IMPLEMENTASI SISTEM PENCARIAN GAMBAR BERBASIS TEKS BAHASA INDONESIA DENGAN FINE-TUNED CLIP TEXT ENCODER DAN VISION TRANSFORMER IMAGE ENCODER

SKRIPSI

Muhammad Nur Irfan
2207411056

PROGRAM STUDI TEKNIK INFORMATIKA
JURUSAN TEKNIK INFORMATIKA DAN KOMPUTER
POLITEKNIK NEGERI JAKARTA
2026

IMPLEMENTASI SISTEM PENCARIAN GAMBAR BERBASIS TEKS BAHASA INDONESIA DENGAN FINE-TUNED CLIP TEXT ENCODER DAN VISION TRANSFORMER IMAGE ENCODER

SKRIPSI

Dibuat untuk Melengkapi Syarat-Syarat yang Diperlukan untuk Memperoleh Diploma Empat Politeknik

Muhammad Nur Irfan

2207411056

PROGRAM STUDI TEKNIK INFORMATIKA

JURUSAN TEKNIK INFORMATIKA DAN KOMPUTER

POLITEKNIK NEGERI JAKARTA

2026

SURAT PERNYATAAN BEBAS PLAGIARISME

Saya yang bertanda tangan di bawah ini:

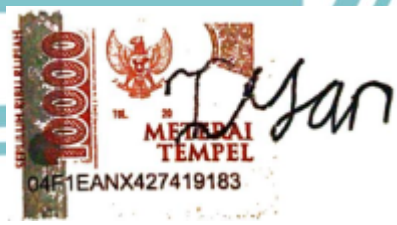
Nama	: Muhammad Nur Irfan
NIM	: 2207411056
Jurusan/ProgramStudi	: T.Informatika dan Komputer / Teknik Informatika
Judul skripsi	:IMPLEMENTASI SISTEM PENCARIAN GAMBAR BERBASIS TEKS BAHASA INDONESIA DENGAN FINE-TUNED CLIP TEXT ENCODER DAN VISION TRANSFORMER IMAGE ENCODER

Menyatakan dengan sebenarnya bahwa skripsiini benar-benar merupakan hasil karya saya sendiri, bebas dari peniruan terhadap karya dari orang lain. Kutipan pendapat dan tulisan orang lain ditunjuk sesuai dengan cara-cara penulisan karya ilmiah yang berlaku.

Apabila di kemudian hari terbukti atau dapat dibuktikan bahwa dalam skripsi ini terkandung cirri-ciri plagiat dan bentuk-bentuk peniruan lain yang dianggap melanggar peraturan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Jakarta, 9 Juni 2026

Yang membuat pernyataan



Muhammad Nur Irfan

NIM 2207411056

SURAT PERNYATAAN PERSETUJUAN PUBLIKASI SKRIPSI UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademik Politeknik Negeri Jakarta, saya bertanda tangan dibawah ini :

Nama : Muhammad Nur Irfan
NIM : 2207411056
Jurusan/ProgramStudi : T.Informatika dan Komputer / Teknik Informatika

Demi pengembangan ilmu pengetahuan , menyetujui untuk memberikan kepada Politeknik Negeri Jakarta Hak Bebas Royalti Non-Eksklusif atas karya ilmiah sayayang berjudul :

IMPLEMENTASI SISTEM PENCARIAN GAMBAR BERBASIS TEKS BAHASA INDONESIA DENGAN FINE-TUNED CLIP TEXT ENCODER DAN VISION TRANSFORMER IMAGE ENCODER

Beserta perangkat yang ada (jika diperlukan). Dengan Hak Bebas Royalti Non-Eksklusif ini Politeknik Negeri Jakarta Berhak menyimpan, mengalihmediakan/formatkan, mengelola dalam bentuk pangkalan data (database), merawat, dan mempublikasikan skripsi saya tanpa meminta izin dari saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyatan ini saya buat dengan sebenarnya.

Jakarta, 9 Juni 2026

Yang membuat pernyataan

Muhammad Nur Irfan

2207411056

LEMBAR PENGESAHAN

Skripsi diajukan oleh :

Nama : Muhammad Nur Irfan
NIM 2207411056
Program Studi : Teknik Informatika
Judul Skripsi : Implementasi Sistem Pencarian Gambar Berbasis Teks Bahasa Indonesia Dengan *Fine-Tuned* CLIP *Text Encoder* dan Vision Transformer *Image Encoder*

Telah diuji oleh tim penguji dalam Sidang Skripsi pada hari Senin, Tanggal 22, Bulan Juni, Tahun 2026 dan dinyatakan **LULUS**.

Disahkan oleh

Pembimbing I : Euis Oktavianti, S.Si., M.T.I (. .....)

Penguji I : Dr. Dewi Yanti Liliana, S.Kom., M.Kom. (. .....)

Penguji II : Dr. Indra Hermawan, M.Kom. (. .....)

Penguji III : Anggi Mardiyono, S.Kom.,M.Kom. (. .....)

Mengetahui:
Jurusan Teknik Informatika dan Komputer
Ketua

Dr. Anita Hidayati, S.Kom., M.Kom.
NIP. 197908032003122003

KATA PENGANTAR

Assalamualaikum Warahmatullahi Wabarakatuh, Puji Syukur saya panjatkan kepada Tuhan Yang Maha Esa, yang telah memberikan karunia dan rahmat-Nya kepada penulis, sehingga penulis dapat menyelesaikan skripsi yang berjudul "IMPLEMENTASI SISTEM PENCARIAN GAMBAR BERBASIS TEKS BAHASA INDONESIA DENGAN FINE-TUNED CLIP TEXT ENCODER DAN VISION TRANSFORMER IMAGE ENCODER" sebagai salah satu syarat untuk mencapai gelar Sarjana Terapan Program Studi Teknik Informatika Politeknik Negeri Jakarta.

Penulis sangat menyadari bahwa tanpa bantuan dan bimbingan dari berbagai pihak, sangatlah sulit bagi penulis untuk menyelesaikan skripsi ini. Dengan demikian, penulis ingin mengucapkan terima kasih kepada pihak-pihak yang terlibat. Secara khusus, penulis menyampaikan ucapan terimakasih kepada:

1. Ibu Dr. Anita Hidayati, S. Kom., M. Kom., selaku Ketua Jurusan Teknik Informatika dan Komputer.
2. Ibu Euis Oktavianti, S.Si., M.TI., selaku Kepala Porgram Studi Teknik Informatika sekaligus Dosen Pembimbing yang sudah bersedia meluangkan waktu untuk membimbing, mengarahkan dan menyemangati dalam proses penyelesaian skripsi ini.
3. Seluruh Bapak/Ibu dosen yang sudah mendidik penulis sehingga menjadi pribadi yang lebih baik
4. Kedua orang tua penulis yang senantiasa mendukung, memberikan tempat cerita, memberikan doa dan semangat serta kasih sayang tanpa henti kepada penulis
5. Azzamahdy Ahmad Rafiqri selaku saudara penulis yang memberikan semangat, nasihat dan tempat untuk berdiskusi kepada penulis
6. Sahabat Prajurit selaku teman seperjuangan dan teman teman lainnya yang memberikan dukungan, solusi dan menemani penulis dalam penyelesaian skripsi ini

Akhir kata, penulis berharap skripsi yang ditulis ini dapat memberikan manfaat kepada berbagai pihak. Penulis menyadari bahwa skripsi ini jauh dari kata sempurna, oleh karena itu, penulis mengharapkan segala bentuk kritik, saran, dan masukan yang membangun yang dapat memperbaiki serta menyempurnakan skripsi ini.

Wassalamualaikum Warahmatullahi Wabarakatuh.

Jakarta, 9 Juni 2026

Muhammad Nur Irfan

IMPLEMENTASI SISTEM PENCARIAN GAMBAR BERBASIS TEKS BAHASA INDONESIA DENGAN FINE-TUNED CLIP TEXT ENCODER DAN VISION TRANSFORMER IMAGE ENCODER

ABSTRAK

Sistem pencarian gambar konvensional umumnya masih bergantung pada metadata manual yang kurang mampu menangkap makna semantik visual. Pendekatan cross-modal retrieval menggunakan model CLIP menawarkan solusi, namun performanya menurun pada bahasa non-Inggris seperti Bahasa Indonesia. Penelitian ini melakukan studi komparatif terkontrol terhadap tiga konfigurasi text encoder pada arsitektur CLIP (ViT-B/32), yaitu CLIP standar (bawaan), CLIP dengan IndoBERT, dan CLIP dengan XLM-RoBERTa untuk mengevaluasi efektivitas contrastive learning berbasis Bahasa Indonesia menggunakan dataset Flickr8K dan Flickr30K Indonesia. Hasil evaluasi metrik retrieval dua arah menunjukkan bahwa CLIP standar secara konsisten mencatatkan performa tertinggi pada kedua dataset (rata-rata Recall@10 mencapai 90,52% pada Flickr8K), diikuti oleh CLIP XLM-RoBERTa, sementara CLIP IndoBERT menghasilkan performa terendah. Temuan ini membuktikan bahwa keselarasan ruang embedding (embedding space alignment) sejak fase pre-training jauh lebih mendominasi performa dibandingkan faktor spesialisasi linguistik text encoder monolingual dalam kondisi data terbatas. Model terbaik (CLIP standar) berhasil diimplementasikan ke dalam sistem Smart Gallery berbasis web menggunakan FastAPI dan NextJS. Pengujian usability menggunakan System Usability Scale (SUS) menghasilkan skor rata-rata 91,36 (kategori Excellent, Grade A), menegaskan bahwa aplikasi sangat layak dan mudah dioperasikan oleh pengguna.

Kata Kunci: Bahasa Indonesia, *Contrastive Learning*, *Cross-Modal Retrieval*, CLIP, IndoBERT, Pencarian Gambar, *Smart Gallery*, XLM-RoBERTa.

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1
PENDAHULUAN

1.1. Latar Belakang

Bahasa Indonesia merupakan bahasa dengan jumlah penutur lebih dari 198 juta jiwa dan menjadi salah satu bahasa dengan penetrasi internet tertinggi di Asia Tenggara, dengan 185,3 juta pengguna internet aktif pada tahun 2024 (Kemp, 2024). Pertumbuhan pengguna digital ini mendorong peningkatan produksi konten visual yang diproduksi dan dikonsumsi, mulai dari galeri pribadi hingga platform berbagi media. Namun, sistem pencarian gambar yang umum digunakan saat ini masih bergantung pada metadata manual seperti nama berkas dan tag subjektif yang tidak mampu menangkap makna semantik dari konten visual (Aske & Giardinetti, 2023). Kesenjangan ini menyulitkan pengguna berbahasa Indonesia untuk menemukan gambar yang relevan menggunakan deskripsi teks dalam bahasa Indonesia.

Pendekatan cross-modal retrieval berbasis *deep learning*, khususnya CLIP (Radford et al., 2021), telah terbukti efektif dalam menghubungkan representasi gambar dan teks dalam satu ruang semantik melalui *contrastive learning*. Namun, Dataset pelatihan CLIP didominasi oleh deskripsi berbahasa Inggris, sehingga bahasa-bahasa yang kurang terwakili termasuk bahasa-bahasa Asia Tenggara mengalami kesenjangan performa dibandingkan Bahasa Inggris (Santos et al., 2023).

Beranjak dari kebutuhan akan sistem *cross modal retrieval* yang mampu memahami konteks Bahasa Indonesia secara lebih mendalam, penelitian ini melakukan studi komparatif terhadap tiga pendekatan text encoder, yaitu CLIP, CLIP dengan XLM-RoBERTa, dan CLIP dengan IndoBERT untuk mengevaluasi efektivitas model dengan pendekatan *contrastive learning*. Oleh karena itu, hasil penelitian ini diharapkan mampu memberikan landasan empiris tentang efektivitas komparatif text encoder untuk arsitektur CLIP dalam konteks cross-modal retrieval berbahasa Indonesia, sekaligus menghadirkan sistem pencarian gambar yang

memungkinkan pengguna melakukan pencarian secara intuitif menggunakan teks Bahasa Indonesia sehari-hari tanpa bergantung pada tag atau nama file.

1.2 Perumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, rumusan masalah pada penelitian ini adalah

1. Bagaimana perbandingan performa konfigurasi text encoder CLIP, CLIP dengan XLM-RoBERTa , dan CLIP dengan IndoBERT dengan *contrastive learning* untuk tugas pencarian gambar dengan kueri berbahasa Indonesia?
2. Bagaimana mengimplementasikan model *text encoder* terbaik ke dalam sistem Smart Gallery untuk mendukung pencarian gambar multimodal berbahasa Indonesia?

1.3 Batasan Masalah

Dalam penelitian ini, batasan masalah diperlukan agar penelitian yang dilakukan tetap terarah dan sistematis sesuai dengan perumusan masalah. Berikut adalah beberapa batasan masalah dalam penelitian ini:

1. Penelitian ini menggunakan dua dataset: Flickr8K Indonesia sebagai eksperimen pendahuluan dan Flickr30K Indonesia sebagai dataset utama evaluasi.
2. Penelitian ini menggunakan arsitektur Vision Transformer ViT-B/32 sebagai image encoder yang dibekukan (frozen) pada seluruh konfigurasi model.
3. Penelitian ini membandingkan tiga konfigurasi *text encoder*: CLIP standar, CLIP dengan XLM-RoBERTa, dan CLIP dengan IndoBERT sebagai *text encoder*.
4. Seluruh konfigurasi model dilatih menggunakan metode *contrastive learning* dengan fungsi loss InfoNCE, *batch size*, *optimizer* dan jumlah *epoch* yang identik, serta *learning rate* yang disesuaikan per arsitektur.
5. Konfigurasi freeze layer pada *text encoder* disesuaikan per arsitektur sehingga jumlah *layer trainable* setara di angka 6 *layer* untuk setiap model BERT, meskipun proporsi *freeze* berbeda akibat perbedaan kedalaman arsitektur.

6. Penelitian ini menggunakan metrik Recall@K dan MRR (*Mean Reciprocal Rank*) untuk mengukur efektivitas *cross-modal retrieval* secara kuantitatif.
7. Sistem Smart Gallery yang dikembangkan mendukung pencarian gambar berbasis kueri teks dan kueri gambar, tanpa fitur penambahan gambar oleh pengguna.

1.4 Tujuan dan Manfaat

1.4.1 Tujuan

Berdasarkan latar belakang dan rumusan masalah, tujuan utama dari penelitian ini adalah membandingkan efektivitas konfigurasi *text encoder* CLIP, CLIP dengan XLM-RoBERTa, dan CLIP dengan IndoBERT dengan *contrastive learning* untuk tugas *cross-modal retrieval* berbahasa Indonesia dan mengimplementasi model ke dalam sistem pencarian gambar.

1.4.2 Manfaat

Adapun manfaat dari penelitian ini, dapat dijabarkan sebagai berikut:

1. Memudahkan proses pencarian gambar menggunakan bahasa alami tanpa ketergantungan pada *metadata*, *tag*, atau nama file manual.
2. Mempercepat proses pencarian gambar dengan memberikan pencarian gambar berdasarkan fitur visual gambar.

1.5 Sistematika Penulisan

Untuk mempermudah penulisan penelitian untuk penulis dan pembaca, penulisan disusun secara sistematis dengan struktur penulisan berikut ini:

BAB I PENDAHULUAN

Bab ini berisi uraian mengenai latar belakang penelitian, perumusan masalah, batasan masalah, tujuan dan manfaat penelitian, serta sistematika penulisan yang menjadi gambaran umum arah dan ruang lingkup penelitian.

BAB II LANDASAN TEORI

Bab ini memuat landasan teori yang mendukung penelitian, meliputi konsep smart gallery, metadata gambar, *cross-modal retrieval*, *embedding* teks dan gambar, Contrastive Language–Image Pre-trained (CLIP), metode CRISP-DM, Software Development Life Cycle (SDLC), serta kajian penelitian terdahulu yang relevan sebagai dasar ilmiah penelitian.

BAB III PEMBAHASAN

Bab ini membahas metodologi dan perancangan penelitian, meliputi rancangan penelitian, tahapan penelitian, sumber data, teknik pengumpulan data, teknik analisis data, perancangan model, evaluasi model, implementasi sistem smart gallery berbasis web, jadwal pelaksanaan penelitian, serta rincian biaya yang diperlukan dalam penelitian.

BAB IV HASIL DAN ANALISIS

Bab ini berisi hasil implementasi dan pengujian sistem yang telah dikembangkan. Pembahasan meliputi hasil fine-tuning model, hasil pengujian pencarian gambar berbasis teks, evaluasi performa sistem menggunakan metrik recall dan serta analisis terhadap efektivitas sistem dalam mendukung pencarian gambar berbasis *cross-modal retrieval*.

BAB V PENUTUP

Bab ini memuat kesimpulan yang diperoleh berdasarkan hasil penelitian serta saran yang dapat digunakan sebagai acuan untuk pengembangan sistem dan penelitian selanjutnya.

BAB 2
TINJAUAN PUSTAKA

2.1 Smart Gallery

Penyimpanan digital merupakan proses penyimpanan, pengelolaan, dan pemeliharaan data dalam bentuk digital menggunakan media berbasis perangkat keras dan perangkat lunak, baik secara lokal maupun terdistribusi melalui jaringan internet. Galeri digital merupakan platform berbasis sistem informasi yang digunakan untuk menyimpan, mengelola, menampilkan, dan mendistribusikan koleksi gambar atau karya visual dalam bentuk digital. Konsep smart gallery muncul sebagai evolusi dari galeri digital konvensional dengan mengintegrasikan teknologi *computer vision* dan *multimodal learning*. Smart gallery dirancang untuk memahami konten visual gambar secara otomatis dan menyediakan mekanisme pencarian yang lebih intuitif, seperti pencarian berbasis teks deskriptif. Sistem ini memanfaatkan model pembelajaran mendalam untuk mengekstraksi representasi visual dan menghubungkannya dengan representasi teks dalam satu ruang semantik yang sama.

