PENDAHULUAN

1.1 Latar Belakang

Perkembangan *Artificial Intelligence* (AI) modern khususnya pada model generatif telah mencapai tahap signifikan melalui kemunculan model difusi *text-to-image,* dengan model Stable Diffusion sebagai salah satu pendekatan yang paling dominan. Stable Diffusion dibangun di atas arsitektur *Latent Diffusion Models* (LDM) yang diperkenalkan oleh (Rombach dkk., 2022) dan mampu menghasilkan gambar visual berkualitas tinggi dari masukan teks alami (*prompt*). Sejumlah studi menunjukkan bahwa struktur, detail, dan kompleksitas *prompt* memiliki pengaruh signifikan terhadap kepatuhan semantik (Chefer et al., 2023) dan kualitas gambar yang dihasilkan (He et al., 2025) oleh model difusi berbasis teks. Oleh karena itu, *prompt* dapat menjadi variabel utama dalam mengevaluasi kinerja dan respons model generatif

Dalam konteks sintesis *text-to-image,* khususnya pada objek sensitif seperti wajah manusia, *prompt* merepresentasikan informasi semantik dan struktur deskriptif yang kompleks (Sun dkk., 2024). Proses sintesis wajah memadukan berbagai bagian wajah untuk menghasilkan gambar yang realistik dan konsisten, sehingga wajah harus disintesis secara koheren dari bagian-bagian visual tersebut (Sun dkk., 2022). Kesulitan dan banyaknya bagian visual yang harus dipertimbangkan membuat wajah menjadi domain yang tepat untuk menguji model generatif yang sensitif terhadap detail semantik dan visual (Kale & Altun, 2023).

Evaluasi kinerja model *text-to-image* harus mempertimbangkan tiga aspek krusial diantaranya, kualitas gambar, interpretasi semantik, dan efisiensi

komputasi. Ketiga aspek ini saling memengaruhi, dan seringkali tidak dapat dioptimalkan secara bersamaan, bahkan terdapat *trade-off* yang inheren. Model yang menghasilkan kualitas gambar lebih baik cenderung membutuhkan sumber daya komputasi dan waktu inferensi lebih besar (Ho et al., 2020), sementara upaya optimasi efisiensi komputasi dapat menyebabkan penurunan kualitas (Salimans & Ho, 2022). Oleh karena itu, penelitian untuk mengukur dan menganalisis *trade-off* ini sangat penting.

Sebagian besar studi saat ini telah fokus pada peningkatan dan interpretasi semantik. Studi oleh (Mañas et al., 2024) dan (Chefer et al., 2023) menunjukkan bahwa peningkatan interpretasi semantik (*semantic guidance*) atau optimasi *prompt* (OPT2I) secara otomatis dapat secara signifikan meningkatkan konsistensi teks dengan gambar (*prompt fidelity*). Studi (Chefer et al., 2023) khususnya menunjukkan bahwa kesetiaan semantik dapat dicapai melalui manipulasi *token* dalam mekanisme *cross-attention*, yang secara teknis membebani komputasi. Namun, studi-studi ini secara konsisten mengabaikan dampak manipulasi *prompt* yang kompleks tersebut terhadap beban sumber daya *hardware host* seperti, CPU dan RAM.

Penilaian aspek komputasi sangat krusial untuk implementasi model di lingkungan sumber daya yang terbatas. Keterbatasan sumber daya telah mendorong penelitian ke arah efisiensi (Wan et al., 2024), sehingga tantangan komputasi telah diakui sebagai hambatan yang signifikan untuk penerapan model generatif *text-to-image*, terutama model dengan skala besar seperti model difusi (Yang et al., 2024). Kurangnya studi yang membahas secara eksplisit hubungan antara kompleksitas

prompt dan beban sumber daya *host* (CPU/RAM) mengindikasikan adanya kebutuhan data empiris di tingkat operasional. Meskipun beberapa forum dan praktisi menduga bahwa *prompt* yang panjang meningkatkan penggunaan memori dan memperlambat efisiensi, belum ada validasi ilmiah kuantitatif yang mengukur dan menganalisis *trade-off*.

Stable Diffusion v1.4 dan v1.5 hingga saat ini masih menjadi *checkpoint* yang banyak digunakan dalam berbagai penelitian dan aplikasi. Meskipun demikian, evaluasi kuantitatif yang secara khusus membandingkan kinerja kedua versi tersebut masih terbatas, terutama dalam konteks sintesis wajah manusia dan efisiensi penggunaan sumber daya komputasi. Oleh karena itu, penelitian ini memanfaatkan lingkungan *hardware host* yang terbatas sebagai kondisi eksperimental yang sensitif untuk mengisolasi dan mendeteksi *bottleneck* kinerja komputasi yang dipicu oleh perbedaan versi model dan kompleksitas *prompt*.

