

BAB I PENDAHULUAN

1.1 Latar Belakang

Teknologi bantu navigasi berbasis kecerdasan buatan memiliki potensi besar dalam mendukung mobilitas serta kemandirian individu tunanetra [1], [2]. Namun, sistem pemahaman visi-bahasa modern menghadapi tantangan teknis berat di lingkungan nyata, seperti degradasi gambar akibat guncangan kamera (*motion blur*), pencahayaan buruk, dan kesalahan pembingkai objek karena pengguna tidak dapat melihat layar bidik [3]. Meskipun arsitektur mutakhir seperti BLIP mampu menjembatani teks dan citra dengan optimal [4], kinerjanya menurun drastis saat memproses foto otentik yang penuh derau. Hal ini menjadi kendala kritis bagi pengembangan alat bantu mandiri yang harus tetap menjaga privasi penggunanya [5].

Keterbatasan tersebut terlihat jelas pada kumpulan data aksesibilitas VizWiz yang bersumber langsung dari tangkapan pengguna tunanetra [6]. Foto-foto otentik di dalamnya sering mengalami kerusakan visual ekstrem yang memicu munculnya bias dan diidentifikasi sebagai "Paradoks Kualitas" [3]. Kesalahan model dalam memproses gambar berderau menuntut adanya teknik deteksi efek *blur* serta restorasi yang lebih presisi agar informasi yang diberikan tidak keliru [7]. Selama ini, sebagian besar penelitian berusaha mengatasi masalah tersebut melalui modifikasi arsitektur yang kompleks (*model-centric*) [8], [9]. Padahal, tanpa kurasi data yang tepat, pemaksaan model untuk belajar langsung dari data yang rusak justru akan mengacaukan kemampuan penalaran semantik sistem secara keseluruhan [10].

Penelitian terdahulu juga menunjukkan bahwa teknik pelatihan ulang (*fine-tuning*) konvensional membutuhkan kapasitas memori dan daya komputasi yang sangat besar [11]. Untuk mengatasi hambatan tersebut, pendekatan optimasi parameter seperti *Low-Rank Adaptation* (LoRA) mampu meningkatkan efisiensi dengan mengurangi jumlah parameter terlatih hingga 10.000 kali lipat, serta

menurunkan kebutuhan memori secara drastis tanpa menambah latensi inferensi [11]. Di sisi lain, analisis dalam domain frekuensi menggunakan *Fast Fourier Transform* (FFT) dapat menjadi solusi objektif untuk mendeteksi tingkat *blur* spasial gambar secara otomatis [12]. Oleh karena itu, pengujian ketangguhan model terhadap data berderau menjadi agenda penting dalam pengembangan sistem ini [13].

Berdasarkan celah permasalahan tersebut, penelitian ini mengajukan solusi kebaruan berupa alur kerja *data-centric* yang menggabungkan penyaringan FFT dengan metode *Parameter-Efficient Fine-Tuning* (PEFT) LoRA [11]. Pendekatan ini dirancang untuk mengeliminasi bias "Paradoks Kualitas" secara efisien menggunakan pengoptimal AdamW [14] guna menjaga konsumsi memori komputasi (VRAM) pada perangkat keras tingkat konsumen. Kinerja model selanjutnya divalidasi menggunakan metrik evaluasi leksikal dan semantik mutakhir seperti BERTScore [15] serta CLIPScore [16]. Hasil penelitian ini diharapkan mampu mengoptimalkan keandalan sistem aksesibilitas visual yang benar-benar akurat dan mandiri bagi individu tunanetra.

1.2 Rumusan Masalah

Berdasarkan latar belakang tersebut, penelitian ini berfokus pada kesulitan model visi-bahasa dalam memproses foto berkualitas rendah yang sering diambil oleh pengguna tunanetra. Oleh karena itu, rumusan masalah yang akan diselesaikan adalah sebagai berikut:

1. Bagaimana efektivitas kombinasi penyaringan kekaburan *Fast Fourier Transform* (FFT) dan metode *Parameter-Efficient Fine-Tuning* (PEFT) *Low-Rank Adaptation* (LoRA) dalam mengeliminasi bias "Paradoks Kualitas" pada dataset VizWiz?
2. Seberapa besar peningkatan kualitas *caption* yang dihasilkan model BLIP setelah dioptimalkan menggunakan pendekatan tersebut, jika dievaluasi menggunakan metrik CIDEr, BERTScore, ROUGE-L, dan BLEU-4?
3. Bagaimana efisiensi penggunaan memori komputasi (VRAM) dan stabilitas konvergensi saat melatih model visi-bahasa menggunakan arsitektur LoRA dan pengoptimal AdamW pada perangkat keras tingkat konsumen?