2.1.1 Metadata

Metadata merupakan data yang digunakan untuk menjelaskan, memberikan konteks, atau informasi tambahan tentang data lain. dalam konteks penelitian ini, Metadata gambar merupakan informasi tambahan yang melekat pada berkas gambar dan digunakan untuk mendeskripsikan karakteristik non-visual, seperti nama file, waktu pengambilan, lokasi, dan deskripsi teks. Pada sistem galeri digital konvensional, metadata umumnya bersifat manual dan subjektif, sehingga kurang mampu merepresentasikan makna visual secara semantik. Penelitian terkini menunjukkan bahwa metadata berbasis konten visual yang dihasilkan secara otomatis melalui model *deep learning* mampu meningkatkan kualitas pencarian gambar secara signifikan (Aske & Giardinetti, 2023). Dalam penelitian ini, metadata semantik dihasilkan melalui proses *embedding* visual dan teks.

2.2 Cross-Modal Retrieval

cross-modal retrieval didefinisikan sebagai pendekatan temu kembali informasi yang bertujuan untuk mengatasi kesenjangan semantik antar modalitas dengan memungkinkan pencarian data pada satu modalitas menggunakan kueri dari modalitas lain, seperti pencarian gambar berbasis teks atau pencarian teks berbasis gambar (Han et al, 2023). Prinsip dasar pencarian gambar berbasis teks dalam *cross-modal retrieval* adalah melakukan transformasi teks kueri dan data gambar ke dalam representasi vektor yang dapat dibandingkan secara langsung.

2.2.1 *Embedding* pada Cross-Modal Retrieval

Embedding merupakan representasi numerik berdimensi tinggi yang digunakan untuk memetakan data kompleks, seperti gambar dan teks, ke dalam ruang vektor yang sama. Dalam konteks *cross-modal retrieval*, *embedding* berfungsi sebagai jembatan antara modalitas visual dan tekstual sehingga keduanya dapat dibandingkan secara langsung. Menurut Wang et al. (2022), keberhasilan sistem *cross-modal retrieval* sangat bergantung pada kualitas *embedding* yang mampu menangkap keselarasan semantik antar gambar dan teks.

2.3 Contrastive Learning

Contrastive learning merupakan pendekatan pembelajaran representasi yang melatih model untuk memetakan data semantis serupa ke posisi yang berdekatan dalam ruang *embedding*, sekaligus mendorong data yang tidak relevan agar saling berjauhan (Chen et al., 2020). Pendekatan ini bekerja dengan mendekatkan pasangan data yang memiliki hubungan semantis (positive pairs) dan menjauhkan pasangan data yang tidak memiliki hubungan semantis (negarive pairs), lalu *encoder* dioptimasi agar relasi tersebut terpetakan dengan jelas di dalam ruang vektor. Proses pelatihan ini dioptimalkan menggunakan fungsi *loss* InfoNCE yang bekerja dengan cara menemukan pasangan positif yang tepat di antara seluruh kandidat melalui evaluasi skor *cosine similarity* yang disesuaikan oleh parameter temperature. Penggunaan temperature yang rendah akan mempertajam sensitivitas

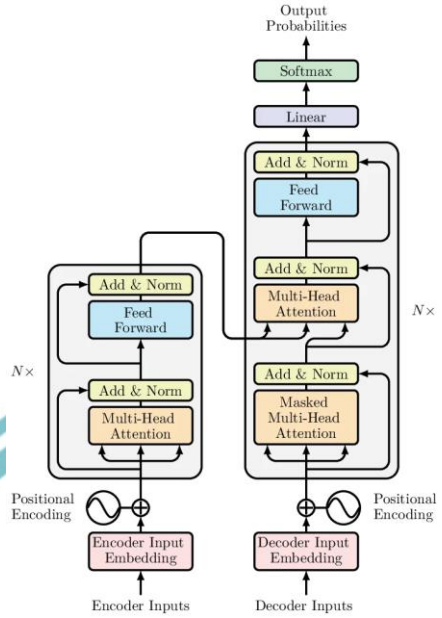
model dalam membedakan pasangan data, meskipun nilai yang terlampau kecil dapat mengganggu stabilitas pelatihan akibat gradien yang terlalu tajam.

2.4 Arsitektur Model

Penelitian ini melibatkan tiga konfigurasi model yang dibandingkan secara langsung untuk tugas pencarian gambar berbasis kueri Bahasa Indonesia. Ketiga konfigurasi tersebut adalah CLIP standar, CLIP dengan XLM-RoBERTa , dan CLIP dengan IndoBERT sebagai *text encoder*.

2.4.1 Transformer dan self attention

Transformer merupakan arsitektur jaringan saraf yang digunakan untuk mengubah urutan token input menjadi representasi vektor kontekstual yang kaya makna, dan setiap lapisannya terdiri dari lima komponen utama yang bekerja secara berurutan (Vaswani et al., 2017). Pada umumnya, arsitektur transformer memiliki 5 komponen utama. Komponen pertama adalah Input *Embedding* dan Positional Encoding, yang bertugas mengubah setiap token input menjadi vektor angka dan memberikan informasi posisi token dalam kalimat agar model mengetahui urutan setiap kata. Komponen kedua adalah *Multi-Head Self-Attention* yang bertugas menghitung tingkat keterkaitan setiap token dengan seluruh token lain dalam kalimat secara paralel sehingga setiap token memperoleh konteks dari keseluruhan kalimat (Tay et al., 2022). Komponen ketiga adalah *Add and Norm* pertama, yang bertugas menggabungkan hasil self-attention dengan representasi masukan awal lalu menstabilkan distribusi nilai antar lapisan agar proses pelatihan berjalan stabil. Komponen keempat adalah *Feed-Forward Network*, yang bertugas memproses setiap token secara terpisah untuk memperkaya dan memperhalus representasi yang telah dihasilkan oleh self-attention. Komponen kelima adalah Add and Norm kedua, yang bertugas kembali menggabungkan hasil feed-forward dengan nilai sebelumnya lalu menstabilkan distribusi sebelum diteruskan ke lapisan berikutnya. Kelima komponen ini ditumpuk secara berulang sehingga setiap pengulangan menghasilkan representasi yang semakin abstrak dan kaya konteks, dan pada lapisan terakhir representasi dapat diagregasi menjadi satu vektor kalimat melalui mekanisme pooling.



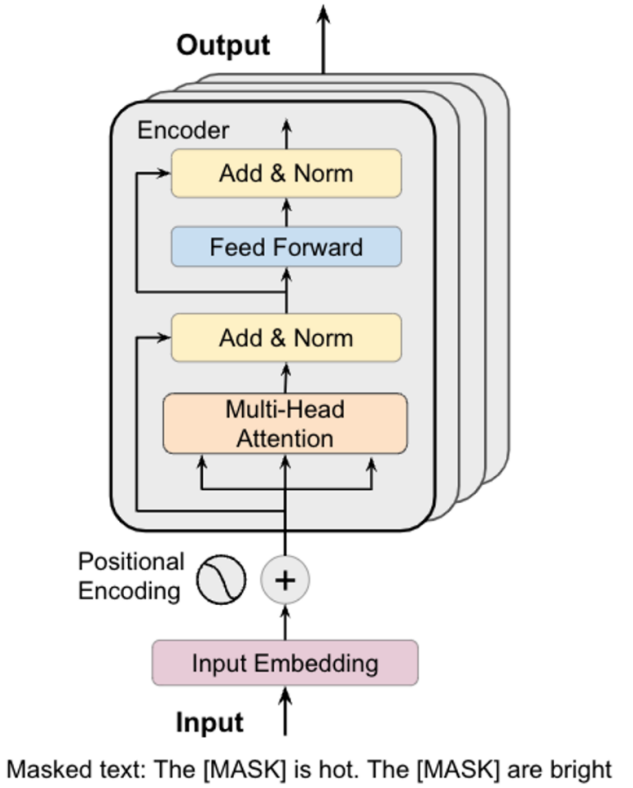
Gambar 2.1 Arsitektur Transformer

Sumber: Vaswani, et al,. 2017

2.4.2 Bidirectional Encoder Representations from Transformers (BERT)

BERT merupakan model bahasa yang digunakan untuk memahami makna semantik teks secara bidirectional dengan hanya memanfaatkan bagian encoder dari arsitektur Transformer, berbeda dari Transformer original yang menggunakan encoder sekaligus decoder (Devlin et al., 2019). Arsitektur ini terdiri dari layer Input *Embedding*, yang bertugas mengubah setiap token menjadi satu vektor angka yang merupakan gabungan dari token *embedding* untuk merepresentasikan identitas token, positional *embedding* untuk menyandikan posisi token dalam kalimat, dan segment *embedding* untuk membedakan kalimat pertama dan kedua. Lapisan selanjutnya adalah 12 Lapisan Transformer Encoder dengan Bidirectional Attention, yang bertugas memproses setiap token dapat mempelajari konteks dalam kalimat baik dari arah kiri maupun kanan secara bersamaan, menghasilkan representasi kontekstual yang jauh lebih kaya. BERT juga dilatih menggunakan dua objektif khusus yaitu Masked Language Modeling untuk memprediksi token yang disembunyikan dan Next Sentence Prediction untuk memahami hubungan antarkalimat (Areshey et al., 2023). Tahapan keempat adalah CLS Token Pooling, yang bertugas mengambil vektor pada posisi token [CLS] dari lapisan terakhir

sebagai representasi agregat seluruh kalimat yang menjadi dasar ekstraksi *embedding* pada IndoBERT dan XLM-RoBERTa dalam penelitian ini.



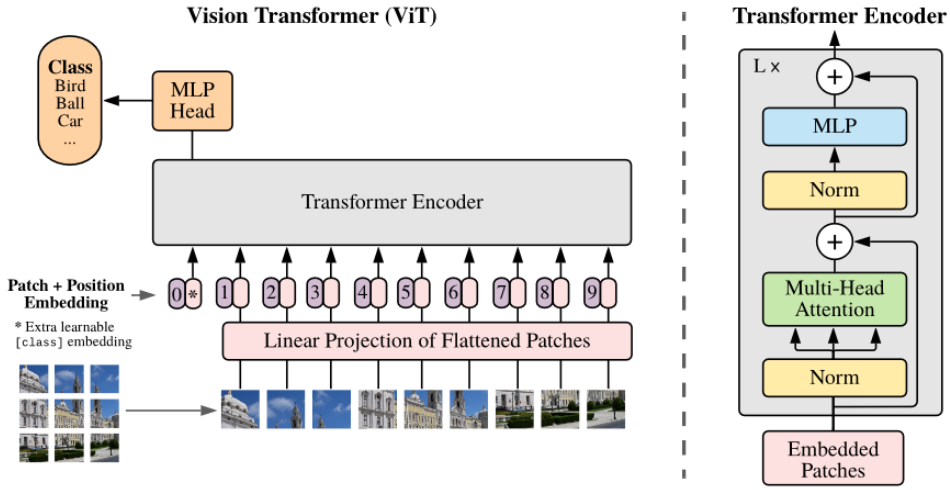
Gambar 2.2 Arsitektur BERT

Sumber: velog (2026)

2.4.3 Vision Transformer (ViT-B/32)

Vision Transformer (ViT-B/32) merupakan arsitektur image encoder yang digunakan untuk mengubah gambar menjadi representasi vektor yang dapat disejajarkan dengan representasi teks dengan cara memperlakukan potongan-potongan kecil gambar layaknya token dalam kalimat, dan pipeline pemrosesannya terdiri dari lima tahapan utama (Dosovitskiy et al., 2021). Arsitektur ini menggunakan konsep yang hamper serupa dengan BERT, dimana arsitektur transformer yang digunakan merupakan lapisan transformer encoder tanpa decoder dan menggunakan token [CLS]. Tahapan pertama adalah Patch Extraction, yang bertugas membagi gambar masukan berukuran 224x224 piksel menjadi 49

potongan atau patch berukuran 32x32 piksel dalam susunan grid 7x7, sesuai dengan konfigurasi /32 pada ViT-B/32. Tahapan kedua adalah Patch *Embedding* yang bertugas meratakan setiap patch menjadi vektor satu dimensi lalu memproyeksikannya ke dalam ruang *embedding* menghasilkan 49 vektor yang masing-masing mewakili satu potongan gambar. Tahapan ketiga adalah penambahan Token [CLS] dan Positional *Embedding*, di mana token [CLS] yang dapat dipelajari disisipkan di awal sekuens patch sebagai calon representasi agregat seluruh gambar, sementara positional *embedding* ditambahkan ke setiap vektor patch untuk mempertahankan informasi posisi spasial. Tahapan keempat adalah 12 Lapisan Transformer Encoder, yang bertugas memproses seluruh vektor patch secara paralel menggunakan mekanisme self-attention dengan 12 attention head sehingga setiap patch dapat memperhatikan patch lain di seluruh bagian gambar secara global, di mana B merujuk pada konfigurasi ini dengan 12 lapisan. Tahapan kelima adalah CLS Token Pooling, yang bertugas mengambil vektor pada posisi [CLS] dari lapisan terakhir sebagai representasi global gambar berdimensi 768 yang selanjutnya diproyeksikan ke ruang *embedding* bersama CLIP. Pada penelitian ini, seluruh parameter ViT-B/32 dibekukan sepenuhnya selama pelatihan agar untuk mempertahakan kemampuan *pre-train* yang sudah dilatih pada jutaan gambar.

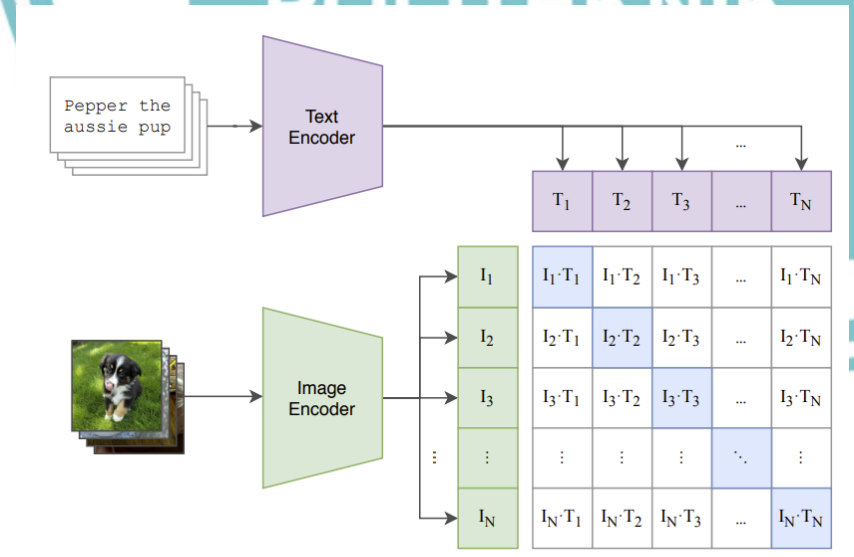


Gambar 2.3 Arsitektur Vision Transformer

Sumber: Dosovitskiy, et al., 2021

2.4.4 Contrastive Language Image Pre-trained (CLIP)

CLIP (*Contrastive Language–Image Pre-training*) merupakan arsitektur multimodal untuk tugas *cross-modal retrieval* berbasis *contrastive learning*. Pendekatan ini memaksimalkan kesamaan *embedding* pasangan gambar-teks yang relevan dan meminimalkan yang tidak relevan melalui fungsi *loss* kontrastif, sehingga kedua modalitas terproyeksi ke dalam satu ruang semantik bersama (Radford et al., 2021). Arsitekturnya terdiri dari dua *encoder* terpisah yang dilatih secara simultan menggunakan dataset berpasangan skala besar. *Image encoder* umumnya menggunakan Vision Transformer (ViT) yang memproses gambar sebagai kumpulan *patch* dan memanfaatkan *self-attention* untuk menangkap hubungan spasial global (Dosovitskiy et al., 2021). Sementara itu, *text encoder* menggunakan arsitektur Transformer untuk memodelkan hubungan kontekstual antar token via *self-attention*, menghasilkan *embedding* berdimensi tinggi yang selaras dengan *embedding* visual dalam ruang laten (Vaswani et al., 2017). Saat ini, banyak penelitian mengombinasikan *text encoder* dengan model lintas bahasa yang kuat seperti XLM-RoBERTa (Conneau et al., 2020). Berbagai riset telah membuktikan efektivitas kombinasi kedua *encoder* ini dalam tugas *cross-modal retrieval*.

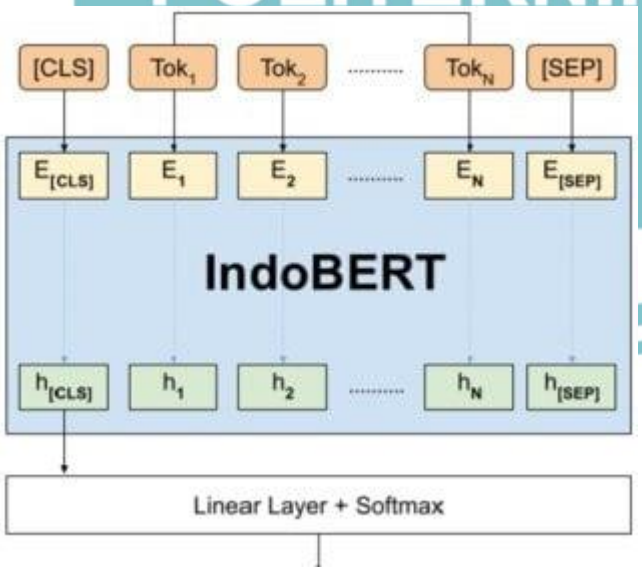


Gambar 2.4 Arsitektur CLIP

Sumber: Radford, et al., 2021

2.4.5 IndoBERT

IndoBERT merupakan model bahasa berbasis arsitektur BERT (*Bidirectional Encoder Representations from Transformers*) yang dirancang khusus untuk Bahasa Indonesia. Model ini mengadopsi arsitektur BERT Base dengan 12 lapisan *Transformer encoder*, 12 *attention head*, dan dimensi *hidden state* sebesar 768, dilatih menggunakan lebih dari 220 juta kata yang bersumber dari artikel berita, Wikipedia Bahasa Indonesia, dan teks web umum (Koto et al., 2020). Kemampuan pemodelan bahasa dua arah yang diwarisi dari BERT (Devlin et al., 2019) memungkinkan IndoBERT menangkap konteks semantik dari kedua arah dalam satu kalimat, menghasilkan representasi token yang sensitif terhadap konteks kalimat secara keseluruhan. Secara teoretis, spesialisasi pada satu bahasa target terbukti menghasilkan representasi semantik yang lebih presisi dibandingkan model multilingual pada tugas NLP dalam bahasa yang sama (Koto et al., 2020). Namun, IndoBERT tidak dirancang untuk tugas multimodal, sehingga berpotensi menghadapi kesenjangan saat diintegrasikan dengan *image encoder* CLIP. Dalam penelitian ini, IndoBERT digunakan sebagai representasi pendekatan monolingual untuk menguji secara empiris apakah spesialisasi linguistik Bahasa Indonesia memberikan keunggulan dalam tugas *cross-modal image retrieval*.