Menindaklanjuti *gap* tersebut, penelitian ini bertujuan melakukan evaluasi kuantitatif terarah pada Stable Diffusion v1.4 dan v1.5 dalam menghasilkan gambar wajah manusia menggunakan *dataset CelebA* sebagai referensi visual, karena *dataset* ini secara luas digunakan dan merepresentasikan atribut wajah manusia secara konsisten (Liu dkk., 2015). Evaluasi difokuskan pada analisis hubungan *trade-off* antara kualitas generatif dan efisiensi kinerja komputasi. Variasi *prompt* berdasarkan tingkat kompleksitas deskriptif digunakan sebagai indikator utama untuk mengukur respons model.

Metrik yang digunakan mencakup metrik kualitas dan kesesuaian semantik gambar yaitu, *Face Fréchet Inception Score* (*Face-FID*), *Inception Score* (IS),

Structural Similarity Index Measure (SSIM), dan *CLIP score*, serta metrik efisiensi komputasi yang meliputi waktu inferensi dan penggunaan sumber daya perangkat keras (CPU, GPU dan RAM). Diharapkan penelitian ini dapat memberikan gambaran komprehensif mengenai performa model dan *trade-off* operasional yang relevan sebagai dasar pengambilan keputusan dalam implementasi teknologi *text-to-image*.

1.2 Rumusan Masalah

Berdasarkan latar belakang penelitian, rumusan masalah penelitian ini diantaranya:

1. Bagaimana dampak kompleksitas *prompt* terhadap kualitas gambar dan efisiensi komputasi pada Stable Diffusion v1.4 dan v1.5?
2. Bagaimana pola *trade-off* kinerja model yang teridentifikasi antara kualitas gambar dan efisiensi komputasi pada Stable Diffusion v1.4 dan v1.5?
3. Bagaimana perbedaan respons model terhadap kompleksitas *prompt* pada Stable Diffusion v1.4 dan v1.5?

1.3 Tujuan Penelitian

Tujuan penelitian ini diantaranya untuk:

1. Menganalisis dan mengukur dampak kuantitatif kompleksitas *prompt* terhadap metrik kualitas gambar dan efisiensi komputasi pada setiap model yang diuji.
2. Menganalisis dan mengidentifikasi pola *trade-off* antara kualitas gambar dan efisiensi komputasi pada setiap model yang diuji.

3. Membandingkan dan mengevaluasi respons model berdasarkan kompleksitas *prompt* pada setiap model yang di uji.

1.4 Manfaat Penelitian

Manfaat teoretis dan praktis yang diperoleh dalam penelitian ini diantaranya:

1. Menghasilkan kontribusi literatur dengan menyediakan metodologi evaluasi kuantitatif yang mengintegrasikan metrik kualitas gambar dan beban sumber daya *hardware host*.
2. Menyajikan data empiris terperinci mengenai pola *trade-off* kinerja antara kualitas gambar dan efisiensi komputsi pada Stable Diffusion v1.4 dan v1.5 di lingkungan sumber daya terbatas.
3. Memberikan analisis komparatif yang ketat dan terukur pada respons model dalam memperkaya pemahaman mengenai perbedaan arsitektur dan sensitivitas Stable Diffusion v1.4 dan v1.5.
4. Menyediakan rekomendasi berbasis data untuk pengambilan keputusan strategis mengenai pemilihan model Stable Diffusion dan *prompt* yang paling optimal.

1.5 Batasan Penelitian

Batasan masalah yang perlu diperhatikan dalam penelitian ini diantaranya:

1. Penggunaan *dataset CelebA* sebagai *ground truth* evaluasi dibatasi pada 500 sampel gambar.

2. Variasi *prompt* yang digunakan dibatasi pada 10 variasi tingkat kompleksitas deskriptif, yang disusun dari *prompt* paling sederhana hingga kompleks, tanpa melibatkan optimasi atau *prompt engineering* lanjutan.
3. Evaluasi kualitas gambar dibatasi pada metrik objektif *Face Fréchet Inception Score* (*Face-FID*), *Inception Score* (IS), dan *Structural Similarity Index Measure* (SSIM).
4. Evaluasi kesesuaian semantik antara *prompt* dan gambar dilakukan menggunakan *CLIP score*.
5. Evaluasi efisiensi komputasi dibatasi pada pengukuran waktu inferensi dan penggunaan sumber daya *hardware host,* yaitu CPU, GPU, dan RAM.
6. Penelitian dilakukan pada lingkungan komputasi dengan keterbatasan sumber daya *hardware host*, sehingga hasil evaluasi tidak dimaksudkan untuk digeneralisasi pada sistem komputasi berskala besar.

Penelitian ini bersifat eksploratif dan bertujuan mengidentifikasi pola kecenderungan serta identifikasi pola *trade-off* antar metrik evaluasi, tanpa mengasumsikan hubungan kausal secara langsung.