BAB II

TINJAUAN PUSTAKA

2.1 Kecerdasan Buatan dan Model Generatif

Kecerdasan buatan atau *Artificial Intelligence* (AI) merupakan cabang ilmu komputer yang berfokus pada pengembangan sistem yang mampu melakukan tugas yang membutuhkan kecerdasan manusia, seperti pengambilan keputusan, pemrosesan bahasa, dan pengenalan pola. Salah satu perkembangan paling signifikan dalam AI modern adalah model generatif (*generative models*), yaitu sistem untuk menghasilkan data baru dengan mempelajari distribusi probabilitas data sehingga mampu mensintesis data yang mirip dengan data asli (Goodfellow et al., 2014). Model generatif berkembang dalam berbagai bentuk, seperti *Generative Adversarial Networks* (GAN) (Goodfellow et al., 2014), *Variational Autoencoders* (VAE) (Kingma & Welling, 2013), dan *Diffusion Models* (Ho et al., 2020).

GAN menghasilkan data melalui mekanisme berlawanan antara generator dan diskriminator, sedangkan VAE menggunakan pendekatan probabilistik melalui ruang laten. *Diffusion Models* bekerja dengan menghilangkan *noise* secara bertahap untuk menghasilkan gambar berkualitas tinggi dari representasi laten. Model ini menjadi sangat populer karena kemampuannya menghasilkan gambar dengan detail dan stabilitas yang lebih baik dibandingkan model generatif sebelumnya.

2.2 *Diffusion Models*

Diffusion models bekerja melalui dua tahap utama yaitu *forward diffusion* dan *reverse diffusion*. Pada tahap *forward diffusion*, gambar asli secara bertahap ditambahkan *noise* hingga menjadi distribusi Gaussian murni. Sebaliknya, pada

reverse diffusion, model mempelajari cara menghilangkan *noise* satu per satu hingga menghasilkan gambar baru (Ho et al., 2020).

Arsitektur *diffusion model* umumnya menggunakan UNet sebagai inti pemrosesan, dilengkapi *noise scheduler* dan teknik pelatihan *variational*. Jika dibandingkan dengan GAN, *diffusion models* lebih stabil dan menghasilkan gambar lebih baik, khususnya pada domain wajah yang sensitif terhadap distorsi detail (Dhariwal & Nichol, 2021). Sebagai kompensasinya, *diffusion models* membutuhkan komputasi lebih tinggi karena proses sampling yang iteratif.

Dalam perkembangan model generatif *text-to-image*, *diffusion models* yang memiliki keunggulan dari aspek kontrol, kualitas, dan fleksibilitasnya menjadi fondasi utama berbagai model mutakhir seperti Stable Diffusion, DALL-E, Imagen, dan sebagainya.

2.3 Sintesis *Text-to-Image*

Sintesis *text-to-image* bertujuan menghasilkan gambar berdasarkan deskripsi teks (*prompt*). Model memanfaatkan teks *encoder* untuk memahami konten teks, kemudian menghasilkan gambar yang sesuai melalui proses generatif. Kompleksitas *prompt* memiliki peran penting. Semakin detail dan panjang *prompt*, semakin banyak informasi yang harus dipetakan oleh model ke dalam ruang visual (Saharia et al., 2022).

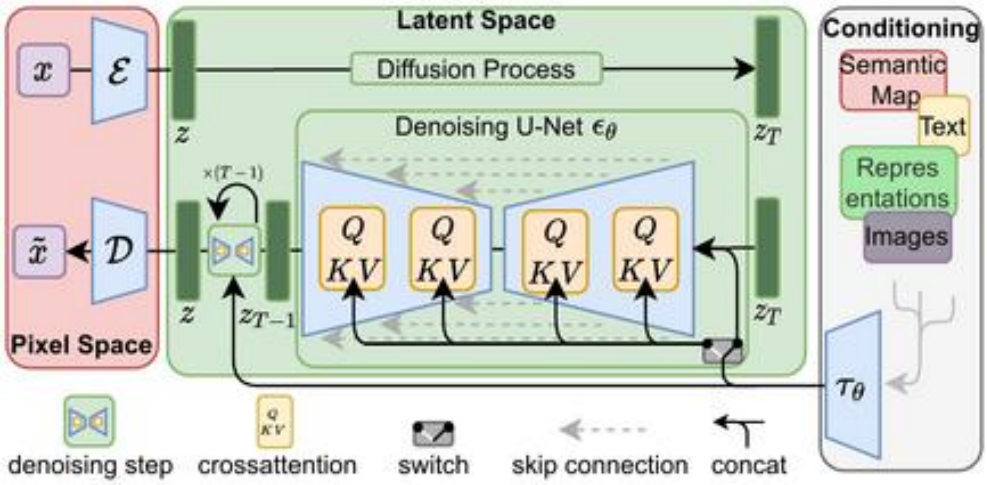
Dalam kontes sintesis *text-to-image*, wajah manusia merupakan salah satu objek yang paling kompleks dan sensitif untuk dievaluasi. Sintesis wajah mengacu pada proses pembangkitan gambar wajah baru oleh model generatif yang mempelajari distribusi data wajah nyata (Sun dkk., 2022). Proses ini menuntut

konsistensi visual dan semantik yang tinggi antar bagian wajah, karena setiap bagian wajah saling bergantung secara struktural (Kale & Altun, 2023). Kompleksitas struktur wajah menjadikan domain ini relevan sebagai objek performa model generatif berbasis teks.

Tantangan utama *text-to-image* tertuju pada cara model menafsirkan struktur semantik dalam *prompt*, seperti atribut wajah, gaya, ekspresi, dan pencahayaan. Kompleksitas *prompt* dapat memengaruhi kualitas gambar dan beban komputasi, karena model membutuhkan lebih banyak *attention computation* pada token yang lebih panjang (Vaswani, 2017). Hal ini membuat hubungan antara *prompt*, kualitas gambar, dan efisiensi komputasi menjadi sangat relevan untuk dievaluasi, terutama pada domain wajah manusia.