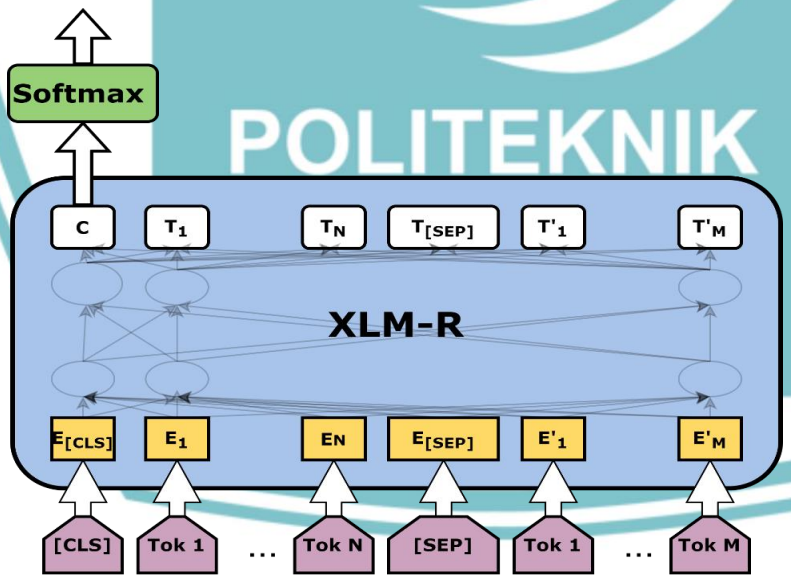


Gambar 2.5 Arsitektur IndoBERT

Sumber: Anadra, et al., 2025

2.4.6 XLM-RoBERTa

XLM-RoBERTa (*Cross-lingual Language Model–RoBERTa*) merupakan model bahasa berbasis arsitektur *Transformer encoder* yang dikembangkan sebagai perluasan multilingual dari RoBERTa, dilatih menggunakan metode *Masked Language Modeling* (MLM) pada korpus teks yang mencakup 100 bahasa dengan total lebih dari 2,5 terabyte data (Conneau et al., 2020). Model ini memiliki 24 lapisan *Transformer encoder*, 16 *attention head*, dan dimensi *hidden state* sebesar 1.024. XLM-RoBERTa menghapus objektif *Next Sentence Prediction* (NSP) yang digunakan BERT dan menggantinya dengan *dynamic masking*, di mana pola *masking* diubah setiap *epoch* sehingga model terekspos pada variasi konteks yang lebih beragam selama pelatihan (Liu et al., 2019). Selain itu, XLM-RoBERTa menggunakan tokenizer berbasis SentencePiece dengan kosakata 250.000 token. Dalam penelitian ini, XLM-RoBERTa Large digunakan sebagai *text encoder* pengganti encoder bawaan CLIP sebagai representasi pendekatan multilingual yang diuji secara empiris terhadap pendekatan monolingual dan model transformer.



Gambar 2.6 Arsitektur XLM-RoBERTa

Sumber: Ranasinghe and Zampieri, 2020

2.5. Cross Industry Standard Process - Data Mining (CRISP-DM)

CRISP-DM merupakan sebuah *process model* standar yang berfungsi sebagai kerangka kerja sistematis dalam pengembangan proyek *data mining*, *analytics*, dan *data science*. Menurut Abasova (2021) CRISP DM membagi proyek menjadi enam fase utama yang saling terkait secara iteratif: *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modeling*, *Evaluation*, dan *Deployment*.

2.5.1 Business Understanding

Business Understanding berfokus pada pemahaman tujuan bisnis dan terjemahan ke dalam tujuan data mining untuk menentukan rencana desain dan sumber daya yang diperlukan (Pambudi, 2023) Menurut Plotnikova et al. (2022), kegagalan banyak proyek data science disebabkan oleh ketidaksesuaian antara tujuan teknis model dan kebutuhan bisnis, sehingga tahap Business Understanding berperan penting sebagai fondasi konseptual.

2.5.2 Data Understanding

Tahap Data Understanding berfokus pada eksplorasi awal dataset untuk memahami struktur, karakteristik, kualitas, dan potensi permasalahan data yang tersedia. Menurut Martínez-Plumed et al. (2021), pemahaman data yang baik berpengaruh langsung terhadap kualitas model yang dihasilkan, karena kesalahan dalam interpretasi data dapat menyebabkan bias dan penurunan performa model.

2.5.3 Data Preparation

Tahap Data Preparation berfokus pada transformasi data mentah menjadi format yang siap digunakan untuk pelatihan model. Dalam penelitian ini, preprocessing mencakup dua komponen utama yaitu preprocessing citra dan tokenisasi teks yang disesuaikan dengan kebutuhan masing-masing konfigurasi model. Preprocessing citra untuk seluruh konfigurasi model dilakukan secara seragam menggunakan pipeline bawaan CLIP, meliputi resize ke resolusi 224×224 piksel, center crop, dan normalisasi nilai piksel sesuai distribusi data pelatihan CLIP (Radford et al., 2021).

2.5.4 Modelling

Tahap *modelling* melibatkan perancangan dan implementasi pendekatan pembelajaran yang digunakan untuk menghasilkan representasi *embedding* lintas-modal yang selaras antara teks Bahasa Indonesia dan gambar. Seluruh konfigurasi model dalam penelitian ini mengadopsi strategi pembekuan *image encoder* (*frozen image encoder*) selama proses pelatihan. Strategi ini didasarkan pada prinsip *Locked-image Tuning* (LiT) yang diusulkan oleh Zhai et al. (2022), di mana representasi visual CLIP ViT-B/32 yang telah dipelajari dari data berskala besar dipertahankan sepenuhnya, sementara hanya komponen teks yang disesuaikan terhadap domain Bahasa Indonesia. Strategi pelatihan yang diterapkan pada *text encoder* dengan *fine-tuning* layer teratas pada kedua model BERT. Pendekatan ini didasarkan pada temuan yang konsisten dalam literatur bahwa lower layers encoder transformer cenderung mengandung representasi linguistik yang bersifat umum dan transferable lintas tugas, sementara upper layers lebih bersifat task-specific dan mengalami perubahan paling signifikan selama fine-tuning (Hwang et al., 2025).

2.5.5 Evaluation

Tahap *Evaluation* bertujuan untuk menilai kinerja model dan kesesuaiannya dengan tujuan penelitian.

1. *Cosine similarity*

   Cosine similarity digunakan sebagai fungsi scoring untuk mengukur tingkat kemiripan semantik antara *embedding* teks dan *embedding* gambar dalam ruang vektor bersama. Secara umum, *cosine similarity* antara dua vektor didefinisikan sebagai:

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} \quad \text{sims}[i] = \mathbf{image}_i \cdot \mathbf{query}$$

Gambar 2.3 Rumus *Cosine similarity*

   Di mana A adalah vektor *embedding* gambar dan B adalah vektor *embedding* teks. Dalam implementasi CLIP, seluruh output encoder telah

dinormalisasi secara L2 sehingga ||A|| = ||B|| = 1 (Radford et al., 2021). Sehingga dapat disederhakan dengan rumus image_i . query. Secara Nilai *cosine similarity* berada pada rentang −1 hingga 1, dimana nilai yang mendekati 1 menunjukkan tingkat kemiripan yang tinggi, sedangkan nilai mendekati 0 atau negatif menunjukkan hubungan semantik yang lemah

2. Recall pada Retrieval Metric

$$R@k = \frac{Number\ of\ Relevant\ Predictions\ among\ Top - k}{Total\ Number\ of\ Relevant\ Items}$$

Gambar 2.4 Rumus Recall@K

Sumber:Mohammadi, et al., 2023

Recall dalam penelitian ini mengacu pada pengukuran proporsi gambar relevan yang berhasil ditemukan oleh sistem pencarian dari jumlah gambar yang ditampilkan sistem. (Mohammadi, et al., 2023). Recall digunakan dalam bentuk Recall@k untuk mengukur berapa banyak gambar relevan yang ditemukan dalam hasil teratas, yang penting untuk pengalaman pengguna dalam pencarian gambar yang cepat dan akurat (Ramzi et al, 2023).

3. Mean Reciprocal Rank (MRR)

Mean Reciprocal Rank (MRR) merupakan metrik evaluasi sistem retrieval yang mengukur rata-rata nilai kebalikan dari posisi pertama hasil yang relevan dalam daftar peringkat (Fumero et al., 2025).

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}.$$

Gambar 2.5 Rumus MRR

Sumber:Yi, et al., 2025

MRR dihitung dengan menjumlahkan nilai kebalikan (*reciprocal*) dari posisi pertama hasil relevan untuk setiap query, kemudian dibagi dengan total jumlah query yang dievaluasi. Dengan ini, MRR bisa menunjukkan dengan lebih detail seberapa tepat urutan gambar yang dicari, sangat cocok untuk pengujian di penelitian yang memasangkan satu kueri tepat dengan satu gambar asli (Nguyen et al., 2025)

2.5.6 Deployment

Tahap deployment didefinisikan sebagai fase di mana model hasil proses data mining yang telah dievaluasi diimplementasikan ke lingkungan operasional agar dapat digunakan oleh pengguna atau sistem bisnis secara nyata, baik dalam bentuk laporan, aplikasi, maupun integrasi ke sistem yang sudah ada, dengan tujuan memastikan hasil analisis memberikan nilai dan dampak langsung bagi organisasi (Martínez-Plumed et al., 2021).

2.6 Tokenization

Tokenization atau tokenisasi merupakan proses pengubahan input yang berupa teks maupun simbol menjadi urutan unit leksikal diskret yang dapat direpresentasikan sebagai unit yang dikenal sebagai token (Gastaldi, et al., 2025). Setiap konfigurasi model dalam penelitian ini menggunakan metode tokenisasi yang berbeda sesuai dengan arsitektur *text encoder* masing-masing, sebagai berikut:

- **Byte Pair Encoding (BPE)** adalah tokenisasi untuk model CLIP dengan vocabulary sebesar 49.408 token dan panjang konteks maksimum 77 token. BPE bekerja dengan menggabungkan pasangan karakter atau subkata yang paling sering muncul sehingga mampu menangani kata di luar kosakata (out-of-vocabulary) dengan merepresentasikannya sebagai kombinasi subword (Kong et al., 2023).

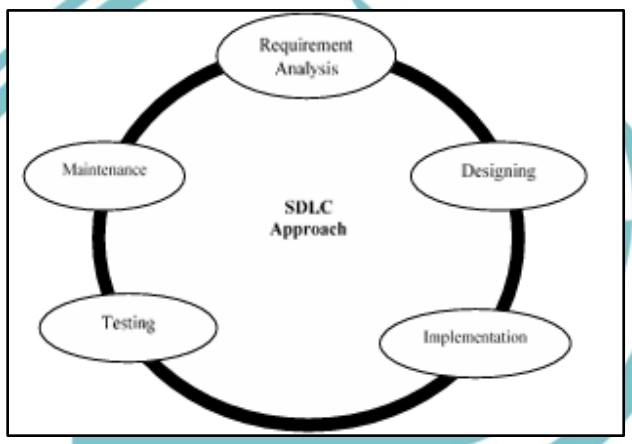
- **WordPiece Tokenization** adalah tokenisasi milik model CLIP dengan IndoBERT dengan vocabulary sebesar 50.000 token yang dibangun dari korpus Bahasa Indonesia. WordPiece memecah kata menjadi subword berdasarkan probabilitas kemunculan, sehingga mampu merepresentasikan morfologi Bahasa Indonesia yang kaya termasuk imbuhan prefiks dan sufiks secara lebih granular (Koto et al., 2020).

- **SentencePiece Tokenization** adalah tokenisasi milik model CLIP dengan XLM-RoBERTa berbasis Unigram Language Model dengan vocabulary sebesar 250.000 token yang mencakup 100 bahasa. Vocabulary yang jauh lebih besar ini memungkinkan model merepresentasikan token Bahasa

Indonesia tanpa bergantung pada dekomposisi subword yang terlalu agresif (Conneau et al., 2020).

2.7 Software Development Life-cycle (SDLC)

Model Waterfall digambarkan sebagai proses pendekatan linier terhadap pengembangan, di mana setiap fase harus diselesaikan sebelum fase berikutnya dapat dimulai. (Hossain, 2023). Model ini sangat relevan diterapkan pada pengembangan sistem yang membutuhkan alur kerja terstruktur.



Gambar 2.6 Siklus SDLC

Sumber: Yas et al, 2023

2.7.1 Planning

Tahap perencanaan merupakan fase awal dalam SDLC yang berfokus pada identifikasi tujuan sistem dan ruang lingkup pengembangan. Perencanaan yang matang bertujuan untuk meminimalkan risiko kegagalan selama proses pengembangan (Ghutmar & Date, 2023).

2.7.2 Analysis

Tahap ini bertujuan untuk mengumpulkan dan mendokumentasikan kebutuhan sistem, baik kebutuhan fungsional maupun non-fungsional. (Ghutmar & Date, 2023).

1. Functional Requirement: kebutuhan yang menjelaskan fungsi spesifik yang harus dilakukan oleh sistem (Huang et al, 2023) seperti, Menerima input

gambar dan teks, melakukan inferencing model dan menampilkan gambar hasil dari kueri pencarian

2. Non-Function Requirement: Model yang diintegrasikan ke dalam sistem Smart Gallery ditetapkan memenuhi threshold minimum Recall@10 ≥ 75% berdasarkan penelitian dari Shi et al. (2024) dan waktu inferencing 2000ms per kueri berdasarkan penelitian dari Nielsen (1993) yang menyebutkan respons di bawah 1–2 detik masih menjaga alur penggunaan tetap mulus tanpa mengganggu konsentrasi pengguna.

2.7.3 Design

Tahap desain berfokus pada penerjemahan kebutuhan pengguna ke dalam bentuk rancangan teknis. Perancangan mencakup arsitektur sistem, struktur database, alur proses, implementasi sistem, antarmuka pengguna, serta interaksi antar komponen. (Ghutmar & Date, 2023)

1. Activity Diagram: Activity Diagram memodelkan alur kerja atau proses bisnis secara rinci, menunjukkan urutan aktivitas dan keputusan yang terjadi dalam sistem. (Gedam & Meshram, 2023)
2. Use Case Diagram: Diagram ini menunjukkan interaksi antara aktor dengan sistem melalui use case yang merepresentasikan layanan atau fungsi yang disediakan sistem (Razinkas et al, 2024).
3. Sequence Diagram: Diagram ini menunjukkan interaksi antar objek dalam sistem berdasarkan urutan waktu, menggambarkan pesan yang dikirim dan diterima antar komponen untuk menyelesaikan suatu skenario tertentu. (Maina, et al., 2021)

2.7.4 Implementation

Pada tahap implementasi, rancangan sistem diterjemahkan ke dalam bentuk kode program menggunakan bahasa pemrograman dan teknologi yang telah ditentukan. Setiap modul dikembangkan sesuai spesifikasi desain dan kemudian diintegrasikan menjadi satu kesatuan sistem (Ghutmar & Date, 2023).

2.7.5 Testing

Tahap pengujian dilakukan untuk memastikan sistem berjalan sesuai kebutuhan pengguna. Pengujian mencakup pengujian integrasi, pengujian sistem, serta validasi terhadap keseluruhan fungsi aplikasi. Fase ini bertujuan menemukan kesalahan, meningkatkan kualitas perangkat lunak, dan membangun kepercayaan sebelum sistem diterapkan secara resmi (Yas et al, 2023).

2.8 Penelitian Terkait

Tabel 2.1 Penelitian Terkait

Judul	Contrastive Language–Image Pre-training for the Italian Language (Bianchi et al, 2021)	CLIPSE – A Minimalistic CLIP-Based Image Search Engine For Research, (Steve Göring, 2025)	Neural-Based Cross-Modal Search and Retrieval of Artwork (Gong et al, 2023)
Tujuan	Mengembangkan model CLIP yang diadaptasi untuk Bahasa Italia dengan melatih text encoder berbasis BERT Italia menggunakan contrastive learning pada dataset gambar-teks berbahasa Italia, sebagai alternatif terhadap model CLIP multilingual generalis.	Menyediakan mesin pencari gambar yang sederhana, open-source, dan mudah diatur (self-hosted) untuk keperluan penelitian pada dataset kecil	Membangun sistem pencarian dan *retrieval* cerdas (BoonArt) khusus untuk karya seni lukisan yang mampu menangkap hubungan semantik entitas dalam gambar
Metode	Model CLIP-Italian dilatih pada lebih dari 1,4 juta pasangan gambar-teks berbahasa Italia dengan menggabungkan berbagai sumber data, menggunakan text encoder berbasis BERT Italia dan image encoder ViT dengan fungsi symmetric contrastive loss.	Mengembangkan aplikasi CLIPSE menggunakan Python dan OpenCLIP dengan Mengubah gambar dan kueri teks menjadi embedding menggunakan CLIP, lalu menghitung kemiripan menggunakan dot product.	Mengembangkan mesin pencari BoonArt yang menggunakan jaringan VSE bernama VITR (ViT-Relation-focus network). Menggunakan encoder CLIP yang di-fine-tune pada dataset RefCOCOg untuk memahami hubungan area gambar. Diuji menggunakan dataset ArtUK
Hasil	CLIP-Italian mengungguli model CLIP multilingual pada tugas image retrieval dan zero-shot classification dalam Bahasa Italia.	Berhasil membuat mesin pencari minimalis yang mendukung antarmuka baris perintah (command line) dan web. Waktu pengindeksan tumbuh secara linear, dan waktu pencarian sangat cepat (sekitar 57ms untuk ~25rb gambar). Cocok untuk dataset kecil, namun disarankan menggunakan pendekatan terdistribusi untuk dataset besar	BoonArt mencapai skor Recall@10 sebesar 97,4% untuk pencarian teks-ke-gambar dan 97% untuk gambar-ke-teks. Mampu melakukan pencarian cross-modal (gambar-ke-teks dan sebaliknya) dan mengungguli sistem pencarian standar website ArtUK

Ketiga penelitian terdahulu secara kolektif membangun urgensi dan arah penelitian ini dari tiga sudut pandang yang saling melengkapi. CLIP-Italian membuktikan bahwa adaptasi CLIP untuk bahasa non-Inggris melalui *contrastive learning* dengan text encoder berbasis model bahasa spesifik dapat memberikan performa yang cukup baik pada tugas image retrieval, memberikan landasan metodologis yang relevan bagi penelitian ini dalam konteks Bahasa Indonesia. CLIPSE menunjukkan bahwa sistem pencarian gambar berbasis CLIP dapat diimplementasikan secara efektif pada skala kecil hingga menengah tanpa infrastruktur komputasi berat, sehingga fokus penelitian dapat diarahkan pada peningkatan kualitas pemahaman semantik melalui optimasi text encoder. Penelitian *Neural-Based Cross-Modal Search and Retrieval of Artwork* oleh Gong, 2013 membuktikan bahwa fine-tuning CLIP pada domain dan karakteristik linguistik tertentu secara signifikan meningkatkan performa *cross-modal retrieval* dibandingkan pendekatan generalis. Ketiga penelitian sudah membuktikan bahwa arsitektur CLIP sudah terbukti dapat diimplementasikan ke dalam sistem pencarian gambar dengan menggunakan *text encoder* yang sesuai dan dilatih dengan domain spesifik. Dengan demikian, terdapat celah penelitian yang belum dijawab, yaitu belum ada studi yang membandingkan secara langsung efektivitas *text encoder* dalam konteks pencarian gambar menggunakan bahasa Indonesia secara semantik.