1.3 Batasan Masalah

Berdasarkan rumusan masalah yang telah ditetapkan, penelitian ini memiliki beberapa batasan masalah agar pembahasan lebih terfokus, tidak meluas, dan parameter yang digunakan dapat terukur dengan jelas. Adapun batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Batasan Dataset dan Evaluasi: Penelitian ini secara khusus menggunakan dataset aksesibilitas VizWiz-Captions. Evaluasi kinerja model hanya dilakukan pada data validasi sebanyak 7.750 gambar. Set pengujian resmi sebanyak 8.000 gambar tidak disertakan dalam pengukuran karena anotasi *ground-truth* dari pembuat dataset tidak tersedia secara publik karena ditahan untuk keperluan kompetisi global.
2. Batasan Arsitektur dan Modifikasi Model: Penelitian difokuskan pada penggunaan arsitektur model dasar BLIP (ViT-B/16 sebagai *visual encoder*

dan dekoder teks berbasis BERT). Penelitian ini tidak melakukan *fine-tuning* pada seluruh parameter model sekitar 224 juta parameter, melainkan membatasinya menggunakan teknik *Parameter-Efficient Fine-Tuning* (PEFT) *Low-Rank Adaptation* (LoRA) yang hanya melatih sekitar 1.179.648 parameter (0,52%) untuk mencegah beban komputasi yang berlebihan.

3. Batasan Prapemrosesan Gambar: Konfigurasi masukan gambar diasumsikan konstan dan diatur secara baku pada resolusi 384×384 piksel menggunakan metode interpolasi Lanczos dan *letterbox padding*. Konfigurasi ini ditetapkan agar sesuai dengan pengaturan default *Vision Transformer* (ViT) pada model BLIP.
4. Batasan Metrik Evaluasi: Pengukuran keberhasilan model dibatasi pada evaluasi ablasi internal menggunakan metrik CIDEr, BERTScore, ROUGE-L, BLEU-4, dan CLIPScore. Penelitian ini tidak menyertakan perbandingan dengan arsitektur *State-of-the-Art* (SOTA) eksternal yang berbeda seperti BLIP-2 atau OFA, demi menjaga ketelitian ilmiah pada dampak kurasi data.
5. Batasan Lingkungan Perangkat Keras: Pelatihan model dibatasi pada perangkat keras tingkat konsumen berupa GPU NVIDIA GeForce RTX 3050 dengan VRAM maksimal 6GB.
6. Batasan Kondisi Gambar Ekstrem: Penelitian ini tidak menargetkan penyelesaian kesalahan prediksi pada gambar yang mengalami oklusi fisik ekstrem seperti lensa kamera tertutup penuh atau gambar yang diambil dalam kondisi gelap total tanpa cahaya. Pada kondisi tersebut, sinyal optik diasumsikan telah hilang secara permanen dan tidak dapat dipulihkan. Penelitian ini juga berjalan secara mandiri tanpa bantuan modul *Optical Character Recognition* (OCR) tambahan.
7. Batasan Akses API dan Kuota Token: Penelitian ini menggunakan Google Gemini API pada tingkatan akses gratis (*free tier*). Oleh karena itu, penggunaan API dibatasi oleh kuota *token* harian dan *rate limit* (*requests*

per minute). Fokus penggunaan API dibatasi pada analisis semantik dan validasi hasil pada sampel data terpilih, yang dioptimasi melalui mekanisme *batching* dan *caching* lokal agar tetap akurat meskipun dilakukan di luar pemrosesan *real-time* berskala masif.

1.4 Tujuan Penelitian

Tujuan yang akan dicapai oleh peneliti dalam penelitiannya adalah menghasilkan solusi yang terukur untuk mengatasi permasalahan degradasi gambar pada sistem visi-bahasa. Secara spesifik, tujuan penelitian ini dijabarkan sebagai berikut:

1. Menghasilkan *pipeline* kurasi data yang divalidasi secara statistik dengan menggabungkan penyaringan kekaburan *Fast Fourier Transform* (FFT) dan metode *Parameter-Efficient Fine-Tuning* (PEFT) *Low-Rank Adaptation* (LoRA) untuk mengeliminasi bias "Paradoks Kualitas" pada dataset aksesibilitas VizWiz.
2. Mengukur peningkatan kualitas *caption* yang dihasilkan model BLIP setelah dioptimalkan menggunakan pendekatan tersebut, dengan menggunakan metrik evaluasi CIDEr, BERTScore, ROUGE-L, BLEU-4, dan CLIPScore.
3. Menganalisis efisiensi penggunaan memori komputasi (VRAM) dan stabilitas konvergensi pelatihan model visi-bahasa menggunakan arsitektur LoRA dan pengoptimal AdamW pada perangkat keras tingkat konsumen.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan kontribusi nyata yang dapat dimanfaatkan langsung oleh berbagai pihak, baik secara praktis maupun teoritis. Adapun manfaat dari penelitian ini adalah sebagai berikut:

1. Manfaat Praktis (Bagi Pengguna dan Pengembang Teknologi)

- a. Meningkatkan keselamatan dan navigasi tunanetra: Sistem yang dioptimalkan ini dirancang untuk memastikan pemahaman semantik yang tangguh, sehingga pengguna tunanetra berpotensi menerima deskripsi navigasi yang lebih akurat meskipun gambar diambil dalam kondisi kamera yang menantang seperti sangat buram.
- b. Alternatif navigasi mandiri: Model ini dirancang untuk mengenali objek dalam kondisi visual terbatas tanpa memerlukan tambahan modul *Optical Character Recognition* (OCR), sehingga berpotensi digunakan sebagai alat bantu navigasi yang mandiri di dunia nyata.
- c. Implementasi perangkat seluler yang terjangkau: Efisiensi komputasi yang dicapai dalam penelitian ini menunjukkan bahwa kecerdasan buatan bantu (*assistive AI*) tingkat lanjut tidak memerlukan biaya server yang mahal. Hal ini berpotensi mendukung penciptaan teknologi navigasi bantuan waktu nyata (*real-time*) yang dapat berjalan langsung pada perangkat seluler.

2. Manfaat Teoritis (Bagi Akademisi dan Peneliti Selanjutnya)

- a. Solusi empiris untuk "Paradoks Kualitas": Penelitian ini memberikan kontribusi empiris dengan membuktikan secara statistik bahwa performa model dapat diseimbangkan pada gambar yang sangat buram maupun tajam melalui penyaringan derau visual yang tidak dapat dipulihkan.
- b. Efisiensi pelatihan model kompleks: Penelitian ini menunjukkan bahwa menggabungkan pelatihan hemat memori menggunakan metode *Parameter-Efficient Fine-Tuning* (PEFT) LoRA dengan

kurasi data lebih efisien dibandingkan *fine-tuning* konvensional secara penuh. Hal ini menjembatani kesenjangan antara kebutuhan akurasi model yang tinggi dengan keterbatasan perangkat keras tingkat konsumen.

- c. Landasan untuk pengembangan multimodal lanjutan: Arsitektur yang hemat memori ini membuka peluang bagi penelitian selanjutnya untuk mengintegrasikan antarmuka interaksi suara serta *Large Language Models* (LLM) guna menciptakan sistem percakapan interaktif yang lebih baik bagi pengguna tunanetra.
- d. Ekspansi ke domain spesifik lainnya: Metode pengoptimalan berpusat pada data yang dikembangkan dalam penelitian ini dapat disesuaikan dan diperluas untuk teknologi bantuan di domain lain yang lebih terspesialisasi, seperti *Visual Question Answering* (VQA) dan analisis citra medis.



1.6 Sistematika Penulisan

Sistematika penulisan ini menguraikan secara garis besar kerangka dan gambaran isi dari masing-masing bab yang menyusun dokumen penelitian ini. Pembahasan disusun menjadi lima bab utama dengan rincian sebagai berikut:

BAB I PENDAHULUAN, berisi latar belakang masalah terkait keterbatasan model visi-bahasa dalam memproses foto berkualitas rendah dari dataset aksesibilitas, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, serta sistematika penulisan.

BAB II TINJAUAN PUSTAKA, berisi kajian penelitian-penelitian terdahulu yang relevan di bidang teknologi aksesibilitas dan sistem *image captioning*, serta landasan teori yang digunakan sebagai fondasi penelitian meliputi arsitektur model BLIP, konsep penyaringan *Fast Fourier Transform* (FFT), metode *Low-Rank Adaptation* (LoRA), dan metrik evaluasi yang digunakan.

BAB III METODE PENELITIAN, berisi uraian tahapan penelitian yang dilakukan secara sistematis, mencakup deskripsi dataset VizWiz-Captions, proses kurasi dan prapemrosesan data, spesifikasi arsitektur model, serta konfigurasi pelatihan yang digunakan dalam eksperimen.

BAB IV HASIL DAN PEMBAHASAN, berisi pemaparan hasil evaluasi empiris model yang telah dioptimalkan, mencakup analisis konvergensi pelatihan, perbandingan metrik kinerja antara model *baseline* dan model yang diusulkan melalui studi ablasi, serta analisis dampak kualitas gambar terhadap performa model secara statistik.

BAB V PENUTUP, berisi kesimpulan yang dirangkum dari hasil penelitian secara keseluruhan serta saran bagi peneliti selanjutnya terkait pengembangan sistem lebih lanjut.