2.4 *Stable Diffusion*

Stable Diffusion dirilis tahun 2022 oleh CompVis, Stability AI sebagai model generatif *text-to-image* berbasis *Latent Diffusion Model* (LDM). Model ini bekerja melalui proses generasi pada ruang laten yang lebih rendah resolusi sehingga lebih efisien secara komputasi dibanding *diffusion models* berbasis piksel. Arsitektur utamanya terdiri atas UNet untuk proses *denoising*, *Variational Autoencoder* (VAE) untuk *encoding-decoding* gambar, dan *CLIP Text Encoder* untuk memahami *prompt* (Rombach et al., 2022). Ilustrasi cara kerja pembangkitan gambar di balik model Stable Diffusion (LDM) ditunjukkan pada Gambar 2.1.



Gambar 2. 1 Arsitektur *Diffusion Model* di Ruang Laten (Rombach, dkk., 2022)

Gambar 2.1 menunjukkan kinerja LDM yang dimulai mereduksi dimensi gambar ke dalam ruang laten melalui VAE untuk efisiensi komputasi, kemudian melakukan proses *denoising* iteratif menggunakan jaringan U-Net yang dikondisikan oleh fitur semantik dari *CLIP Text Encoder*. Hasil dari proses rekonstruksi laten tersebut kemudian dikonversi kembali leh *decoder* VAE menjadi gambar beresolusi tinggi di ruang piksel sesuai dengan instruksi tekstual (Rombach dkk., 2022).

Stable Diffusion memungkinkan gambar yang dihasilkan berkualitas tinggi dan mencapai proses yang efisien pada perangkat keras dengan sumber daya terbatas. Perbedaan antara Stable Diffusion v1.4 dan v1.5 terletak pada data pelatihan dan optimasi yang lebih baik pada v1.5, sehingga versi tersebut cenderung menghasilkan gambar lebih konsisten, terutama pada wajah dan tekstur detail. Peningkatan ini relevan untuk penelitian yang mengevaluasi stabilitas model terhadap variasi komputasi.

Meskipun demikian, model generatif pada domain wajah memiliki keterbatasan seperti potensi bias *dataset*, distorsi struktur, *facial artifacts*, atau pemetaan identitas yang tidak akurat (Schramowski et al., 2023). Maka dari itu, evaluasi wajah memerlukan pendekatan metrik evaluasi kualitas gambar yang lebih khusus seperti *Face-FID* dan SSIM.

2.5 *Dataset* Wajah Manusia

Dataset CelebA (Liu dkk., 2015) adalah salah satu *dataset* wajah manusia yang paling banyak digunakan dalam penelitian generatif. *Dataset* ini berisi lebih dari 200.000 gambar wajah dengan variasi identitas, ekspresi, pose, dan atribut visual. *Dataset* ini digunakan dalam penelitian ini karena memiliki struktur wajah yang konsisten, ukuran data yang besar, serta anotasi atribut yang lengkap.

2.6 Variasi *Prompt* dalam Sintesis Wajah

Prompt adalah instruksi atau isyarat yang diberikan kepada *AI Image Generator* untuk memandu pembuatan gambar berdasarkan masukan teks yang diberikan, sehingga *prompt* memiliki pengaruh besar terhadap hasil yang diperoleh (Chen et al., 2024). Studi (Liu & Chilton, 2022) menemukan bahwa detail *prompt* dapat meningkatkan kualitas kontrol tetapi juga meningkatkan risiko distorsi jika model tidak mampu menafsirkan *prompt* secara konsisten. Penemuan lain oleh (Jiao et al., 2025) membuktikan bahwa *prompt* yang panjang dan detail dapat mengurangi akurasi model dalam mengenali atribut dan struktur visual.

2.7 Evaluasi Kualitas Gambar

Kualitas gambar yang dihasilkan oleh model generatif biasanya dievaluasi menggunakan metrik otomatis seperti *Face Fréchet Inception Distance (Face-*

FID), Inception Score (IS), *Structural Similarity Index Measure* (SSIM), dan *CLIP Score*.

2.7.1. *Face Fréchet Inception Distance* (*Face-FID*)

Fréchet Inception Distance (FID) terbukti sebagai salah satu metrik yang paling andal untuk mengevaluasi kualitas dan kemiripan distribusi gambar. Secara umum metrik ini digunakan untuk mengukur jarak antara distribusi gambar yang dihasilkan dan gambar nyata dalam ruang fitur (Heusel et al., 2017). *Face Fréchet Inception Distance (Face-FID*) merupakan adaptasi dari FID, fitur yang diekstrak *Face-FID* diambil dari model yang lebih relevan untuk wajah. *Face-FID* lebih sensitif terhadap atribut wajah yang penting, seperti identitas dan detail wajah dibanding FID standar yang menggunakan fitur dari model umum (Saritas & Ekenel, 2024). Perhitungan *Face-FID* ditunjukkan pada Persamaan (3.1):

$$FID = \left\|\mu_r - \mu_f\right\|^2 + \mathrm{T}_r\left(\sum_r + \sum_f - 2\left(\sum_r \sum_f\right)^{1/2}\right) \tag{3.1}$$

Keterangan variabel:

μ_r : Nilai rata-rata (*mean*) dari *embedding* wajah asli.

μ_f : Nilai rata-rata (*mean*) dari *embedding* wajah hasil generatif.

$\sum_r$: Nilai *covariance* dari *embedding* wajah asli.

$\sum_f$: Nilai *covariance* dari *embedding* wajah hasil generatif.

T_r : Simbol *trace* (penjumlahan elemen diagonal matriks).

Skor *Face-FID* yang lebih rendah mengindikasikan kualitas dan kesamaan yang lebih baik dengan gambar asli secara statistik dan perseptual (Kabra & Balakrishnan, 2023).