BAB 3
PEMBAHASAN

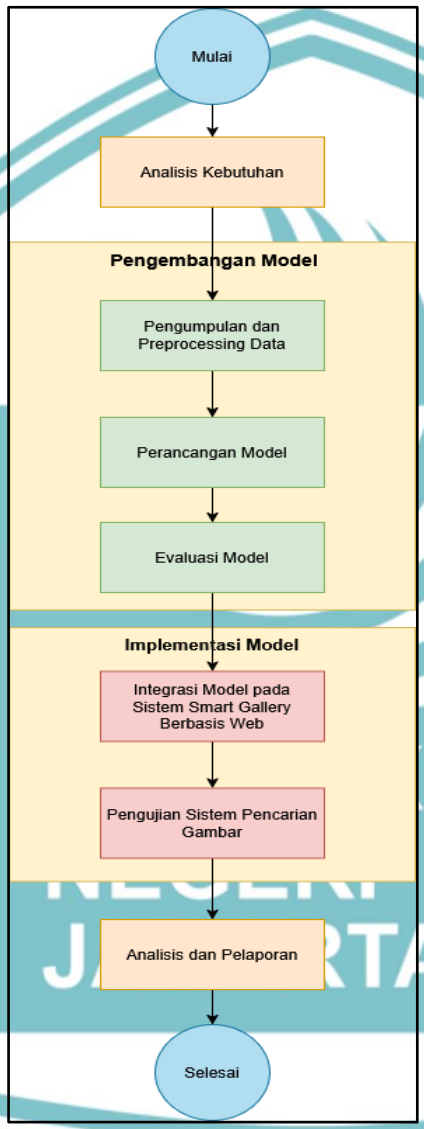
3.1 Rancangan Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode rekayasa sistem yang melakukan perancangan dan mengimplementasikan sistem pencarian gambar cerdas pada website Smart Gallery berbasis *cross-modal retrieval* dengan memanfaatkan arsitektur Contrastive Language–Image Pre-trained (CLIP). Model yang paling efektif dalam mendukung pencarian gambar berbasis kueri Bahasa Indonesia diuji dengan melakukan studi komparatif terhadap tiga konfigurasi text encoder pada arsitektur Contrastive Language–Image Pre-trained (CLIP), yaitu CLIP standar, arsitektur CLIP dengan XLM-RoBERTa dan CLIP dengan IndoBERT sebagai *text encoder*.

Studi komparatif dirancang sebagai eksperimen terkontrol di mana ketiga model dilatih menggunakan metode fine-tuning yang identik, yakni Contrastive Loss dengan frozen image encoder, pada dataset yang sama yaitu Flickr8k Indonesia. Keseragaman metode training ini memastikan bahwa perbedaan performa yang terukur dapat diatribusikan sepenuhnya pada pilihan arsitektur text encoder, bukan pada perbedaan paradigma training. Model dengan performa terbaik berdasarkan hasil evaluasi kemudian diintegrasikan ke dalam sistem Smart Gallery sebagai model final.

Sistem Smart Gallery diimplementasikan dalam bentuk website berbasis web yang menyediakan fitur pencarian gambar cerdas menggunakan kueri teks dan gambar berbahasa Indonesia. Sistem dapat menerima input teks maupun gambar dari pengguna, kemudian memetakannya ke dalam ruang *embedding* yang sama sehingga pengguna dapat melakukan pencarian gambar secara lebih efisien dan intuitif tanpa bergantung pada metadata manual.

Penelitian ini menggabungkan framework CRISP-DM sebagai landasan metode pengembangan model dan framework Waterfall sebagai landasan implementasi model pada perangkat lunak untuk memastikan pengerjaan penelitian lebih terstruktur dengan visualisasi sebagai berikut:



Gambar 3.1 Rancangan Penelitian

3.1.2 Tahapan Penelitian

Tahapan penelitian dalam mengembangkan dan menerapkan model CLIP pada sistem pencarian gambar dilakukan dengan mengikuti alur yang sistematis dan runtut seperti yang tertera pada Gambar 3.1 dengan rincian setiap bagian sebagai berikut:

1. Analisis Kebutuhan

Tahap analisa kebutuhan mencangkup kebutuhan perangkat keras, kebutuhan perangkat lunak, kebutuhan dataset, kebutuhan model, kebutuhan fungsional dan kebutuhan non-fungsional aplikasi web dan kebutuhan perangkat lunak untuk pengembangan aplikasi web.

2. Persiapan dan Preprocessing Data

Tahap ini dilakukan dengan menggunakan dataset Flickr8k dan Flickr30K Indonesia yang masing masing berisi 8.091 dan 31.014 gambar dengan masing-masing lima caption berbahasa Indonesia. Tahap preprocessing disesuaikan dengan kebutuhan masing-masing model yang dibandingkan.Tiap model memiliki tokenisasi yang berbeda, CLIP menggunakan tokenisasi standart dari CLIP yaitu BPE Tokenization, CLIP dengan IndoBERT menggunakan Tokenisasi Word Piece dan CLIP dengan XLM-RoBERTa menggunakan Tokenisasi Setence Piece. Preprocessing citra untuk seluruh model meliputi resizing, cropping, dan normalisasi sesuai konfigurasi pretrained CLIP. Dataset kemudian disiapkan pada data loader yang seragam untuk ketiga model guna memastikan konsistensi eksperimen komparatif.

3. Perancangan model

Pada tahap ini dirancang skema perbandingan tiga konfigurasi model dengan arsitektur image encoder yang sama, yaitu Vision Transformer (ViT-B/32) yang dilakukan frozen sesuai dengan penelitian yang dilakukan oleh Zhai et al. (2022). Perbedaan ketiga konfigurasi terletak pada pilihan text encoder sebagai berikut:

- CLIP standar: menggunakan text encoder bawaan CLIP berbasis Transformer yang di-fine-tune menggunakan *contrastive learning* pada dataset Flickr8k Indonesia, berfungsi sebagai baseline utama.
- CLIP dengan IndoBERT: menggunakan IndoBERT sebagai text encoder yang diintegrasikan ke dalam arsitektur CLIP melalui projection layer untuk menyesuaikan dimensi *embedding*, 6 dari 12 lapisan / layer dari IndoBERT dilakukan *freeze* untuk mencegah *catastrophic forgetting* ketika melakukan fine tuning model karena lapisan bawah mengandung semantik dasar.

Kemudian, model dan *projection layer* di-*fine-tune* menggunakan *contrastive learning*, sebagai kombinasi model monolingual.

- CLIP dengan XLM-RoBERTa : menggunakan text encoder multilingual yang dirancang untuk kompatibilitas dengan ruang *embedding* CLIP, 1-18 lapisan / layer dari model XLM-RoBERTa dilakukan freeze untuk mencegah *catastrophic forgetting* atau kehilangan pengetahuan dasar semantik dari berbagai bahasa. Lapisan 19 hingga 24 dilakukan fine tuning bersama dengan *projection head* menggunakan *contrastive loss* pada dataset yang sama, berfungsi sebagai kombinasi model multilingual.

Ketiga model akan dikonfigurasi dengan batch size dan jumlah epoch yang identik, sedangkan *learning rate* ditentukan secara individual per arsitektur. Hal ini dilakukan karena perbedaan skala parameter dan karakteristik pre-training masing-masing model berpotensi menyebabkan instabilitas pelatihan jika menggunakan *learning rate* yang seragam. InfoNCE loss dipilih sebagai fungsi loss yang seragam karena dirancang secara native untuk cross-modal alignment pada arsitektur dual-encoder.

4. Evaluasi model

Evaluasi performa ketiga model dilakukan menggunakan dua metrik retrieval yang saling melengkapi, yaitu Recall@K dan Mean Reciprocal Rank (MRR). Recall@K mengukur proporsi gambar relevan yang berhasil ditemukan dalam top-K hasil pencarian, sedangkan MRR mengukur rata-rata posisi pertama hasil yang relevan dalam daftar peringkat sehingga memberikan gambaran yang lebih sensitif terhadap kualitas ranking. Pemilihan metrik evaluasi Recall@K dan MRR karena sesuai dengan skenario penggunaan sistem pencarian gambar. Model dengan nilai MRR dan Recall@K tertinggi dipilih sebagai model terbaik dan dilakukan export menggunakan format PyTorch untuk diintegrasikan ke dalam sistem Smart Gallery.

5. Implementasi Sistem Smart Gallery Berbasis Web

Tahap ini mencakup implementasi sistem website Smart Gallery berdasarkan metode SDLC Waterfall, yang meliputi analisis kebutuhan pengguna, perancangan desain sistem dan antar muka, pembangunan backend menggunakan python untuk

pengelolaan *embedding* gambar dan teks, serta penyediaan framework Fast API pencarian gambar berbasis kueri teks. Selain itu, dilakukan pembangunan frontend menggunakan NextJS sebagai antarmuka pengguna untuk melakukan pencarian gambar menggunakan deskripsi teks berbahasa Indonesia dan menampilkan hasil pencarian berupa daftar gambar paling relevan.

6. Evaluasi Sistem Pencarian Gambar

Evaluasi sistem dilakukan melalui dua pendekatan yang saling melengkapi. Pertama, evaluasi kuantitatif menggunakan metrik Recall@K, MRR, dan waktu inferensi untuk menilai akurasi pencarian dan responsivitas sistem secara objektif. Kedua, evaluasi kualitatif dilakukan melalui pengujian *usability* menggunakan System Usability Scale (SUS) yang melibatkan pengguna nyata untuk menilai kemudahan penggunaan dan kepuasan terhadap sistem Smart Gallery yang dikembangkan.

7. Analisis, dan Pelaporan

Tahap akhir meliputi pengujian fungsional website Smart Gallery untuk memastikan seluruh fitur pencarian, penyimpanan data, dan penampilan hasil pencarian berjalan sesuai dengan tujuan penelitian. Hasil pengujian kemudian dianalisis untuk menilai performa sistem, relevansi hasil pencarian, serta efektivitas penerapan fine-tuned CLIP dalam pencarian gambar berbasis konten. Selanjutnya dilakukan penarikan kesimpulan dan perumusan saran untuk pengembangan sistem di masa mendatang.

3.2 Sumber Data

Sumber data yang digunakan dalam penelitian ini dibagi menjadi dua bagian, yaitu data untuk penulisan dan data untuk pengembangan model.

Data untuk penulisan berasal dari studi literatur yang dilakukan dengan mengumpulkan referensi dari jurnal, artikel ilmiah, dan dokumentasi teknis terkait CLIP, CLIP dengan arsitektur BERT, XLM-RoBERTa, IndoBERT, Vision Transformer, *cross-modal retrieval*, serta pengelolaan media digital. Tahap ini bertujuan untuk memahami pendekatan arsitektural, metode *embedding* lintas-modal, karakteristik masing-masing model yang dibandingkan, dan metrik evaluasi kesamaan vektor, sekaligus memperkuat landasan teoritis penelitian.

Data untuk pengembangan dan evaluasi model menggunakan dataset Flickr8k Indonesia dan Flickr30K Indonesia yang merupakan terjemahan caption dari dataset Flickr8k dan Flickr30K ke dalam Bahasa Indonesia menggunakan *machine translation*. Dataset ini terdiri dari 8.091 dan 31.014 gambar yang masing-masing gambar memiliki caption yang mendeskripsikan konten visual gambar menggunakan Bahasa Indonesia. Dataset ini digunakan secara seragam untuk melatih dan mengevaluasi ketiga konfigurasi model yang dibandingkan, sehingga hasil komparasi bersifat valid dan tidak bias terhadap perbedaan data.

3.3 Teknik Pengumpulan Data

Teknik pengumpulan yang digunakan dalam penelitian ini dilakukan dengan cara pengambilan dataset publik dan studi dokumentasi

Dataset publik yang digunakan adalah Flickr8k Indonesia yang dapat diakses dan diunduh melalui platform Kaggle (https://www.kaggle.com/datasets/joykaihatu/image-caption-indonesia) dab Flickr30K dapat diunduh atau diakses melalui platform *Huggingface* (https://huggingface.co/datasets/indrad123/flickr30k-google-translate-id). Secara lengkap, dataset dapat dijabarkan pada tabel 3.1 berikut ini

Tabel 3.1 Perbandingan Dataset

Kode	Nama Dataset	Jumlah Dataset	Sumber
D-1	Flickr8K Indonesia	8.091 Gambar 40.455 Caption	Kaggle
D-2	Flickr30K Indonesia	31.014 Gambar 155.070 Caption	HuggingFace

Dataset ini digunakan untuk proses pelatihan model dan evaluasi ketiga konfigurasi model pada eksperimen komparatif. Data untuk penulisan penelitian ini berasal dari studi literatur pada jurnal penelitian yang relevan dengan topik penelitian, dengan penelusuran terhadap jurnal ilmiah yang membahas teori dan pengembangan sistem *cross-modal retrieval*, model bahasa, dan evaluasi sistem pencarian gambar. Data

hasil penelitian diperoleh dari hasil evaluasi model komparatif dan pengujian sistem pencarian gambar.

3.4 Teknik Analisis Data

Teknik analisis data dalam penelitian ini menggunakan pendekatan kuantitatif dengan analisis statistik deskriptif serta evaluasi performa model dan efektivitas sistem pencarian gambar.

Analisis performa model dilakukan terhadap hasil evaluasi ketiga konfigurasi text encoder menggunakan dua metrik utama. Recall@K digunakan untuk mengukur proporsi gambar relevan yang berhasil ditampilkan dalam sejumlah K hasil pencarian teratas, yang secara langsung mencerminkan pengalaman pengguna dalam menemukan gambar yang dicari. Kemudian, Mean Reciprocal Rank (MRR) digunakan untuk mengukur rata-rata posisi pertama hasil yang relevan dalam daftar peringkat, memberikan gambaran yang lebih sensitif. Hasil evaluasi ketiga model dianalisis secara komparatif untuk mengidentifikasi konfigurasi text encoder yang paling efektif untuk pencarian gambar berbasis Bahasa Indonesia.

Analisis efektivitas sistem dilakukan melalui dua pendekatan yang saling melengkapi. Evaluasi kuantitatif sistem menggunakan Recall@K, MRR, dan waktu inferensi untuk mengukur akurasi dan responsivitas sistem dalam kondisi operasional nyata. Evaluasi kualitatif dilakukan menggunakan instrumen System Usability Scale (SUS) untuk mengukur tingkat kemudahan penggunaan dan kepuasan pengguna terhadap antarmuka dan fungsionalitas sistem Smart Gallery yang dikembangkan. Hasil seluruh analisis digunakan untuk menjawab kedua tujuan penelitian: efektivitas komparatif text encoder dan keberhasilan implementasi sistem Smart Gallery berbasis model terpilih

BAB 4
HASIL DAN PEMBAHASAN

4.1 Identifikasi Kebutuhan

Berdasarkan studi literatur serta diskusi bersama dosen dan guru selaku target pengguna akhir, telah berhasil mengidentifikasi beberapa kebutuhan untuk penelitian ini. Kebutuhan tersebut mencakup kebutuhan pengembangan model, kebutuhan pengembangan web, kebutuhan perangkat lunak dan kebutuhan perangkat keras. Semua kebutuhan tersebut perlu dilengkapi untuk menjamin kelancaran penelitian.

4.1.1 Kebutuhan Pengembangan Model

A. Kebutuhan Dataset

Dataset merupakan komponen utama dalam pelatihan ketiga konfigurasi model pada penelitian ini. Setiap model memerlukan pasangan gambar-teks berbahasa Indonesia agar proses *contrastive learning* agar dapat mengoptimalkan representasi visual dan tekstual. Kriteria dataset yang diperlukan dalam penelitian ini dapat diidentifikasi pada Tabel 4.1 berikut.

Tabel 4.1 Kebutuhan Dataset

Kode	Nama	Deskripsi
KD-1	Translation Based dataset	Dataset merupakan hasil terjemahan caption bahasa Inggris ke bahasa Indonesia
KD-2	Deskripsi Gambar	Setiap gambar disertai caption deskriptif berbahasa Indonesia yang menggambarkan konten visual gambar secara informatif
KD-3	Isi Data	Setiap entri data terdiri atas pasangan gambar (JPEG/PNG) dan caption teks berbahasa Indonesia yang tersimpan dalam file CSV dengan kolom nama file gambar dan teks caption

B. Kebutuhan Model

Penelitian ini membandingkan tiga konfigurasi model yang masing-masing memiliki arsitektur *text encoder* berbeda namun berbagi *image encoder* yang sama. Kebutuhan model diidentifikasi pada Tabel 4.2 berikut.

