/. Diambil dari https://thegameawards.com/nominees/game-of-the-year

Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, I. S. K. (2024). A Survey of Text-to-Image *Diffusion* Models in Generative AI. *Proceedings of the 14th International Conference on Cloud Computing, Data Science and Engineering, Confluence 2024*, *14*(8), 73–78. doi: 10.1109/Confluence60223.2024.10463372

Howarth, J. (2024). *How Many Gamers Are There? (New 2024 Statistics)*. EXploding Topics. Diambil dari https://explodingtopics.com/blog/number-of-gamers

Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., & Kumar, S. (2023). *Rethinking FID: Towards a Better Evaluation Metric for Image Generation*. 9307–9315. doi: 10.1109/CVPR52733.2024.00889

Jin, Z., & Song, Z. (2023). *Generating coherent comic with rich story using ChatGPT and Stable diffusion*. Diambil dari http://arxiv.org/abs/2305.11067

Kwon, Y., Wu, E., Wu, K., & Zou, J. (2024). Datainf: Efficiently Estimating Data Influence in Lora-Tuned Llms and *Diffusion* Models. *12th International Conference on Learning Representations, ICLR 2024*.

Li, Y., Jiang, H., Wu, Y., & Engineering, A. P. (n.d.). *Semantic Draw Engineering for Text-to-Image Creation*. *1*(1), 1–6.

Luo, Z., Xu, X., Liu, F., Koh, Y. S., Wang, D., & Zhang, J. (2024). *Privacy-Preserving Low-Rank Adaptation for Latent Diffusion Models*. 1–19. Diambil dari http://arxiv.org/abs/2402.11989

Mehrafrooz, B. (2024). *10 Most Common Challenges of Designing Great Game Environments*. pixune.com. Diambil dari https://pixune.com/blog/most-common-challenges-in-game-environment-design/

Peebles, W., & Xie, S. (2023). Scalable *Diffusion* Models with Transformers. *Proceedings of the IEEE International Conference on Computer Vision*, 4172–4182. doi: 10.1109/ICCV51070.2023.00387

Peng, Y. (2024). A Comparative Analysis Between GAN and *Diffusion* Models in

Image Generation. *Transactions on Computer Science and Intelligent Systems Research*, *5*(D), 189–195. doi: 10.62051/0f1va465

Pradana, A. G., Setiadi, D. R. I. M., & Muslikh, A. R. (2024). Fine tuning model Convolutional Neural Network EfficientNet-B4 dengan augmentasi data untuk klasifikasi penyakit kakao. *Journal of Information System and Application Development*, *2*(1), 01–11. doi: 10.26905/jisad.v2i1.11899

Rantanen, R. (2023). *Open-source tools for automatic generation of game content*. Diambil dari http://www.cs.helsinki.fi/

Sohn, K., Ruiz, N., Lee, K., Chin, D. C., Blok, I., Chang, H., Barber, J., Jiang, L., Entis, G., Li, Y., Hao, Y., Essa, I., Rubinstein, M., & Krishnan, D. (2023). StyleDrop: Text-to-Image Generation in Any Style. *Advances in Neural Information Processing Systems*, *36*(NeurIPS 2023).

Stability.ai. (2022). *Stable diffusion 2.0 Release*. stability.ai. Diambil dari https://stability.ai/news/stable-*diffusion*-v2-release

Totlani, K. (2023). The Evolution of Generative AI: Implications for the Media and Film Industry. *International Journal For Multidisciplinary Research*, *5*(5). doi: 10.36948/ijfmr.2023.v05i05.7537

Valevski, D., Leviathan, Y., Arar, M., & Fruchter, S. (2024). *Diffusion Models Are Real-Time Game Engines*. Diambil dari http://arxiv.org/abs/2408.14837

Valvano, G., Agostino, A., De Magistris, G., Graziano, A., & Veneri, G. (2024). Controllable Image Synthesis of Industrial Data using *Stable diffusion*. *Proceedings - 2024 IEEE Winter Conference on Applications of Computer Vision, WACV 2024*, 5342–5351. doi: 10.1109/WACV57701.2024.00527

Wahyu, M. (2024). Analisis Landscape dan Kartun pada Karya Lugas Syllabus. *Ars: Jurnal Seni Rupa dan Desain*, *27*(1), 31–38. doi: 10.24821/ars.v27i1.7558

Wang, W., Sun, Y., Yang, Z., Hu, Z., Tan, Z., & Yang, Y. (2024). *Replication in Visual Diffusion Models: A Survey and Outlook*. 1–20. Diambil dari http://arxiv.org/abs/2408.00001

Yin, X.-X. (2022). *2022_Journal of Healthcare Engineering - 2022 - Yin - Retracted U-Net-Based Medical Image Segmentation.pdf*.

Yu, Y., Zhang, W., & Deng, Y. (2021). Frechet inception distance (fid) for evaluating gans. *Researchgate.Net*, *September*, 1–7. Diambil dari https://www.researchgate.net/profile/Yu-Yu-120/publication/354269184_Frechet_Inception_Distance_FID_for_Evaluating_GANs/links/612f01912b40ec7d8bd87953/Frechet-Inception-Distance-FID-for-Evaluating-GANs.pdf