2.7.2. *Inception Score* (IS)

Inception Score (IS) digunakan untuk mengevaluasi kualitas dan keragaman semantik antar gambar yang dihasilkan oleh model generatif *text-to-image*. Metrik ini dihitung berdasarkan distribusi probabilitas kelas yang diprediksi oleh model klasifikasi *inception*, yaitu distribusi kondisional $p(y|x)$ untuk setiap gambar dan distribusi marginal $p(y)$ yang diperoleh dari rata-rata $p(y|x)$ pada semua gambar dalam satu subset. Metrik IS mengukur perbedaan (*divergence*) antara kedua distribusi probabilitas ini menggunakan *Kullback-Leiber (KL) divergence*, yang kemudian dirata-rata secara eksponensial, sebagaimana ditunjukkan pada Persamaan (3.2).

$$IS = exp(\mathbb{E}_x[D_{KL}(p(y|x)\|p(y))]) \tag{3.2}$$

Keterangan variabel:

$\mathbb{E}_x$: Nilai harapan dari seluruh gambar (x) yang dihasilkan.

D_{KL} : *Kullback-Leibler Divergence*.

$p(y|x)$: Distribusi probabilitas label kelas (y) untuk gambar tertentu (x).

$p(y)$: Distribusi marginal dari label kelas untuk seluruh gambar yang dihasilkan.

Nilai IS yang lebih tinggi mengindikasikan kualitas yang baik dan keragaman yang tinggi. Hal ini ditandai distribusi prediksi kelas yang tajam (entropi rendah pada $p(y|x)$) serta keragaman semantik yang tinggi, yang tercermin dari distribusi kelas $p(y)$ yang lebih merata (Salimans et al., 2016).

2.7.3. *Structural Similarity Index Measure* (SSIM)

Structural Similarity Index Measure (SSIM) diperkenalkan pertama kali oleh (Wang et al., 2004) sebagai metode evaluasi kesamaan struktural antara dua citra. Metrik ini dirancang dengan mempertimbangkan persepsi sistem visual manusia yang lebih sensitif terhadap perubahan struktur dibandingkan sekadar perbedaan nilai piksel individual. SSIM mengukur sejauh mana citra hasil generatif text-to-image mendekati citra asli dari segi struktur (*structure*), kekontrasan (*contrast*), dan luminansi (*luminance*) sekaligus memberi informasi sejauh mana citra hasil generatif mempertahankan struktur dan kualitas visualnya. Secara matematis, SSIM ditunjukkan pada Persamaan (3.3):

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{3.3}$$

Keterangan variabel:

μ_x dan μ_y : Rata-rata lokal dari gambar x dan y yang merepresentasikan aspek luminasi.

σ_x^2 dan σ_y^2 : Varians dari gambar x dan y sebagai representasi aspek kontras.

σ_{xy} : Kovarians antara gambar x dan y yang menunjukkan hubungan struktur antar kedua gambar.

C_1 dan C_2 : Konstanta penstabil.

$$C_1 = (K_1 L)^2$$

$$C_2 = (K_2 L)^2$$

L : Rentang dinamis nilai piksel.

Metrik ini menghasilkan nilai rentang -1 dan 1. Nilai yang mendekati 1 mengindikasikan bahwa kedua gambar memiliki tingkat kesamaan yang tinggi atau identik secara struktural.

2.7.4. *CLIP Score*

CLIP Score merupakan metrik evaluasi untuk mengukur tingkat relevansi gambar yang dihasilkan dan *prompt* yang diberikan. Metrik ini dikembangkan berdasarkan model *Contrastive Language-Image Pretraining* (CLIP) yang diperkenalkan oleh (Radford et al., 2021). Secara teknis, *CLIP Score* dihitung menggunakan metode *cosine similarity* antara vektor *prompt* dan vektor gambar yang telah diekstraksi melalui *encoder* masing-masing. Rumus perhitungan *CLIP Score* ditunjukkan pada Persamaan (3.4).

$$CLIPScore = \cos\theta = \frac{v_t \,.\, v_i}{\|v_t\|\|v_i\|} \tag{3.4}$$

Keterangan variabel:

v_t : Vektor fitur (*embedding*) dari *prompt*.

v_i : Vektor fitur (*embedding*) dari gambar yang dihasilkan.

$v_t \,.\, v_i$: Perkalian antara vektor *prompt* dan gambar.

$\|v_t\|\|v_i\|$: Perkalian dari norma (panjang) masing-masing vektor.

Interpretasi nilai *CLIP Score* berkisar antara -1 hingga 1. Skor yang lebih tinggi mengindikasikan bahwa gambar yang dihasilkan memiliki kesesuaian semantik yang sangat kuat dengan *prompt*.

Setiap metrik memiliki kelebihan dan kekurangan, sehingga kombinasi empat metrik dapat menunjang evaluasi kualitas gambar wajah yang lebih komprehensif.

2.8 Evaluasi Efisiensi Komputasi

Evaluasi efisiensi komputasi menjadi aspek penting dalam penelitian model generatif *text-to-image* karena proses inferensi *diffusion models* cenderung memerlukan sumber daya perangkat keras yang intensif. Evaluasi efisiensi komputasi berfungsi untuk meninjau bagaimana pengaruh variasi kompleksitas *prompt* terhadap waktu inferensi dan seberapa optimal model generatif *text-to-image* memanfaatkan sumber daya komputasi terbatas (CPU, GPU, dan RAM). Pada model difusi, proses difusi melibatkan iterasi *sampling* berulang yang secara langsung memengaruhi waktu inferensi dan penggunaan memori GPU (Ho et al., 2020). Sehingga, evaluasi efisiensi komputasi menjadi faktor penting untuk menilai kinerja model difusi. Metrik evaluasi efisiensi komputasi yang digunakan pada penelitian ini adalah pengukuran waktu inferensi, *CPU usage*, *GPU usage*, dan *RAM usage*.