Tabel 4.2 Kebutuhan Model

Kode	Nama	Deskripsi
KM-1	Dukungan Bahasa Indonesia	Model mampu memahami dan merepresentasikan teks berbahasa Indonesia secara akurat ke dalam ruang embedding yang selaras.
KM-2	Kemampuan mengubah Gambar menjadi Embedding	Model harus mampu mengkonversi gambar mentah menjadi vektor representasi numerik berdimensi tetap.
KM-3	Kemampuan mengubah Teks menjadi Embedding	Model harus mampu mengkonversi teks berbahasa Indonesia menjadi vektor representasi berdimensi yang selaras dengan ruang embedding gambar.
KM-4	Performa Komparatif	Ketiga konfigurasi model dievaluasi menggunakan Recall@1, Recall@5, Recall@10, dan MRR untuk mengidentifikasi konfigurasi text encoder terbaik untuk pencarian gambar berbasis Bahasa Indonesia.

4.1.2 Kebutuhan Website

Website dikembangkan bertujuan untuk mengimplementasi sistem pencarian gambar agar pengguna dapat lebih mudah menggunakan model.

A. Kebutuhan Fungsional

Kebutuhan fungsional yang dijelaskan pada tabel merupakan fitur-fitur yang harus tersedia dalam sistem agar tujuan penelitian dapat tercapai. Kebutuhan fungsional mencangkup sistem dapat melakukan pencarian berdasarkan teks, gambar dan menambahkan gambar baru seperti yang dijabarkan pada Tabel 4.3.

Tabel 4.3 Kebutuhan Fungsional

Kode	Nama	Deskripsi
KF-1	Input Teks Pencarian	Pengguna dapat memasukkan kueri teks berbahasa Indonesia untuk mendapatkan top-K gambar yang paling relevan.
KF-2	Input Gambar Pencarian	Pengguna dapat mengunggah gambar sebagai kueri untuk mendapatkan gambar serupa.

B. Kebutuhan Non-Fungsional

Kebutuhan non-fungsional mendefinisikan standar kualitas layanan sistem pencarian gambar, mencakup aspek kecepatan respons dan kenyamanan penggunaan antarmuka seperti yang dijabarkan pada Tabel 4.4.

Tabel 4.4 Kebutuhan Non-Fungsional

Kode	Nama	Deskripsi
KNF-1	Kecepatan Pencarian	Sistem harus mengembalikan hasil pencarian dalam waktu kurang dari 2 detik.
KNF-2	Responsivitas	Antarmuka harus dapat diakses dengan baik pada perangkat desktop maupun mobile, serta kompatibel dengan browser modern.

4.1.3 Kebutuhan Perangkat Keras

Platform Kaggle dipilih sebagai lingkungan komputasi untuk menjalankan perancangan model, pemrosesan data, proses pre-training, fine-tuning model, sekaligus mengevaluasi performa model pencarian gambar yang dikembangkan. Rincian mengenai spesifikasi perangkat keras pendukung yang digunakan pada platform tersebut dapat dilihat pada Tabel 4.5.

Tabel 4.5 Kebutuhan Perangkat Keras Pengembangan Model

Kode	Nama	Deskripsi
PKM-1	GPU	NVIDIA Tesla T4 2x
PKM-2	CPU	Intel(R) Xeon(R) CPU @ 2.00GHz
PKM-3	RAM	16 GB

Berbeda halnya dengan fase pengembangan model yang menuntut performa komputasi tinggi, kebutuhan perangkat keras untuk pengembangan web memerlukan spesifikasi komputer yang jauh lebih ringan. Rincian spesifikasi minimum untuk kebutuhan web tersebut dicantumkan pada Tabel 4.6.

Tabel 4.6 Kebutuhan Perangkat Keras Pengembangan Website

Kode	Nama	Deskripsi
PKW-1	CPU	Minimal 4 Core
PKW-2	RAM	Minimal 8 GB

4.1.4 Kebutuhan Perangkat Lunak

Pengembangan model pada penelitian ini memerlukan pustaka dan framework yang mendukung *deep learning*, pemrosesan bahasa alami, serta komputasi numerik yang diidentifikasi pada Tabel 4.7 berikut.

Tabel 4.7 Kebutuhan Perangkat Lunak Pengembangan Model

Kode	Nama	Deskripsi
PLM-1	Pytorch	*Framework deep learning* utama untuk membangun arsitektur model dan mengelola proses pelatihan.
PLM-2	Numpy	Pustaka komputasi numerik untuk mengubah array dan penyimpanan *embedding*
PLM-3	HuggingFace Transformers	Pustaka untuk memuat model dan tokenizer dari model
PLM-4	openai-clip	Pustaka untuk memuat model CLIP

Kode	Nama	Deskripsi
PLM-5	scikit-learn	Pustaka untuk komputasi metrik evaluasi, termasuk *cosine similarity* antar *embedding*.
PLM-6	Pandas	Pustaka untuk memuat dan memproses file CSV metadata dataset Flickr8K Indonesia.
PLM-7	Kaggle Notebooks	Platform penyedia lingkungan pelatihan model.

Pengembangan sistem web pencarian gambar berbasis model CLIP dan Indobert memerlukan framework dan tools berikut untuk menghasilkan sistem web yang sesuai dengan tujuan penelitian. Kebutuhan perangkat lunak dapat diidentifikasi dengan Tabel 4.8 berikut ini.

Tabel 4.8 Kebutuhan Perangkat Lunak Pengembangan Website

Kode	Nama	Deskripsi
PLW-1	FastAPI	Framework untuk membangun RESTful API.
PLW-2	Uvicorn	Untuk menjalankan aplikasi FastAPI dan menangani permintaan HTTP.
PLW-3	Next JS	Framework untuk membangun antarmuka pengguna (frontend) sistem pencarian gambar.
PLW-4	Docker	Platform kontainerisasi untuk mengemas backend FastAPI dan frontend Next.js agar konsisten.
PLW-5	VSCode	IDE yang digunakan sebagai lingkungan pengembangan utama selama penelitian.

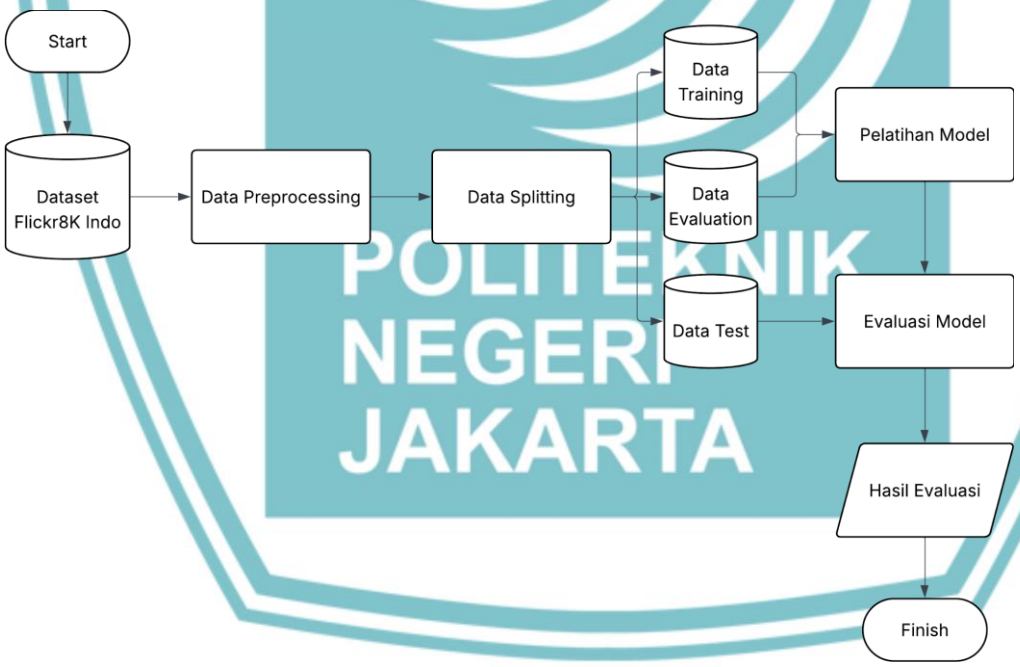
4.2 Perancangan Sistem

Tahap perancangan sistem dilakukan segera setelah analisis kebutuhan selesai dipetakan, guna memberikan panduan terarah dalam proses realisasi dan implementasi sistem. Pada penelitian ini, ruang lingkup perancangan difokuskan pada dua aspek, yaitu perancangan model *deep learning* dan perancangan aplikasi web.

4.2.1 Perancangan Model Deep Learning

Perancangan Alur Pelatihan Model

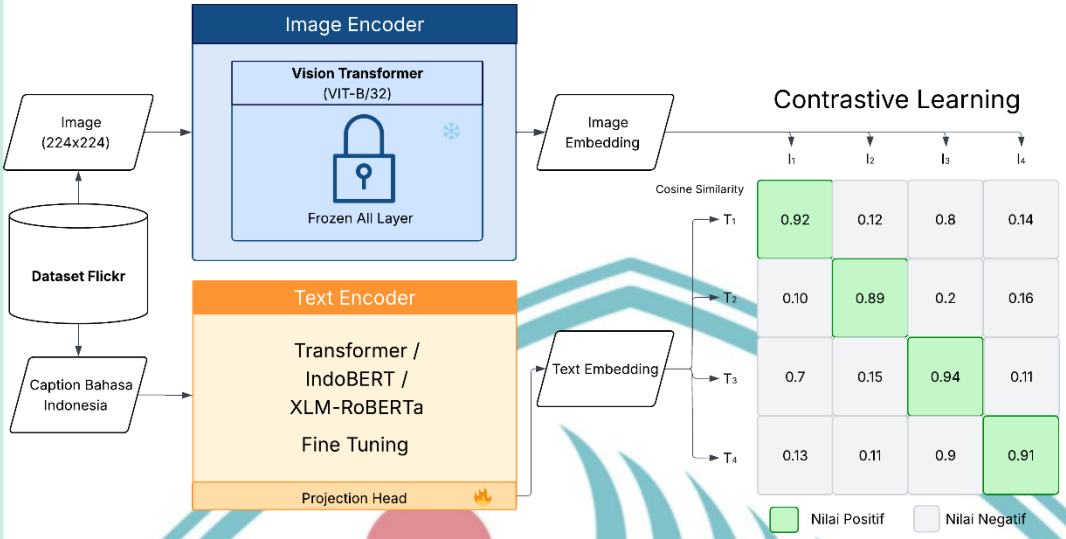
Model yang dibutuhkan dalam penelitian ini adalah model yang dapat melakukan pencarian lintas modal (*cross-modal retrieval*) dengan cara mengubah data gambar dan data teks menjadi vektor *embedding* yang dapat dibandingkan dalam satu ruang vektor yang sama.



Gambar 4.1 Alur Tahapan Pelatihan Model

Gambar 4.1 menjelaskan tentang alur pelatihan yang dimulai dari tahap data preprocessing pada dataset Flickr8K dan Flickr30K yang sudah dikumpulkan. Tahap ini dimulai dengan pembersihan inline meliputi penghapusan karakter khusus dan eliminasi duplikasi *caption*, diikuti preprocessing teks menggunakan tokenizer sesuai arsitektur masing-masing model CLIP menggunakan *tokenizer*

BPE, IndoBERT menggunakan WordPiece, dan XLM-RoBERTa menggunakan SentencePiece. Tahap preprocessing selanjutnya adalah mengolah data gambar menggunakan normalisasi standar CLIP dengan augmentasi pada data training dan tanpa augmentasi secara identik pada ketiga model pada data validasi dan *testing*. Data splitting dilakukan untuk membagi dataset menjadi tiga bagian terpisah: data training, data validasi, data test. Data training bersama dengan data validasi digunakan untuk tahap pelatihan model sedangkan data test digunakan untuk mengevaluasi model. Pembagian dataset Flickr30K menggunakan Karpathy Split dengan perbandingan data train 29.000, data validation 1.014, dan data test 1.000. Flickr8K menggunakan rasio Split sebesar 90%:5%:5%. Pendekatan ini dipilih untuk menjaga konsistensi proporsi antar kedua dataset, meskipun jumlah absolut sampel berbeda. Pada tahap pelatihan, tiga kombinasi arsitektur model *cross-modal retrieval* berbasis *embedding* yaitu CLIP, CLIP dengan Indobert, CLIP dengan XLM-RoBERTa untuk menentukan konfigurasi *text encoder* yang paling efektif dalam memproses kueri Bahasa Indonesia dengan *image encoder* yang sama yaitu Vision Trasnformer (ViT-B/32) yang di bekukan. Setelah proses pelatihan, model diuji pada tahap evaluasi menggunakan data test yang sebelumnya sudah disiapkan dengan metrik evaluasi yang sudah disebutkan pada tahapan penelitian. Tahap terakhir akan menghasilkan hasil evaluasi model kuantitatif yang dapat dibandingkan seperti Recall@K dan MRR yang menunjukkan kemampuan model untuk memberikan hasil kueri bahasa Indonesia yang sesuai.



Gambar 4.2 Alur Pelatihan Contrastive Learning

Gambar 4.2 menunjukkan proses pelatihan *contrastive learning* pada model CLIP dimulai dari tahap ekstraksi representasi (*embedding*) dari dua modalitas data yang berbeda, yaitu data gambar dan data teks. Data gambar yang berukuran 224×224 piksel diproses melalui *Image Encoder* berbasis *Vision Transformer* (ViT-B/32) yang seluruh lapisannya dibekukan (*frozen*) selama pelatihan. Pada data teks, *caption* berbahasa Indonesia diproses melalui *Text Encoder* yang menjalani *fine-tuning* dan dilengkapi dengan *Projection Head* yang dapat dilatih, sehingga menghasilkan *Text Embedding* dalam ruang vektor yang sama dengan *Image Embedding*. Keselarasan ruang vektor ini menjadi prasyarat utama agar *embedding* dapat dihitung dan diproses pada tahap *contrastive learning*.

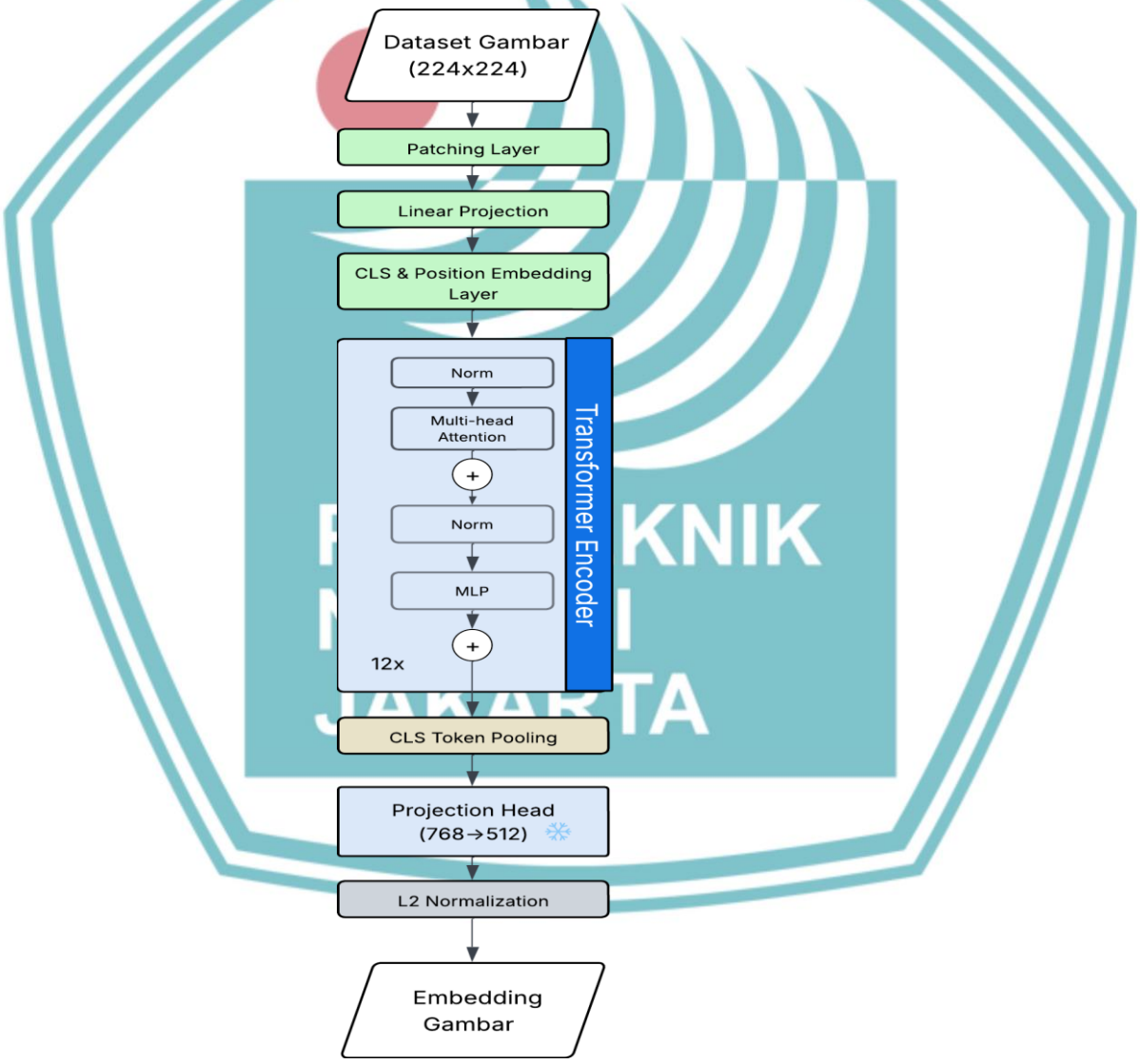
Proses *contrastive learning* dilakukan dengan membentuk *similarity matrix* berukuran N×N menggunakan *cosine similarity*, di mana setiap *image embedding* dihitung kemiripannya terhadap seluruh *text embedding* dalam satu *mini-batch*. Pasangan gambar dan teks yang benar akan menempati posisi diagonal matriks karena gambar ke-i dan teks ke-i yang saling bersesuaian selalu berbagi indeks baris dan kolom yang sama, sementara pasangan yang tidak bersesuaian berada di luar diagonal. Proses pelatihan menggunakan *InfoNCE Loss* secara simetris dari dua arah, yaitu dari gambar ke teks dan dari teks ke gambar, yang mendorong nilai diagonal terus meningkat melalui pembaruan bobot *Text Encoder* dan *Projection Head*, sementara nilai non-diagonal ditekan mendekati nol. Melalui mekanisme ini, kedekatan posisi dalam ruang *embedding* mencerminkan kedekatan makna

semantik antara gambar dan teks yang berpasangan sehingga model dapat memasangkan pasangan gambar dan teks yang sesuai.