2.8.1 Waktu Inferensi

Waktu inferensi (*inference time*) mengukur durasi proses yang dibutuhkan model untuk menghasilkan gambar. Semakin kompleks *prompt* yang diberikan, seringkali semakin tinggi kebutuhan komputasi karena proses *denoising* membutuhkan langkah-langkah tambahan dalam *pipeline* (Rombach et al., 2022). Maka dari itu, metrik ini sangat relevan pada *diffusion models* yang membutuhkan langkah *sampling* dalam jumlah banyak untuk menghasilkan gambar.

2.8.2 *Central Processing Unit* (*CPU Usage*)

Central Processing Unit (*CPU usage*) mengukur persentase pemanfaatan prosesor selama inferensi model. Walaupun sebagian besar proses utama berada di *Graphics Processing Unit* (GPU), CPU tetap memiliki peran krusial dalam menjalankan *pre-processing*, orkestra *pipeline* komputasi, dan manajemen memori (Narayanan et al., 2021). Penggunaan CPU yang tinggi dapat mengindikasikan *bottleneck* pada *pipeline inference*, terutama ketika model berada pada *batch* besar atau penggunaan token input yang kompleks.

2.8.3 *Graphics Processing Unit* (*GPU Usage*)

Graphics Processing Unit (*GPU usage*) menilai intensitas penggunaan GPU saat model melakukan inferensi. GPU merupakan komponen utama dalam menjalankan operasi tensor besar seperti *convolution* dan *attention* pada *diffusion models* (Shoeybi et al., 2019). Pada umumnya, penggunaan GPU pada model seperti Stable Diffusion sering kali stabil, karena struktur arsitekturnya memiliki jumlah operasi tetap setiap iterasi *sampling* (Ho et al., 2020). Metrik ini penting untuk mengevaluasi pengaruh peningkatan kualitas model terhadap beban pemrosesan GPU.

2.8.4 *Random Access Memory* (*RAM Usage*)

Random Access Memory (*RAM usage*) mengukur jumlah memori sistem yang digunakan selama inferensi. Evaluasi terhadap RAM dapat memberikan gambaran tentang kebutuhan memori sistem dan skalabilitas model ketika digunakan dalam jumlah *batch* yang besar. Pada *diffusion model* seperti LDM,

penggunaan memori ditentukan oleh ukuran *latent space*, resolusi *output*, model *weights*, dan kompleksitas *prompt* (Rombach et al., 2022).

2.9 *Trade-Off* pada Model Generatif *Text-to-Image*

Dalam model generatif berbasis difusi, peningkatan kualitas visual umumnya dicapai melalui proses *denoising* bertahap yang kompleks dan berulang, yang secara inheren meningkatkan kebutuhan waktu inferensi dan penggunaan sumber daya (Ho et al., 2020). Kondisi ini memunculkan fenomena *trade-off* antara kualitas gambar dan efisiensi kinerja komputasi, terutama pada lingkungan dengan keterbatasan sumber daya *hardware*.

Sejumlah penelitian menunjukkan bahwa metrik kualitas visual dan semantik tidak selalu meningkat secara linear seiring dengan meningkatnya kompleksitas proses generatif, sementara beban komputasi cenderung meningkat seiring kompleksitas model dan representasi internal (Rombach et al., 2022). Selain itu, distribusi beban komputasi dapat berbeda pada setiap komponen sumber daya, seperti CPU, GPU, dan memori tergantung pada karakteristik representasi visual dan semantik yang diproses (Chefer et al., 2023).

Oleh karena itu, identifikasi hubungan *trade-off* antara kualitas gambar dan efisiensi komputasi menjadi penting untuk memahami batas optimal performa model serta implikasinya dalam penggunaan praktis, khususnya pada sistem dengan keterbatasan sumber daya.

2.10 Respons Model terhadap Kompleksitas *Prompt*

Respons model generatif *text-to-image* terhadap kompleksitas *prompt* dipengaruhi oleh sensitivitas representasi internal model terhadap perubahan

struktur dan kompleksitas input teks. Studi sebelumnya menunjukkan bahwa perubahan kecil pada *prompt* dapat menghasilkan variasi keluaran yang signifikan, terutama pada model yang mengandalkan mekanisme *cross-attention* antara representasi teks dan gambar (Hertz et al., 2022).

Analisis sensitivitas *prompt* umum digunakan untuk mengukur sejauh mana model merespons perubahan input secara stabil atau fluktuatif, yang dapat merepresentasikan stabilitas dan adaptabilitas model terhadap kompleksitas deskriptif (Meng et al., 2022). Selain itu, perbedaan arsitektur dan optimasi antar versi model dapat menyebabkan respons yang berbeda terhadap *prompt* yang sama, sehingga analisis perubahan model menjadi relevan dalam membandingkan karakteristik kinerja generatif antar model (Rombach et al., 2022).

Dengan demikian, analisis respons model melalui sensitivitas *prompt* dan perbandingan antar versi model memberikan dasar konseptual untuk memahami batas pemanfaatan informasi *prompt* serta implikasi desain model terhadap kualitas gambar dan efisiensi komputasi.

2.11 State Of The Art (SOTA)

Beberapa penelitian terkait dan kebaruan penelitian yang sudah dilakukan diuraikan pada tabel 2.1.

Tabel 2. 1 *State-Of-The-Art*

No.	Penulis, Tahun Penelitian	Judul Penelitian	Fokus Penelitian	Hasil Penelitian
1	(Heusel et al., 2017)	*GANs Trained By A Two Time-Scale Update Rule* (TTUR)	Mengatasi *training instability* dan *mode collapse* yang umum terjadi pada model GANs dengan mengusulkan TTUR.	TTUR terbukti meningkatkan kualitas gambar secara signifikan dan meningkatkan stabilitas proses pelatihan. Hasil penelitian ini mengenalkan metrik FID sebagai metrik standar untuk mengevaluasi kualitas *output* model generatif. Meskipun temuan ini sudah teruji secara historis, analisis efisiensi pada penelitian ini masih terbatas pada stabilitas pelatihan GAN. Oleh karena ini, diperlukan penelitian untuk mengukur efisiensi *inference* (Waktu Inferensi) dan penggunaan sistem *host* (CPU/RAM) terutama pada model difusi modern, seperti LDM.
2	(Ho et al., 2020)	*Denoising Diffusion Probabilistic Models*	Memperkenalkan arsitektur *Denoising Diffusion Probabilistic Models* (DDPM) untuk menghasilkan sampel gambar berkualitas tinggi	DDPM berhasil menunjukkan bahwa *diffusion model* dapat menghasilkan gambar yang sangat berkualitas dan realistis hingga melampaui kinerja GAN. Meskipun begitu, proses *sampling* DDPM

No.	Penulis, Tahun Penelitian	Judul Penelitian	Fokus Penelitian	Hasil Penelitian
			dengan *log-likelihood* yang sangat baik.	sangat lambat (1000-4000 langkah *sampling* atau mencapai beberapa menit per gambar) dan membutuhkan biaya komputasi yang sangat tinggi (VRAM/*GPU memory*) karena *denoising* dilakukan langsung di ruang piksel. Stable Diffusion (LDM) memperbaiki hal ini dengan memindahkan proses ke ruang laten yang lebih efisien. Namun, penilaian peningkatan efisiensi tersebut belum banyak dievaluasi secara kuantitatif.
3	(Song et al., 2021)	*Denoising Diffusion Implicit Models*	Memperkenalkan kerangka kerja *Denoising Diffusion Implicit Models* (DDIM) yang mempertahankan arsitektur *denoising* yang sama dengan DDPM tetapi mengubah proses *sampling* menjadi implisit dan non-Markovian.	DDIM berhasil menunjukkan bahwa proses *sampling* dapat dipercepat secara eksponensial dengan menggunakan langkat *sampling* yang lebih sedikit, tetapi mampu mempertahankan kualitas gambar. DDIM mampu mempercepat waktu inferensi 10-50x dari DDPM. Namun, DDIM hanya berfokus pada efisiensi Waktu Inferensi secara abstrak atau berdasarkan jumlah langkah, sehingga diperlukan penelitian untuk membahas pengukuran spesifik pada metrik *hardware* tingkat rendah (CPU dan RAM) yang dapat menjadi sumber

No.	Penulis, Tahun Penelitian	Judul Penelitian	Fokus Penelitian	Hasil Penelitian
				bottleneck sekunder yang tidak dipertimbangkan DDIM.
4	(Rombach et al., 2022)	*High-Resolution Image Synthesis with Latent Diffusion Models*	Mengusulkan arsitektur *Latent Diffusion Models* (LDM) yang menerapkan proses *denoising diffusion model* pada ruang laten dari *autoencoder* yang telah dilatih sebelumnya.	LDM berhasil mengurangi kebutuhan komputasi (VRAM) secara signifikan dan mempercepat inferensi 20-50 langkah atau beberapa detik per iterasi. LDM mampu menghasilkan gambar beresolusi tinggi dengan kualitas baik. Namun, penelitian ini hanya berfokus pada efisiensi VRAM/GPU dan menguji arsitektur LDM secara umum. Maka dari itu, pengembangan evaluasi efisiensi LDM dengan melibatkan evaluasi pengaruh variasi *prompt* untuk menganalisis pengaruhnya terhadap *resource usage* (CPU dan RAM) pada LDM.
5	(Salimans & Ho, 2022)	*Progressive Distillation for Fast Sampling of Diffusion Models*	Mengembangkan teknik *progressive distillation* untuk mengatasi *latency inference* yang sangat tinggi pada *diffusion models* dengan melatih *model student* agar dapat menghasilkan *output* berkualitas tinggi dalam	*Progressive distillation* berhasil mengurangi jumlah *sampling step* (NFE = 4) secara signifikan dan mencapai waktu *runtime* di bawah 0.2 detik per gambar, sambil mempertahankan kualitas gambar dengan skor FID yang kompetitif. Namun, penelitian ini tidak mengevaluasi dampak *overhead* latensi pada sistem *host*

No.	Penulis, Tahun Penelitian	Judul Penelitian	Fokus Penelitian	Hasil Penelitian
			langkah *sampling* yang sangat sedikit.	yang dipicu oleh *text encoder*. Oleh karena itu, diperlukan pengukuran tambahan untuk menganalisis konsistensi performa model yang sudah dioptimalkan tersebut ketika diberikan variasi kompleksitas *prompt*.
6	(Hu et al., 2022)	*LoRA: Low-Rank Adaptation of Large Language Models*	Mengusulkan LoRA untuk meningkatkan efisiensi *fine-tuning* (penyesuaian) model *foundation* besar tanpa membebani daya komputasi (VRAM dan *storage*).	*LoRA* mampu mengurangi kebutuhan VRAM selama *fine-tuning* dan mengurangi ukuran *file* yang disimpan secara signifikan dengan kualitas hasil model terbukti setara dengan *full fine-tuning* pada sebagian besar *benchmark.* Meskipun begitu, penelitian ini belum mengeksplorasi penilaian terhadap beban sistem *host* selama proses inferensi, terutama ketika model dihadapkan pada variasi kompleksitas *prompt*. Hal ini memberikan urgensi untuk mengukur stabilitas efisiensi komputasi *host* pada Stable Diffusion v1.4 dan v1.5 yang telah mengadopsi prinsip efisiensi serupa.
7	(Podell et al., 2024)	*SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis*	Mengembangkan SDXL untuk meningkatkan kualitas gambar secara drastis dan meningkatkan pemahaman *prompt* yang	SDXL menunjukkan kinerja unggul dengan skor FID yang lebih baik dan kemampuan memahami *prompt* kompleks dengan detail struktural yang tinggi,