Pendefinisian Model

Pelatihan model terdiri dari dua komponen utama yaitu *Image encoder* dan Text Encoder yang dapat mengubah modalitas dataset gambar dan teks menjadi *embedding* yang dapat dijelaskan sebagai berikut ini.

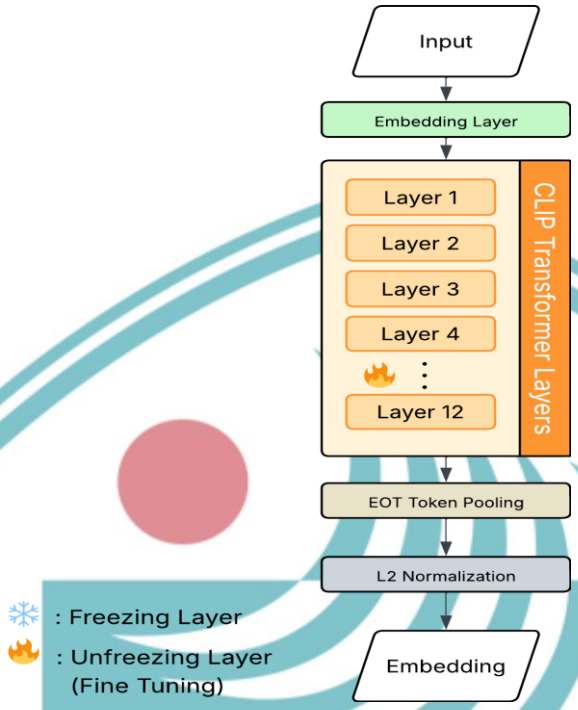
Arsitektur *image encoder* dengan model ViT-B/32 sebagai berikut



Gambar 4.3 Arsitektur Image Encoder ViT-B/32

Gambar 4.3 menunjukkan arsitektur *Image encoder* yang digunakan pada seluruh konfigurasi model dalam penelitian ini adalah ViT-B/32 (*Vision Transformer Base/32*) bawaan CLIP, yang memproses gambar masukan berukuran 224×224 piksel melalui serangkaian tahapan transformasi untuk menghasilkan representasi visual dalam ruang *embedding* bersama. Gambar pertama-tama dipecah menjadi 49 *patch* berukuran 32×32 piksel oleh Patching Layer, lalu setiap *patch* diproyeksikan secara linear menjadi vektor berdimensi 768 oleh Linear Projection. Satu token CLS ditambahkan di awal urutan dan *positional embedding* dijumlahkan ke seluruh token pada CLS & Position *Embedding* Layer agar model mempertahankan informasi posisi spasial. Urutan token ini kemudian diproses oleh 12 blok Transformer Encoder yang masing-masing terdiri dari *Layer Normalization*, *Multi-Head Self-Attention* dengan 12 *head*, *residual connection*, *Layer Normalization* kedua, dan *MLP* dengan *residual connection*, sehingga setiap token membangun representasi kontekstual terhadap seluruh urutan secara iteratif. Setelah seluruh blok selesai, CLS Token Pooling mengekstrak vektor pada posisi token CLS sebagai representasi global gambar, yang kemudian diproyeksikan dari dimensi 768 ke 512 oleh Projection Head untuk disejajarkan ke ruang *embedding* bersama, dan dinormalisasi menjadi unit vektor pada *hypersphere* melalui L2 Normalization sehingga perbandingan representasi dapat dilakukan secara konsisten menggunakan *cosine similarity*. Seluruh parameter *image encoder* dibekukan sepenuhnya selama pelatihan agar untuk mempertahakan kemampuan *pre-train* yang sudah dilatih pada jutaan gambar. Arsitektur *image encoder* ini akan dipasangkan dengan arsitektur *text encoder* yang akan dilatih dengan lingkungan yang berbeda sesuai dengan kebutuhan arsitektur masing masing seperti berikut ini.

Arsitektur *text encoder* dengan Transformer pada CLIP sebagai berikut

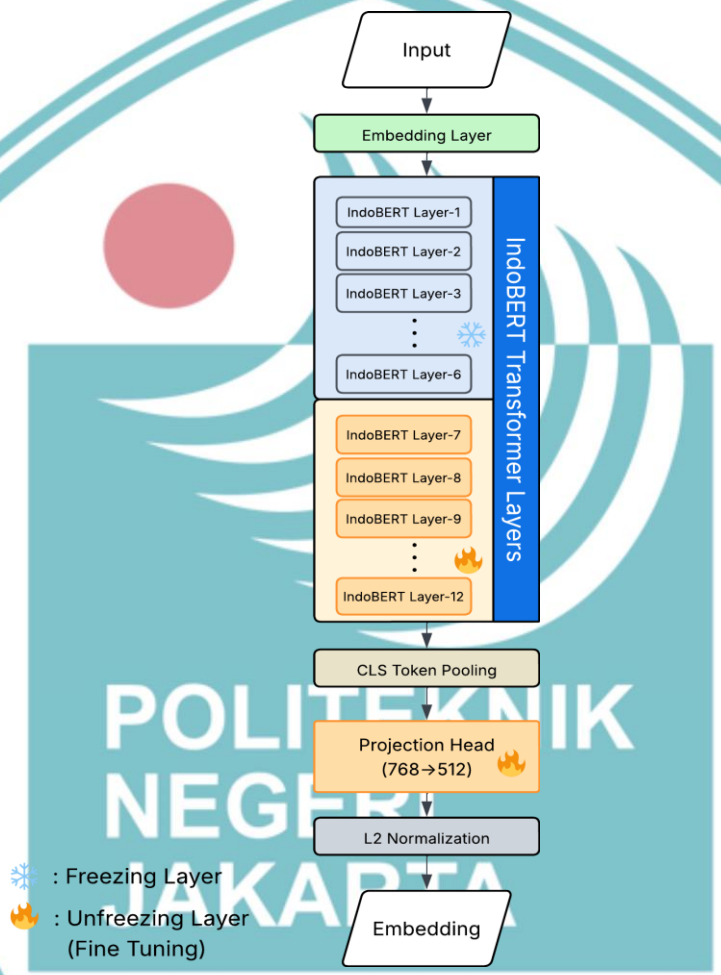


Gambar 4.4 Arsitektur text encoder CLIP standar

Gambar 4.4 menunjukkan arsitektur *text encoder* pada model CLIP menggunakan arsitektur Transformer bawaan CLIP yang dirancang untuk memproses teks secara langsung dalam kerangka *contrastive learning*. Input teks terlebih dahulu diproses melalui *Embedding* Layer yang mengubah token input menjadi representasi vektor sebelum masuk ke dalam 12 lapisan Transformer. Model ini menggunakan tokenizer bawaan CLIP berbasis Byte Pair Encoding (BPE) dengan panjang maksimum 77 token dan menerapkan causal attention mask sehingga setiap token hanya dapat melihat sebelumnya tanpa melihat token setelahnya. Representasi vektor tersebut kemudian masuk pada lapisan CLIP Transformer Layer yang terdiri dari 12 lapisan yang dilatih penuh agar dapat menangkap relasi semantik antar token secara bertahap. Representasi kontekstual yang dihasilkan kemudian diekstraksi melalui EOT Token Pooling untuk menghasilkan vektor representasi yang merepresentasikan seluruh kalimat dalam dimensi 512 secara langsung tanpa projection head tambahan, sehingga *embedding* teks dan gambar dapat dibandingkan langsung dalam ruang vektor yang sama. *Embedding* yang dihasilkan kemudian dinormalisasi menggunakan L2 normalization untuk memastikan seluruh vektor berada pada skala yang seragam sebelum digunakan dalam proses

contrastive learning. Model ini memiliki total 151.277.313 parameter, di mana 63.428.097 parameter (41,9%) dapat dilatih yang seluruhnya berasal dari text encoder, sementara 87.849.216 parameter (58,1%) yang tidak dapat dilatih berasal dari image encoder berbasis ViT-B/32 yang dibekukan seluruh lapisannya.

Arsitektur *text encoder* dengan model IndoBERT pada CLIP sebagai berikut.



Gambar 4.5 Arsitektur text encoder menggunakan IndoBERT

Gambar 4.5 menunjukkan arsitektur *text encoder* pada model CLIP dengan IndoBERT menggunakan IndoBERT, sebuah model bahasa berbasis BERT yang dilatih khusus pada korpus Bahasa Indonesia. *Input* teks terlebih dahulu diproses melalui Embedding Layer yang mengubah token input menjadi representasi vektor padat sebelum masuk ke dalam 12 lapisan Transformer. Strategi pembekuan parsial diterapkan pada text encoder ini, di mana lapisan 1–6 dibekukan untuk mempertahankan representasi linguistik dasar Bahasa Indonesia yang telah dimiliki

model, sementara lapisan 7–12 dibiarkan aktif untuk beradaptasi dengan domain data. Representasi kontekstual yang dihasilkan kemudian diekstraksi melalui CLS Token Pooling untuk menghasilkan vektor representasi kalimat. Karena dimensi output IndoBERT (768) tidak kompatibel dengan ruang *embedding* CLIP (512), sebuah projection head berupa layer linear digunakan untuk mereduksi dimensi tersebut menjadi 512 dan dilatih penuh bersama model. *Embedding* yang dihasilkan. Kemudian output *embedding* dinormalisasi menggunakan L2 normalization untuk memastikan seluruh vektor berada pada skala yang seragam sebelum digunakan dalam proses contrastive learning. Model ini memiliki total 213.275.905 parameter, di mana 82.899.457 parameter (38,9%) dapat dilatih yang mencakup 6 lapisan aktif IndoBERT beserta projection head, sementara 130.376.448 parameter (61,1%) yang tidak dapat dilatih berasal dari ViT-B/32 dan 6 lapisan IndoBERT yang dibekukan.

Arsitektur *text encoder* dengan model XLM-RoBERTa pada CLIP sebagai berikut.

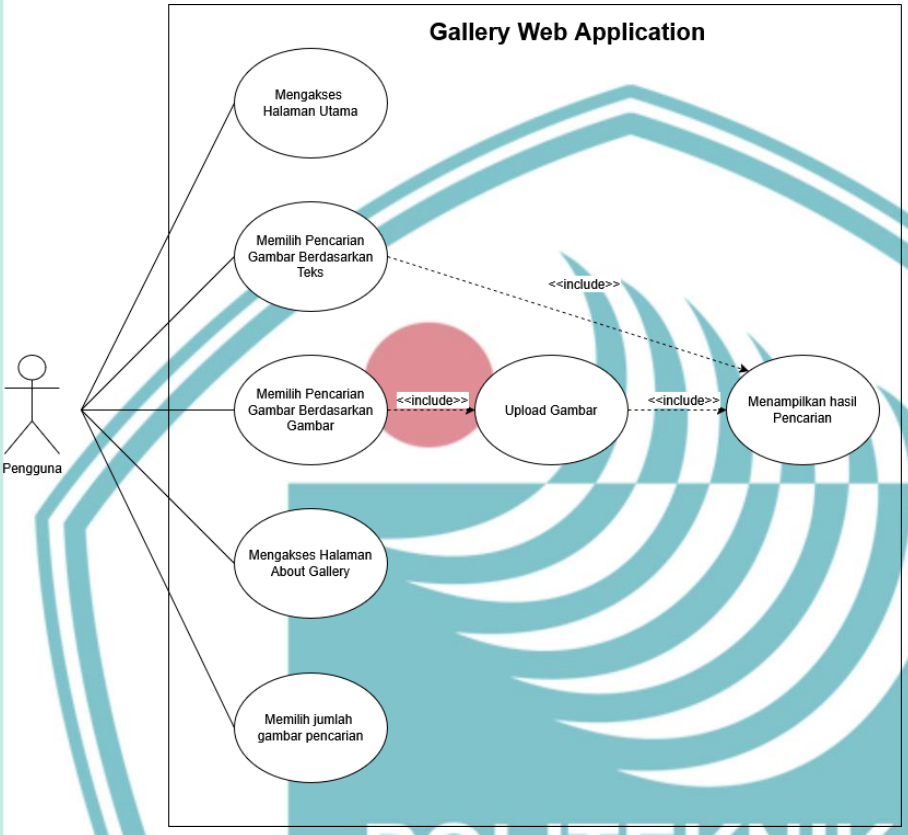
Gambar 4.6 Arsitektur text encoder menggunakan XLM-RoBERTa

Gambar 4.6 menjelaskan arsitektur *text encoder* pada model CLIP dengan XLM-RoBERTa menggunakan XLM-RoBERTa, sebuah model bahasa multibahasa

berskala besar yang dilatih pada korpus 100 bahasa termasuk Bahasa Indonesia. Input teks terlebih dahulu diproses melalui *Embedding* Layer untuk mengubah input menjadi representasi vektor yang kemudian masuk ke dalam 24 lapisan Transformer. Strategi pembekuan yang lebih konservatif diterapkan pada model ini dibandingkan IndoBERT, di mana 18 lapisan pertama dibekukan untuk menjaga representasi multibahasa yang telah dimiliki model, sementara hanya 6 lapisan terakhir diaktifkan untuk beradaptasi dengan domain data. Representasi kontekstual yang dihasilkan kemudian diekstraksi melalui CLS Token Pooling untuk menghasilkan vektor representasi kalimat. Karena dimensi output XLM-RoBERTa (1024) tidak kompatibel dengan ruang *embedding* CLIP (512), sebuah projection head berupa layer linear digunakan untuk mereduksi dimensi tersebut menjadi 512 dan dilatih penuh bersama model. *Embedding* yang dihasilkan kemudian dinormalisasi menggunakan L2 normalization sebelum digunakan dalam proses *contrastive learning*. Model ini memiliki total 560.677.888 parameter, di mana 77.414.400 parameter (13,8%) dapat dilatih yang mencakup 6 lapisan aktif XLM-RoBERTa beserta projection head, sementara 483.263.488 parameter (86,2%) yang tidak dapat dilatih berasal dari ViT-B/32 dan 18 lapisan XLM-RoBERTa yang dibekukan.

4.2.2 Perancangan Web

A. Use Case Diagram



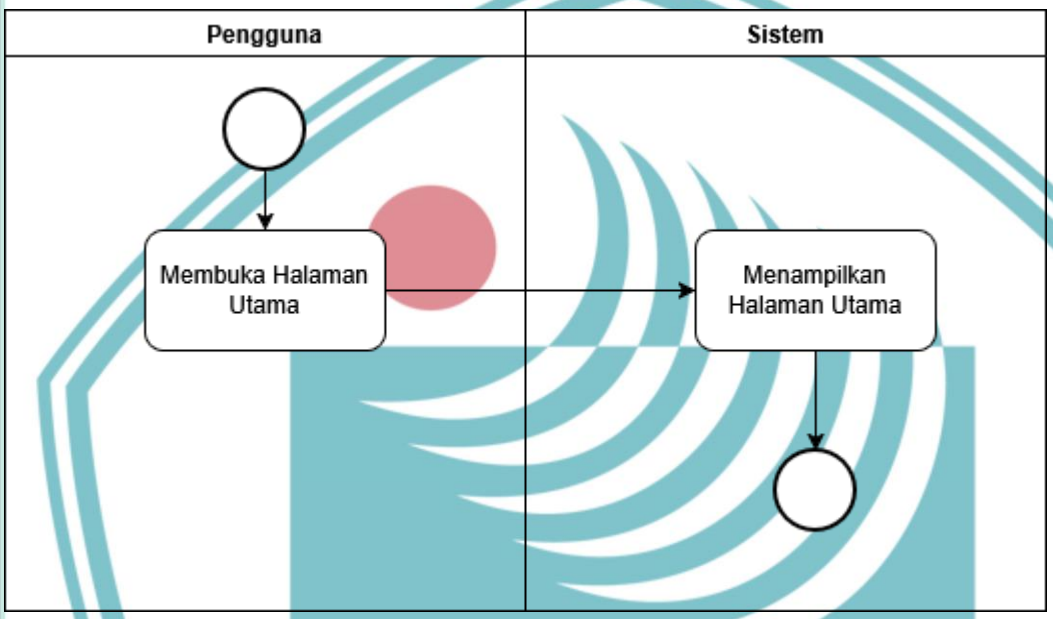
Gambar 4.7 Use Case Diagram

Use case diagram yang dijelaskan pada Gambar 4.7 dibuat berdasarkan kebutuhan fungsional yang sudah ditetapkan sebelumnya. Diagram ini menunjukkan interaksi yang dapat dilakukan aktor pada sistem web yang dikembangkan. Terdapat satu aktor pada sistem web yaitu pengguna (*user*). Interaksi yang dapat dilakukan oleh pengguna pada sistem pencarian gambar mencangkup mengakses halaman utama, memilih gambar berdasarkan teks, memilih gambar berdasarkan gambar, mengakses halaman about gallery dan memilih jumlah gambar pencarian seperti yang dapat diidentifikasi pada Gambar 4.7.