No.	Penulis, Tahun Penelitian	Judul Penelitian	Fokus Penelitian	Hasil Penelitian
			kompleks dengan menggunakan model yang lebih besar dan arsitektur dua tahap (*two-stage pipeline*).	disertai efisiensi *sampling* yang tetap kompetitif. Namun, SDXL membutuhkan VRAM yang jauh lebih besar (16-24 GB), sehingga kurang sesuai untuk evaluasi berbasis *CPU/RAM* pada lingkungan komputasi standar. Oleh karena itu, penelitian yang berfokus pada Stable Diffusion v.1x yang lebih ringan diperlukan untuk mengeksplorasi pengaruh variasi kompleksitas *prompt* berpotensi memicu *bottleneck* sekunder pada efisiensi komputasi.
8	(Borji, 2023)	*Generated Faces in the Wild: Quantitative Comparison of* Stable Diffusion, *MIDjourney and DALL-E 2*	Evaluasi kuantitatif dan komparatif terhadap kualitas sintesis wajah manusia yang dihasilkan Stable Diffusion, Midjourney, dan DALL-E 2.	Secara umum, Midjourney menunjukkan kualitas gambar wajah tertinggi (FID paling rendah) dan realisme yang lebih baik dibanding model lain. Di sisi lain Stable Diffusion menunjukkan keunggulan dalam kepatuhan terhadap *prompt* (*CLIP score*), meskipun realisme gambar yang dihasilkan lebih buruk (FID tinggi). Namun, penelitian ini hanya fokus pada kualitas gambar (FID dan *aesthetic*) dan menggunakan *prompt* standard. Oleh karena itu, diperlukan evaluasi untuk menguji kualitas sintesis wajah secara

No.	Penulis, Tahun Penelitian	Judul Penelitian	Fokus Penelitian	Hasil Penelitian
				lebih ketat serta menambahkan evaluasi efisiensi komputasi menggunakan *prompt* yang lebih bervariatif.
9	(Ruiz et al., 2023)	*DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation*	Personalisasi Model Difusi untuk merepresentasikan subjek spesifik (misal: wajah tertentu) dengan konsistensi identitas tinggi.	*Dreambooth* mampu mempertahankan identitas subjek yang sangat tinggi dalam berbagai konteks *prompt* baru hanya dengan sedikit gambar sampel (3-5 gambar). Namun, metode ini membutuhkan *full fine-tuning* pada seluruh bobot UNet yang sangat memakan waktu, boros VRAM, dan rentan terhadap *language drift*. Oleh karena itu, diperlukan evaluasi untuk mengukur sejauh mana model dasar seperti Stable Diffusion v1.4 dan v1.5 dapat mempertahankan kualitas gambar secara efisien tanpa perlu proses *fine-tuning* yang berat seperti *Dreambooth*, sambil memantau penggunaan CPU/RAM.
10	(Croitoru et al., 2023)	*Diffusion Models in Vision: A Survey*	Memberikan tinjauan sistematis dan komprehensif mengenai evolusi dasar teoretis, dan aplikasi dari model difusi dalam domain *computer vision*.	Penelitian ini menemukan *diffusion model* unggul menghasilkan gambar realistis dan koheren, dengan LDM menjadi solusi yang efisien dalam ruang laten sekaligus memvalidasi penggunaan metrik standar

No.	Penulis, Tahun Penelitian	Judul Penelitian	Fokus Penelitian	Hasil Penelitian
				(FID, SSIM, LPIPS, *CLIP score*) untuk *diffusion models*. Namun, penelitian ini hanya fokus pada arsitektur algoritmik dan tidak menguji beban CPU/RAM akibat variasi kompleksitas *prompt* selama proses inferensi. Oleh karena itu, *gap* pengukuran efisiensi komputasi masih diperlukan dalam konteks ini.
11	(Chefer et al., 2023)	*Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models*	Mengembangkan teknik *semantic guidance* yang memodifikasi proses inferensi *diffusion model* untuk mengatasi masalah kegagalan model dalam merepresentasikan semua objek penting yang disebutkan dalam *prompt* yang panjang atau kompleks.	Teknik *semantic guidance* secara signifikan meningkatkan *prompt fidelity*, khususnya untuk *prompt* dengan banyak objek. Namun, evaluasi dalam penelitian ini berfokus pada kualitas *output* dan kepatuhan semantik, tanpa menganalisis efisiensi komputasi atau *overhead* yang ditimbulkan oleh operasi *attention* tambahan dan kompleksitas *prompt*. Oleh karena itu, diperlukan analisis lanjutan untuk mengukur dampak kompleksitas *prompt* terhadap penggunaan sumber daya komputasi pada model difusi.
12	(Feng et al., 2024)	*PromptMagician: Interactive Prompt Engineering for Text-to-Image Creation*	Mengembangkan alat interaktif yang memfasilitasi proses *prompt engineering* bagi pengguna non-ahli.	*PromptMagician* efektif merekomendasikan kata kunci *prompt* yang relevan, membantu eksplorasi interaktif, serta mendukung pengguna

No.	Penulis, Tahun Penelitian	Judul Penelitian	Fokus Penelitian	Hasil Penelitian
				meningkatkan hasil gambar. Namun, penelitian ini sama sekali tidak mengaudit dampak *prompt* yang dihasilkan terhadap sumber daya (CPU/RAM) selama proses inferensi. Analisis efisiensi penelitian ini berhenti pada kualitas *output*, sehingga penilaian dampak variasi kompleksitas *prompt* terhadap efisiensi komputasi perlu dilakukan.
13	(Croitoru et al., 2024)	*Reverse Stable Diffusion: What Prompt was Used to Generate This Image?*	Mengembangkan algoritma untuk merekonstruksi kembali (*recovery*) *prompt* asli yang digunakan untuk menghasilkan suatu gambar, khususnya pada Stable Diffusion untuk membuktikan bahwa informasi yang terkandung dalam *prompt* tersimpan begitu kuat dalam ruang laten gambar yang dihasilkan.	Metode *reverse Stable Diffusion* mampu memulihkan *prompt* dari gambar dengan akurasi tinggi, membuktikan hubungan yang sangat erat dan kuat antara *prompt* dan gambar yang dihasilkan. Namun, penelitian ini tidak mengevaluasi biaya komputasi atau penggunaan sumber daya yang terlibat dalam proses rekonstruksi *prompt*. Oleh karena itu, diperlukan analisis lanjutan untuk mengukur dampak kompleksitas *prompt* terhadap efisiensi komputasi pada proses generatif Stable Diffusion.
14	(Ghosal et al., 2024)	*IntCoOp: Intepretability-Aware Vision Language Prompt Tuning*	Mengembangkan metode *prompt tuning* otomatis untuk meningkatkan *interpretability*	Metode *interpretability-aware vision-language prompt tunning* dapat mencapai akurasi kinerja yang sebanding dengan

No.	Penulis, Tahun Penelitian	Judul Penelitian	Fokus Penelitian	Hasil Penelitian
			dari keputusan yang dibuat model.	*prompt tuning* lainnya, sekaligus meningkatkan *interpretability* model. Meskipun demikian, penelitian ini hanya menganalisis *trade-off* internal (akurasi vs interpretasi), tanpa mengevaluasi efisiensi komputasi *prompt* yang di *tune* pada fase inferensi. Oleh karena itu, diperlukan analisis lanjutan untuk menilai dampak kompleksitas *prompt* terhadap penggunaan sumber daya komputasi pada model generatif.
15	(He et al., 2025)	*Autmated Blac-Box Prompt Engineering for Personalized Text-to-Image Generation*	Mengembangkan metode otomatis untuk melakukan *prompt engineering* dalam skenario *text-to-image* yang dipersonalisasi.	Metode ini mampu menghasilkan *prompt* yang secara signifikan lebih efektif daripada *prompt* manual, menghasilkan kualitas visual yang lebih tinggi dan kesesuaian semantik yang lebih baik terhadap personalisasi yang diminta. Penelitian ini fokus pada otomatisasi dan kualitas *output*, dan tidak menyertakan analisis empiris mengenai dampak variasi kompleksitas *prompt* terhadap sumber daya (CPU/RAM) selama proses inferensi. Oleh karena itu, pengukuran efisiensi komputasi masih diperlukan.

No.	Penulis, Tahun Penelitian	Judul Penelitian	Fokus Penelitian	Hasil Penelitian
16	(Meng et al., 2022)	*SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations*	Mengembangkan kerangka kerja generatif baru untuk *guided image synthesis* dan *image editing* menggunakan persamaan diferensial Stokastik (SDE).	*SDEdit* berhasil menghasilkan gambar yang diedit secara realistis dan koheren dengan *prompt* baru, sambil mempertahankan struktur spasial dan identitas subjek dari gambar aslinya. Model menunjukkan kemampuan yang kuat dalam mencapai koherensi antara gambar hasil dan *prompt*. Namun, penelitian ini tidak secara eksplisit menganalisis *trade-off* antara kompleksitas *guidance* terhadap beban sumber daya *hardware host* selama proses SDE yang intensif. Maka dari itu, diperlukan penelitian untuk menganalisis pengaruh kompleksitas *prompt* terhadap efisiensi komputasi dan hubungannya.
17	(Saharia et al., 2022)	*Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding*	Mengenalkan Imagen, model difusi yang dirancang untuk meningkatkan kesesuaian semantik antara *prompt* dan gambar yang dihasilkan sekaligus mempertahankan tingkat fotorealisme yang tinggi	Penelitian ini menunjukkan bahwa peningkatan kompleksitas linguistik tidak selalu menghasilkan kualitas visual secara linear. Model dapat mengalami titik jenuh dalam memanfaatkan detail *prompt*. Meskipun berhasil mengkaji pengaruh pemahaman *prompt* terhadap kualitas dan kesesuaian semantik, penelitian ini tidak mengevaluasi dampaknya terhadap

No.	Penulis, Tahun Penelitian	Judul Penelitian	Fokus Penelitian	Hasil Penelitian
				efisiensi komputasi, sehingga diperlukan pendekatan evaluatif yang berorientasi pada aspek komputasi.
18	(Wang et al., 2004a)	*SD-Acc: Accelerating Stable Diffusion through Phase-Aware Sampling and Hardware CO-Optimizations*	Akselerasi proses inferensi Stable Diffusion dengan menggabungkan pendekatan *phase-aware sampling* dan *hardware co-optimization* untuk mengurangi *latency* dan biaya komputasi model tanpa melakukan *retraining* model.	*SD-Acc* berhasil menurunkan waktu inferensi secara signifikan yang dicapai dengan degradasi kualitas visual yang minimal. Namun, penelitian ini tidak menganalisis pengaruh kompleksitas *prompt* terhadap kualitas gambar maupun efisiensi komputasi sehingga hubungan antara struktur *prompt* dan beban sumber daya belum dieksplorasi. Oleh karena itu, diperlukan evaluasi respons model terhadap kompleksitas *prompt*.
19	(Lee et al., 2023)	*Holistic Evaluation of Text-to-Image Models*	Mengusulkan kerangka evaluasi holistik untuk model *text-to-image*, mencakup aspek kualitas visual, kesesuaian semantik, *robustness* terhadap variasi *prompt*, serta kegagalan generatif yang tidak terdeteksi oleh satu metrik tunggal.	Hasil penelitian menunjukkan bahwa respons model terhadap *prompt* bersifat tidak seragam, mengindikasikan adanya sensitivitas terhadap struktur dan formasi *prompt*. Namun, analisis terhadap kompleksitas *prompt* serta implikasinya terhadap aspek sumber daya *hardware host* belum dibahas secara mendalam. Oleh karena itu, diperlukan penelitian lanjutan untuk mengkaji hubungan antar