B. Activity Diagram

Activity diagram berisikan penjabaran dari *use case diagram* untuk menunjukkan aktivitas yang lebih rinci

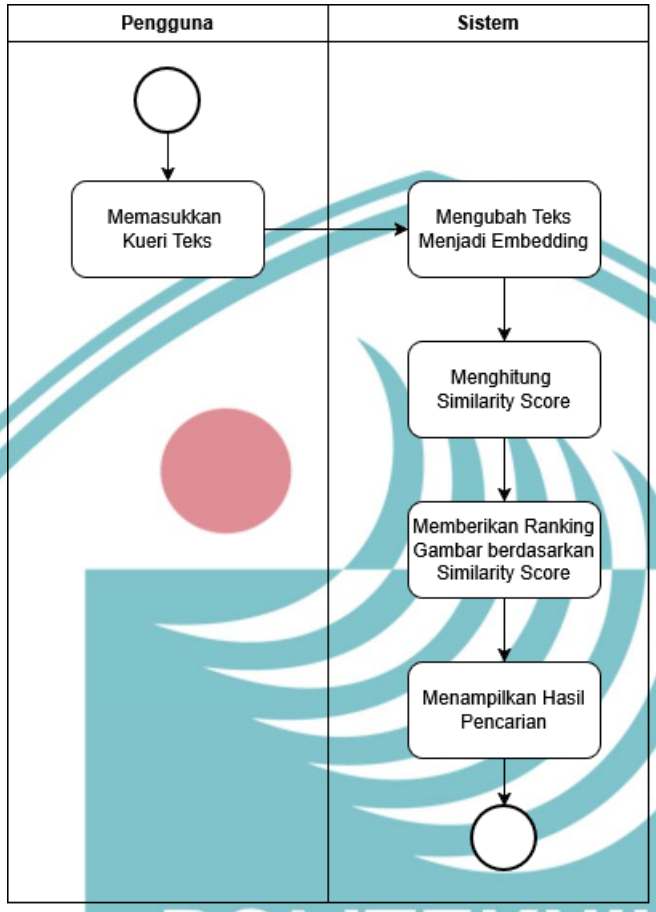
a) Mengakses Halaman Utama



Gambar 4.8 Activity Diagram Halaman Utama

Gambar 4.8 menjelaskan alur awal saat pengguna membuka aplikasi website untuk pertama kalinya. Begitu pengguna masuk, sistem akan langsung memuat seluruh komponen antarmuka (tampilan) halaman. Beberapa elemen yang ditampilkan meliputi menu navigasi (*navigation bar*), judul aplikasi, beserta deskripsi singkatnya. Selain itu, pada halaman utama juga disediakan kotak pencarian (*search bar*) yang memungkinkan pengguna untuk mencari gambar, lengkap dengan opsi pencarian menggunakan teks atau gambar. Setelah semua komponen tersebut berhasil tampil sempurna di layar pengguna, proses awal ini dinyatakan selesai, yang disimbolkan dengan *node* berbentuk lingkaran.

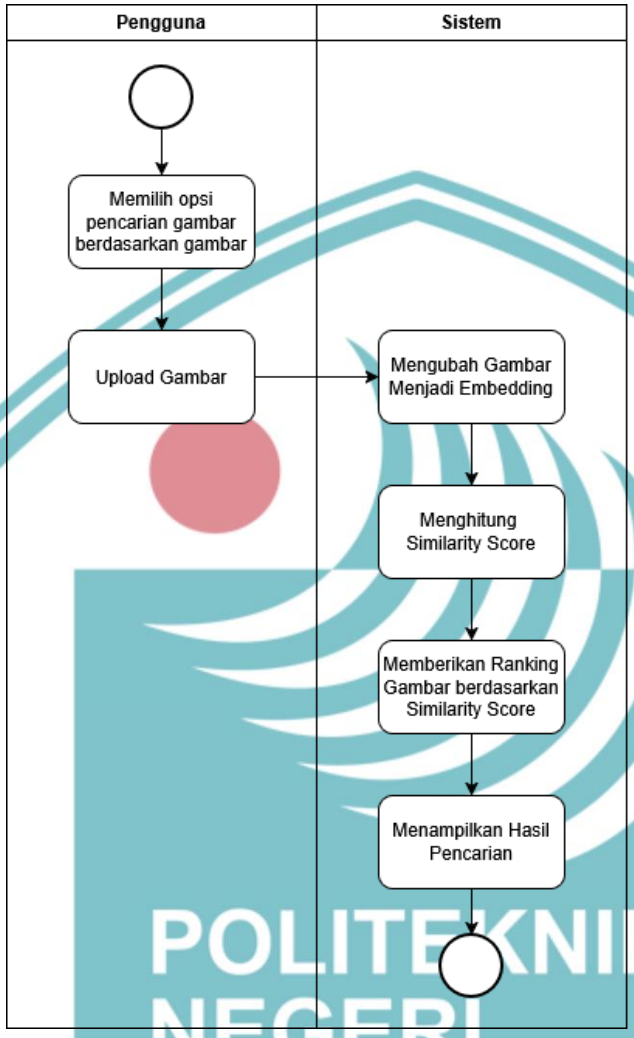
b) Proses Pencarian Gambar Berdasarkan Teks



Gambar 4.9 Activity Diagram Pencarian Berdasarkan Teks

Gambar 4.9 atas menunjukkan proses pengguna ketika melakukan pencarian berdasarkan teks pada aplikasi. Pada tahap ini, pengguna dapat memasukkan input teks sesuai dengan search bar yang disediakan pada halaman. Selanjutnya, sistem mengubah input kueri pengguna menjadi *Embedding* oleh model text encoder. Kemudian model menghitung Similarity Score *embedding* teks dengan *embedding* gambar yang sudah diproses saat aplikasi pertama kali diinisiasi. Kemudian sistem memberikan ranking gambar berdasarkan similarity score yang sudah dihasilkan. Tahap terakhir sistem akan menampilkan gambar gambar relevan sesuai dengan jumlah gambar yang diinginkan pengguna.

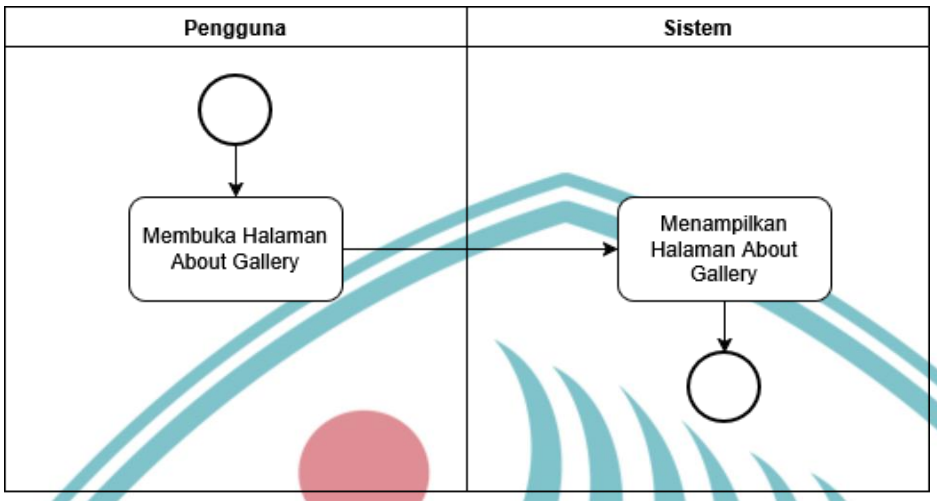
c) Proses Pencarian Gambar Berdasarkan Gambar



Gambar 4.10 Activity Diagram Pencarian Berbasis Gambar

Gambar 4.10 di atas menunjukkan proses pengguna ketika ingin melakukan pencarian gambar berbasis gambar. Pada tahap ini, pengguna dapat memilih opsi pencarian gambar dengan gambar terlebih dahulu karena secara *default* pencarian berdasarkan teks. Setelah itu, pengguna dapat melakukan upload gambar yang pengguna inginkan. Seperti proses pencarian teks, input gambar akan diubah menjadi *embedding* menggunakan model *image encoder*, kemudian *embedding* akan dihitung menggunakan Similarity Score dan diberikan ranking berdasarkan similarity score-nya. Tahap terakhir sistem akan memberikan gambar yang relevan sesuai dengan jumlah gambar yang pengguna inginkan.

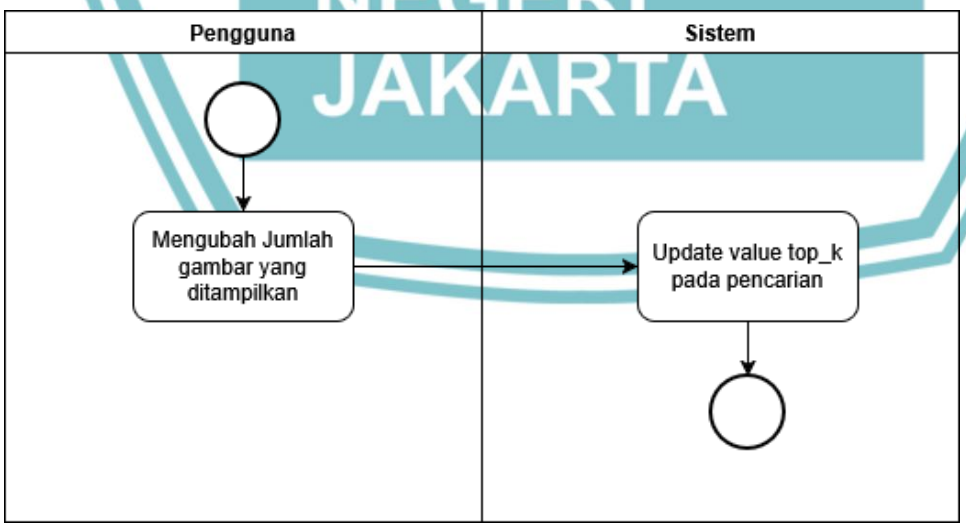
d) Proses Penampilan Halaman About Gallery



Gambar 4.11 Activity Diagram Halaman About Gallery

Gambar 4.11 di atas menunjukkan proses penampilan halaman About Galery saat pengguna melakukan navigasi pada halaman tersebut. Halaman ini berisi informasi terhadap sistem galeri maupun karakteristik gambar dari galeri itu sendiri. Konten di dalamnya mencangkup informasi tentang galeri, karakteristik gambar pada galeri, teknologi yang digunakan dalam galeri serta beberapa contoh dari gambar galeri yang tersedia.

e) Proses mengubah jumlah gambar

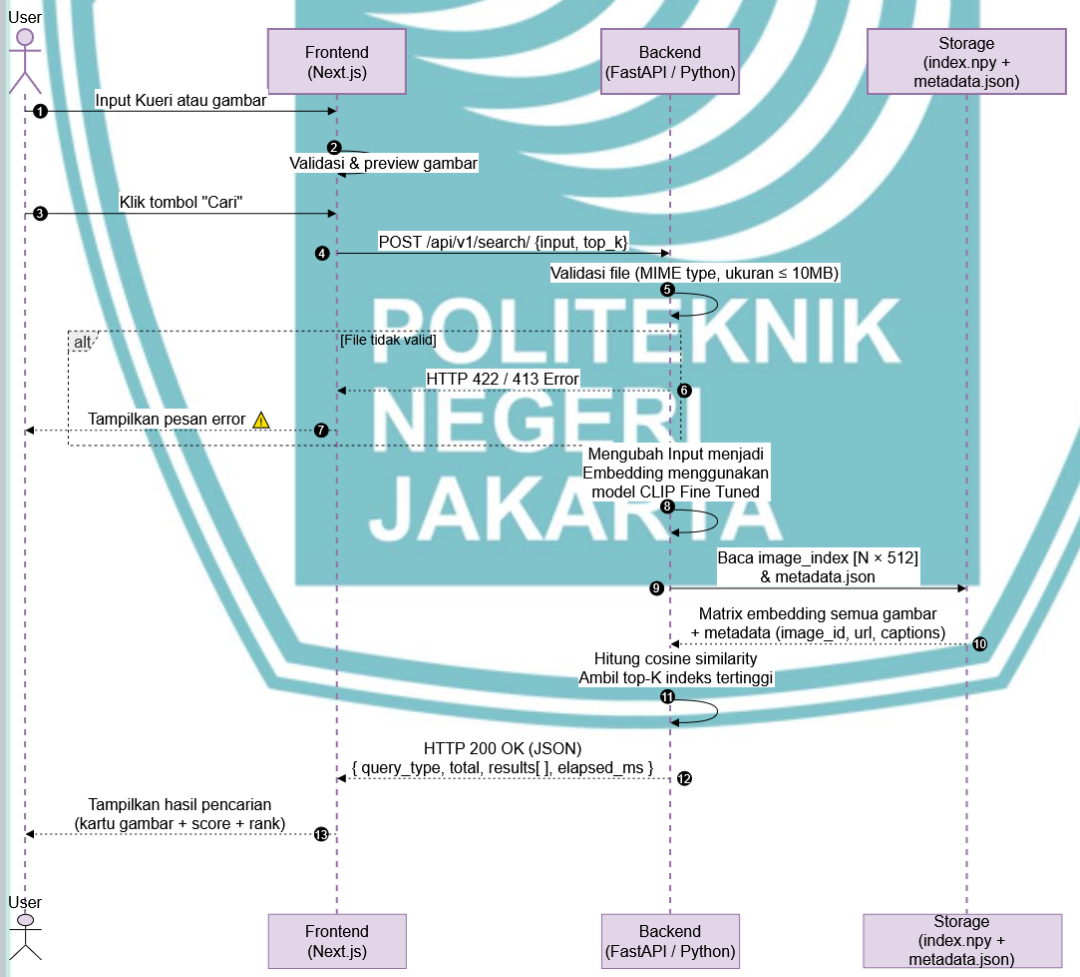


Gambar 4.12 Activity Diagram Pengubahan Jumlah Gambar

Gambar 4.12 di atas merupakan proses perubahan jumlah gambar yang diinginkan pengguna melalui dropdown. Saat pengguna melakukan perubahan, sistem secara otomatis akan mengubah value dari top_k pencarian agar sistem dapat memberikan gambar sesuai dengan jumlah input yang pengguna inginkan.

C. Sequence Diagram

Sequence diagram yang dijelaskan pada Gambar 4.13 menggambarkan urutan interaksi antara pengguna, antarmuka (*frontend*), *backend* dan *storage* dalam sistem pencarian gambar. Proses dimulai pada lapisan antarmuka saat pengguna memasukkan kueri teks atau gambar, kemudian divalidasi formatnya dan ditampilkan sebagai pratinjau. Setelah pencarian dikonfirmasi, kueri diteruskan ke *backend* untuk verifikasi kesesuaian format dan ukuran file.

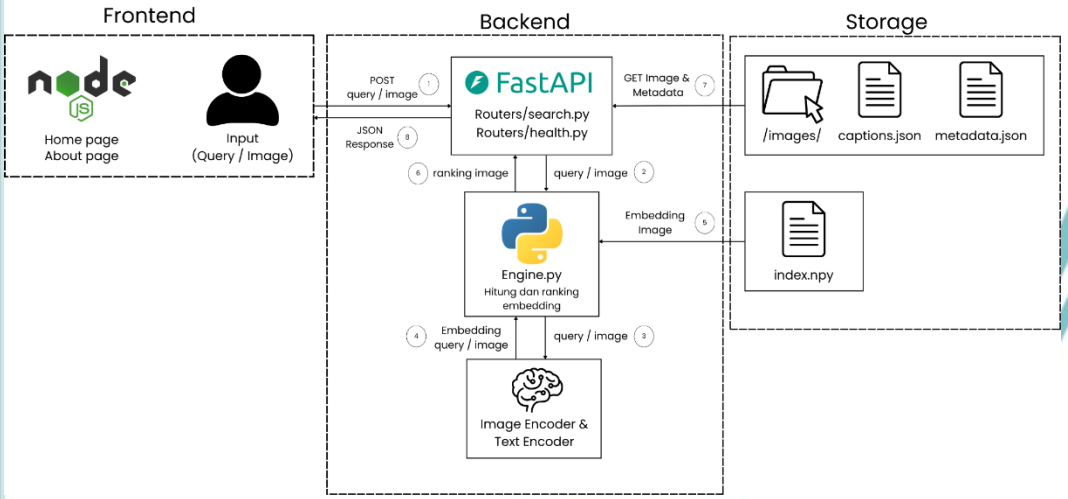


Gambar 4.13 Sequence Diagram

Pada input yang valid, *backend* mengonversi kueri menjadi vektor berdimensi tetap menggunakan model CLIP yang telah di-*fine-tune* dalam bahasa Indonesia. Langkah ini menyelaraskan makna semantik kueri dan gambar ke dalam ruang *embedding* yang sama. Selanjutnya, sistem mengambil seluruh *embedding* serta metadata gambar dari lapisan penyimpanan untuk dihitung tingkat kemiripannya dengan vektor kueri menggunakan metode *cosine similarity*. Terakhir, gambar-gambar dengan skor kemiripan tertinggi diekstraksi sesuai kuota yang diminta, lalu dikirimkan kembali ke antarmuka untuk disajikan kepada pengguna dalam bentuk kartu gambar yang dilengkapi peringkat dan skor relevansi.

D. Arsitektur Sistem

Gambar 4.14 menunjukkan bahwa sistem *Smart Gallery* dibangun dalam tiga lapisan utama yang saling terintegrasi, yaitu frontend berbasis Next.js yang berfungsi sebagai antarmuka pengguna, backend berbasis FastAPI dan Python yang bertugas memproses permintaan dan menghitung kemiripan semantik, serta lapisan penyimpanan yang menyimpan berkas gambar, metadata, dan matriks *embedding*.



Gambar 4.14 Arsitektur Sistem

Proses pencarian diinisiasi ketika pengguna memasukkan kueri berupa teks atau gambar melalui antarmuka frontend, yang kemudian dikirimkan ke lapisan backend melalui permintaan HTTP POST yang diarahkan ke endpoint FastAPI, di mana modul Routers/search.py menerima permintaan tersebut dan meneruskan data kueri ke modul Engine.py untuk diproses lebih lanjut. Setelah menerima data kueri,

Engine.py meneruskannya ke pasangan encoder yang sesuai, yakni Text Encoder untuk masukan berupa teks dan Image Encoder untuk masukan berupa gambar. Encoder kemudian menghasilkan representasi vektor *embedding* berdimensi 512 yang merepresentasikan makna semantik dari kueri pengguna, dan vektor tersebut dikembalikan ke Engine.py untuk tahap perbandingan. Pada tahap perbandingan, Engine.py memuat matriks *embedding* gambar berukuran N×512 yang telah dihitung sebelumnya dari berkas index.npy di lapisan penyimpanan, kemudian menghitung kemiripan kosinus antara vektor *embedding* kueri dengan setiap vektor dalam matriks tersebut. Hasil perhitungan digunakan untuk mengurutkan gambar berdasarkan skor relevansi tertinggi, dan daftar gambar terperingkat tersebut dikembalikan ke FastAPI. Berdasarkan daftar perankingan yang diterima, FastAPI mengambil berkas gambar dari direktori /images/ serta informasi deskriptif yang bersesuaian dari metadata.json dan captions.json. Seluruh hasil pencarian tersebut kemudian disusun ke dalam format JSON dan dikirimkan kembali ke frontend untuk ditampilkan kepada pengguna sebagai hasil pencarian yang telah terurut berdasarkan relevansi.