No.	Penulis, Tahun Penelitian	Judul Penelitian	Fokus Penelitian	Hasil Penelitian
				metrik evaluasi model generatif *text-to-image*.
20	(Mo et al., 2024)	*Dynamic Prompt Optimizing for Text-to-Image Generation*	Mengembangkan metode untuk mengoptimalkan *prompt* teks secara dinamis selama proses atau pra proses untuk mencapai kualitas gambar yang lebih tinggi.	Penelitian ini menunjukkan bahwa *prompt* yang secara semantik lebih kaya, terstruktur, atau dioptimalkan mampu meningkatkan kesesuaian dan kualitas hasil generasi pada model difusi. Namun, penelitian ini tidak mengevaluasi implikasi optimisasi *prompt* terhadap efisiensi komputasi. Oleh karena itu, diperlukan evaluasi yang lebih komprehensif pada aspek komputasi untuk memahami konsekuensi operasional dari peningkatan kompleksitas *prompt*.

2.12 Matriks Penelitian

Matriks penelitian diuraikan pada tabel 2.2.

Tabel 2. 2 Matriks Penelitian

No	Penelitian	Ruang Lingkup											
		Model Generatif *Text-to-Image* (Stable Diffusion)		Objek Wajah Manusia	Variasi *Prompt*	Parameter Uji							
						Kualitas Gambar			Interpretasi Semantik	Efisiensi Kompuasi			
		V1.4	V1.5			Face-FID/FID	IS	SSIM	CLIP Score	Waktu	CPU	GPU	RAM
1	(Heusel et al., 2017)	-	-	✓	-	✓	✓	-	-	-	-	-	-
2	(Ho et al., 2020)	-	-	✓	-	✓	✓	-	-	✓	-	✓	-
3	(Song et al., 2021)	-	-	✓	-	✓	✓	-	-	✓	-	-	-
4	(Rombach et al., 2022)	-	-	✓	✓	✓	✓	-	✓	✓	-	✓	-
5	(Salimans & Ho, 2022)	-	-	✓	-	✓	✓	-	-	✓	-	-	-
6	(Hu et al., 2022)	-	-	-	-	-	-	-	-	-	-	✓	-

No	Penelitian	Ruang Lingkup											
		Model Generatif *Text-to-Image* (Stable Diffusion)		Objek Wajah Manusia	Variasi *Prompt*	Parameter Uji							
						Kualitas Gambar			Interpretasi Semantik	Efisiensi Kompuasi			
		V1.4	V1.5			Face-FID/FID	IS	SSIM	CLIP Score	Waktu	CPU	GPU	RAM
7	(Podell et al., 2024)	-	-	✓	✓	✓	-	✓	✓	✓	-	✓	-
8	(Borji, 2023)	-	-	✓	✓	✓	-	-	✓	-	-	-	-
9	(Ruiz et al., 2023)	-	-	✓	✓	-	-	-	✓	-	-	-	-
10	(Croitoru et al., 2023)	-	-	✓	✓	✓	✓	✓	✓	✓	-	-	-
11	(Chefer et al., 2023)	-	-	✓	✓	-	-	-	✓	✓	-	-	-
12	(Feng et al., 2024)	-	-	✓	✓	✓	✓	-	✓	-	-	-	-
13	(Croitoru et al., 2024)	-	-	-	✓	✓	-	-	✓	-	-	-	-
14	(Ghosal et al., 2024)	-	-	-	✓	-	-	-	-	✓	-	-	-
15	(He et al., 2025)	-	-	✓	✓	-	-	-	✓	✓	-	-	-
16	(Meng et al., 2022)	-	-	✓	✓	✓	✓	-	✓	✓	-	-	-

No	Penelitian	Ruang Lingkup											
		Model Generatif *Text-to-Image* (Stable Diffusion)		Objek Wajah Manusia	Variasi *Prompt*	Parameter Uji							
						Kualitas Gambar			Interpretasi Semantik	Efisiensi Kompuasi			
		V1.4	V1.5			Face-FID/FID	IS	SSIM	CLIP Score	Waktu	CPU	GPU	RAM
17	(Saharia et al., 2022)	-	-	-	✓	✓	✓	-	-	-	-	-	-
18	(Wang et al., 2004a)	-	-	-	-	✓	✓	-	-	✓	-	✓	-
19	(Lee et al., 2023)	-	-	-	✓	✓	✓	-	✓	-	-	-	-
20	(Mo et al., 2024)	-	-	-	✓	-	-	-	✓	-	-	-	-
21	Peneitian Ini	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

2.13 *Gap Analysis*

Berdasarkan tabel *state-of-the-art* dan matriks penelitian, literatur yang ada menunjukkan kemajuan signifikan dalam pengembangan model difusi *text-to-image,* baik dari aspek kualitas visual, kontrol semantik, maupun efisiensi komputasi. Namun, sebagian besar studi tersebut masih berfokus pada kualitas *output* atau *interpretability prompt,* sementara pengaruh kompleksitas *prompt* terhadap efisiensi komputasi dan respons model belum dianalisis secara kuantitatif dan terstruktur. Selain itu, evaluasi perbedaan respons antar versi Stable Diffusion v1.4 dan v1.5 terhadap kompleksitas *prompt,* khususnya pada sintesis wajah manusia dan lingkungan dengan sumber daya terbatas masih jarang dilakukan. Oleh karena itu, penelitian ini mengisi *gap* tersebut dengan mengevaluasi pola *trade-off* dan respons model pada aspek kualitas gambar, interpretasi semantik, dan efisiensi komputasi berdasarkan kompleksitas *prompt*.