E. Struktur Database

Database yang digunakan dalam sistem pencarian gambar ini menggunakan *file-based storage* atau penyimpanan lokal untuk mengoptimalkan pencarian vektor dan kemudahan implementasi. Sistem ini terdiri dari empat komponen utama terintegrasi satu sama lain yang dapat dijelaskan pada tabel 4.9 berikut ini

Tabel 4.9 Komponen Struktur Database

Nama Komponen	Tipe	Isi	Peran
/images/	Folder / Direktori	File gambar asli (JPEG/PNG) dengan nama file unik.	Sumber gambar yang ditampilkan ke pengguna pada antarmuka hasil retrieval.
metadata.json	JSON	Metadata deskriptif tiap gambar dalam format key-value	Diakses saat backend perlu mengembalikan informasi gambar ke frontend setelah retrieval berhasil.
captions.json	JSON	Teks caption berbahasa Indonesia yang berpasangan dengan tiap gambar (Dari dataset)	Digunakan untuk melengkapi informasi gambar pada antarmuka hasil pencarian.
index.npy	Numpy / Array	Matriks *embedding* hasil ekstraksi ViT-B/32 pada dataset	Digunakan untuk perhitungan *cosine similarity* dengan teks *embedding* pada memori

Arsitektur database Smart Gallery terdiri dari empat komponen yang saling melengkapi. Direktori /images/ berfungsi sebagai repositori penyimpanan berkas gambar asli, di mana setiap berkas diidentifikasi menggunakan nama berkas unik yang sekaligus berperan sebagai penghubung implisit antarkompon. Berkas metadata.json menyimpan informasi deskriptif setiap gambar dalam format pasangan kunci-nilai yang diakses backend untuk mengembalikan informasi gambar ke antarmuka pengguna setelah proses retrieval selesai, sedangkan berkas captions.json menyimpan teks caption berbahasa Indonesia yang ditampilkan sebagai teks pendamping pada antarmuka hasil pencarian. Komponen paling kritis dalam arsitektur ini adalah berkas index.npy, yang menyimpan matriks *embedding* berukuran N×512 hasil ekstraksi fitur visual menggunakan *image encoder* ViT-B/32 dan menjadi basis seluruh operasi pencarian kemiripan (*cosine similarity*) antara *query embedding* dan *image embedding* secara langsung di dalam memori sistem.

4.3 Implementasi Sistem

Hasil dari perancangan sistem yang sudah dibahas akan diimplementasikan ke dalam sistem yang dapat diakses oleh pengguna. Implementasi sistem mencangkup tiga proses yaitu persiapan dataset, pelatihan model dan website pencarian gambar. Ketiga proses ini dipisah untuk memastikan dataset yang digunakan sudah diproses sesuai kebutuhan dan model yang digunakan dalam website pencarian model adalah model hasil pelatihan terbaik. Kemudian, Model yang terbaik akan diintegrasikan pada *website* pencarian gambar.

4.3.1 Persiapan Dataset

Pengumpulan dataset

Dataset yang digunakan pada penelitian ini adalah Flickr8K Indonesia, sebuah dataset image captioning berbahasa Indonesia yang dikonstruksi melalui proses penerjemahan dan adaptasi dari bahasa Inggris ke dalam Bahasa Indonesia. Dataset ini terdiri atas 8.091 gambar fotografi yang mencakup beragam subjek dan konteks visual kehidupan sehari-hari, dengan masing-masing gambar dilengkapi oleh 5 caption deskriptif. Sama seperti Flickr8K, Flickr30K berisi terdiri dari 31.014 gambar dengan masing-masing gambar dilengkapi 5 caption deskriptif berbahasa Indonesia. Contoh pasangan gambar dan caption dapat dilihat pada tabel 4.10 berikut ini

Tabel 4.10 Contoh Dataset Gambar dan Caption

Gambar	Caption Bahasa Indonesia
	Seorang anak laki-laki tersenyum di depan tembok batu di kota.
	Seorang anak laki-laki berdiri di jalan sementara seorang lelaki sedang bekerja di dinding batu.
	Seorang anak laki-laki berlari di seberang jalan.
	Seorang anak kecil sedang berjalan di jalan beraspal dengan tiang logam dan seorang

Gambar	Caption Bahasa Indonesia
	pria di belakangnya.
	Anak laki-laki tersenyum dengan kemeja putih dan celana jeans biru di depan dinding batu dengan pria di belakangnya.
	Seekor anjing hitam putih berusaha menangkap benda berwarna kuning dan ungu di halaman yang dipotong rendah.
	Seekor anjing hitam putih melompat mengejar mainan kuning.
	seekor anjing hitam putih melompat untuk mendapatkan Frisbee.
	Seekor anjing hitam melompat untuk menangkap mainan berwarna ungu dan hijau.
	Seekor anjing melompat untuk menangkap mainan.

Seperti yang terlihat pada gambar 4.15, dataset caption sudah berbentuk csv dengan kolom file_name yang menjelaskan nama file gambar dan kolom text yang berisi caption berbahasa Indonesia

```
file_name,text
1000268201_693b08cb0e.jpg,Seorang anak berpakaian merah muda sedang menaiki tangga di jalan masuk.
1000268201_693b08cb0e.jpg,Seorang gadis memasuki sebuah bangunan kayu.
1000268201_693b08cb0e.jpg,Seorang gadis kecil naik ke rumah bermain kayu.
1000268201_693b08cb0e.jpg,Seorang gadis kecil menaiki tangga menuju rumah bermainnya.
1000268201_693b08cb0e.jpg,Seorang gadis kecil berpakaian merah muda masuk ke kabin kayu.
1001773457_577c3a7d70.jpg,Seekor anjing hitam dan anjing tutul sedang berkelahi
1001773457_577c3a7d70.jpg,Seekor anjing hitam dan anjing tiga warna bermain satu sama lain di jalan.
1001773457_577c3a7d70.jpg,Seekor anjing hitam dan seekor anjing putih dengan bintik-bintik coklat saling menatap di jalan.
1001773457_577c3a7d70.jpg,Dua anjing dari ras berbeda saling memandang di jalan.
1001773457_577c3a7d70.jpg,Dua anjing di trotoar bergerak menuju satu sama lain.
1002674143_1b742ab4b8.jpg,Seorang gadis kecil berlumuran cat duduk di depan lukisan pelangi dengan tangannya di dalam mangkuk.
1002674143_1b742ab4b8.jpg,Seorang gadis kecil sedang duduk di depan lukisan pelangi besar.
1002674143_1b742ab4b8.jpg,Seorang gadis kecil di rerumputan bermain dengan cat jari di depan kanvas putih dengan pelangi di atasnya.
1002674143_1b742ab4b8.jpg,Ada seorang gadis berkuncir duduk di depan lukisan pelangi.
1002674143_1b742ab4b8.jpg,Gadis muda dengan kuncir melukis di luar di rumput.
```

Gambar 4.15 Contoh Dataset Caption

Pemrosesan Data

Tahap pemrosesan data gambar dalam dataset ini menggunakan dua skema transformasi yang berbeda antara data latih dan data evaluasi seperti yang terlihat pada gambar 4.16. Pada data latih, diterapkan augmentasi acak berupa *random resized crop*, *horizontal flip*, dan *color jitter* untuk meningkatkan variasi data dan mengurangi risiko *overfitting*. Seluruh gambar kemudian di-*resize* dan di-*crop* ke resolusi 224×224 piksel, lalu dinormalisasi menggunakan nilai *mean* dan *standar deviasi* khusus CLIP agar distribusi piksel sesuai dengan kondisi saat pra-pelatihan model.

```
_CLIP_MEAN = (0.48145466, 0.4578275, 0.40821073)
_CLIP_STD  = (0.26862954, 0.26130258, 0.27577711)

IMG_TRANSFORM = T.Compose([
    T.Resize(224, interpolation=T.InterpolationMode.BICUBIC),
    T.CenterCrop(224),
    T.ToTensor(),
    T.Normalize(mean=_CLIP_MEAN, std=_CLIP_STD),
])

IMG_TRANSFORM_TRAIN = T.Compose([
    T.RandomResizedCrop(224, scale=(0.85, 1.0),
                        interpolation=T.InterpolationMode.BICUBIC),
    T.RandomHorizontalFlip(p=0.3),
    T.ColorJitter(brightness=0.1, contrast=0.1),
    T.ToTensor(),
    T.Normalize(mean=_CLIP_MEAN, std=_CLIP_STD),
])
```

Gambar 4.16 Kode Pemrosesan Data Gambar

Pada proses preprocessing data gambar sebelum dan sesudah dapat divisualisasikan dengan tabel 4.11 berikut ini

Tabel 4.11 Contoh Pemrosesan Dataset Gambar

Kode	Dataset	Gambar
DP-1	Gambar Mentah Dataset	

DP-2	Gambar Preprocessed data Validation / Test (Resize, Center Crop)	
DP-3	Gambar Preprocessed data Train (Flip, ColorJitter, RandomResizeCrop)	

Kemudian, untuk data caption pada dataset ini dilakukan pembersihan special character untuk memastikan bahwa caption untuk memastikan kualitas tokenisasi caption. Tahap tokenization teks disesuaikan dengan kebutuhan masing-masing konfigurasi model: CLIP menggunakan BPE Tokenization, IndoBERT menggunakan WordPiece Tokenization, XLM-RoBERTa menggunakan Sentence Piece Tokenization. Pemilihan tokenisasi ini dilakukan secara otomatis menggunakan *AutoTokenizer* dari *libary* hugging face seperti gambar 4.17

```
tokenizer = AutoTokenizer.from_pretrained(CONFIG["model_name"])
```

Gambar 4.17 Tokenization Sesuai dengan Nama Model

Proses tokenisasi untuk model selain CLIP dilakukan dengan mengubah setiap caption menjadi representasi token berupa input_ids dan attention_mask. Sedangkan model CLIP mengubah caption menjadi representasi token berupa token_ids tanpa attention mask tambahan sesuai dengan kebutuhan arsitektur CLIP.

Pembatasan Dataset bersih kemudian dibagi dengan rasio 90% untuk data training, 5% untuk data evaluasi, dan 5% untuk data tes. Rasio ini dipilih untuk memaksimalkan jumlah data training sesuai dengan pembagian data yang sudah dijadikan benchmark dataset Flickr. Kemudian pembagian dataset disimpan menggunakan data loader sehingga memungkinkan proses pembacaan, pengacakan, dan pembentukan batch data serta pengaturan batch yang tersisa dilakukan secara otomatis dan lebih efisien seperti yang dapat dilihat pada gambar berikut.

```
train_loader = DataLoader(train_dataset, batch_size=CONFIG['batch_size'],
                          shuffle=True,  num_workers=4, pin_memory=True, drop_last=True)
val_loader   = DataLoader(val_dataset,   batch_size=CONFIG['batch_size'],
                          shuffle=False, num_workers=4, pin_memory=True)
test_loader  = DataLoader(test_dataset,  batch_size=CONFIG['batch_size'],
                          shuffle=False, num_workers=4, pin_memory=True)
```

Gambar 4.18 Kode Data Loader Dataset

Gambar 4.18 menunjukkan kode *train loader* yang menggunakan batch_size mengontrol jumlah sampel yang diproses sekaligus dalam satu iterasi pelatihan. Kemudian, shuffle=True dan drop_last=True untuk memastikan variasi urutan data setiap epoch serta membuang batch terakhir jika tidak penuh, sehingga ukuran batch tetap konsisten selama pelatihan, dengan pin_memory=True dan num_workers=4 untuk mempercepat transfer data ke GPU. Validation dan test loader menggunakan shuffle=False karena urutan data tidak mempengaruhi hasil evaluasi.

4.3.2 Implementasi Model

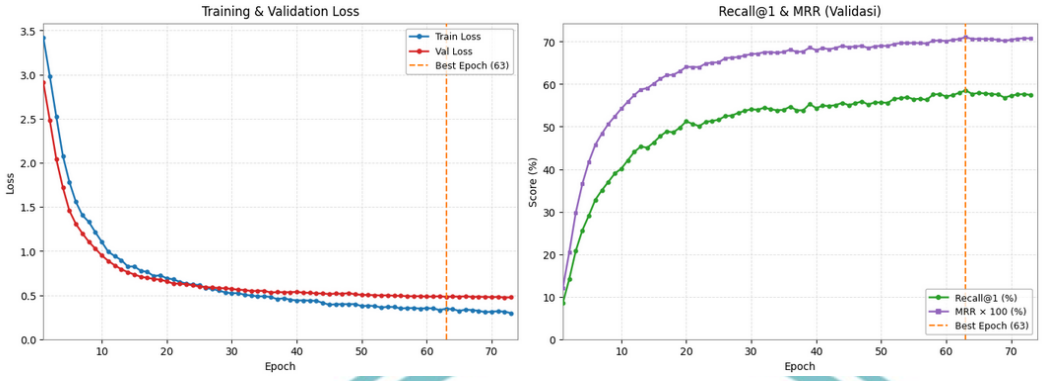
Implementasi Model CLIP

Model CLIP standar dilatih dengan menggunakan arsitektur yang sudah dijelaskan pada tahapan penelitian sebelumnya yaitu transformer yang seluruh lapisannya dilakukan pelatihan penuh. Parameter yang digunakan dalam pelatihan model CLIP standar ini dapat dijelaskan pada Tabel 4.12 berikut ini

Tabel 4.12 Parameter Model CLIP

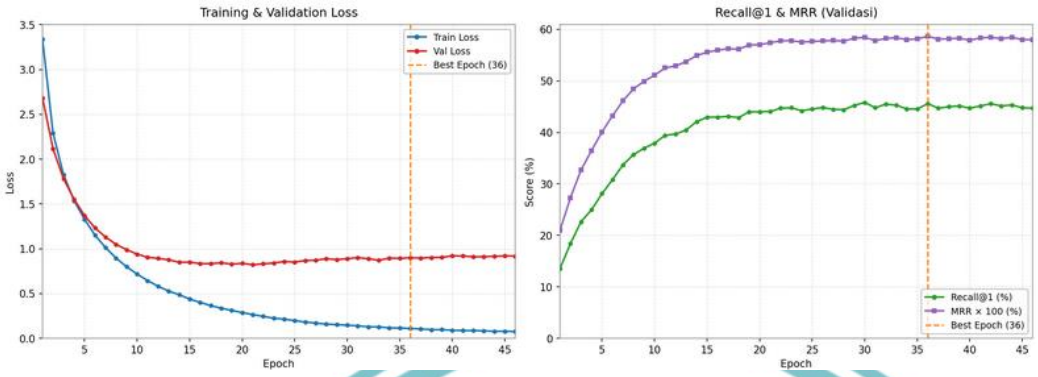
Nama Parameter	Keterangan
Optimizer	AdamW
Weight decay	0.01
Max Length	77
Batch Size	64
Learning Rate Model	5e-7
Temperature InfoNCE	0.07
Epochs	100
EarlyStop	Patience = 10 dan restore_best_weight = True

Hyperparameter yang dijelaskan pada Tabel 4.12 diterapkan secara konsisten pada proses fine-tuning model CLIP. Optimizer AdamW digunakan karena mampu mengatasi overfitting melalui mekanisme weight decay serta memberikan stabilitas pada proses pelatihan model. Nilai learning rate sebesar 5e-7 dipilih agar proses pembaruan bobot berlangsung secara bertahap sehingga tidak merusak representasi pretrained yang sudah dimiliki model CLIP. Temperature InfoNCE sebesar 0.07 digunakan untuk mengatur sensitivitas perhitungan similarity antara *embedding* gambar dan teks pada *contrastive learning*. Jumlah epoch ditetapkan sebanyak 100 agar model memiliki cukup iterasi untuk mempelajari pola data secara optimal. Mekanisme Early Stopping diterapkan dengan memantau nilai Recall@1 pada data validasi, di mana pelatihan dihentikan apabila tidak terjadi peningkatan dalam 10 epoch berturut-turut dan bobot model terbaik dikembalikan melalui restore_best_weight = True untuk mencegah overfitting.



Gambar 4.19 Grafik Pelatihan Model CLIP pada Dataset Flickr8K

Gambar 4.19 menunjukkan proses pelatihan model CLIP pada dataset Flickr8K berlangsung selama 73 epoch dengan menunjukkan tren konvergensi yang konsisten dan stabil. Nilai *training loss* mengalami penurunan yang signifikan dari 3.4200 pada epoch pertama hingga mencapai 0.2965 pada epoch terakhir, sementara *validation loss* juga menurun secara berkelanjutan dari 2.9100 menuju 0.4752 tanpa menunjukkan tanda pembalikan arah yang berarti. Gap antara *training loss* dan *validation loss* cenderung menyempit secara bertahap seiring bertambahnya epoch, yang mengindikasikan bahwa model mampu melakukan generalisasi dengan baik terhadap data yang belum pernah dilihat sebelumnya tanpa kecenderungan *overfitting* yang signifikan. Sejalan dengan penurunan loss tersebut, metrik Recall@1 meningkat secara progresif dari 8.50% hingga mencapai kisaran 57–58% pada epoch-epoch akhir, begitu pula MRR yang naik dari 0.1200 menuju 0.7077 dengan laju peningkatan yang paling tajam terjadi pada 30 epoch pertama sebelum berangsur melambat dan memasuki fase plateau yang landai. Pola ini menunjukkan bahwa kapasitas representasi CLIP pada dataset skala kecil seperti Flickr8K masih terus berkembang sepanjang pelatihan, meskipun marginal *gain* semakin mengecil di epoch lanjutan.



Gambar 4.20 Grafik Pelatihan Model CLIP pada Dataset Flickr30K

Gambar 4.20 menunjukkan pelatihan model CLIP pada dataset Flickr30K berlangsung selama 46 epoch dan memperlihatkan pola divergensi antara *training loss* dan *validation loss* yang mulai tampak sejak pertengahan pelatihan. *Training loss* turun secara konsisten dari 3.3380 hingga 0.0752, namun *validation loss* yang semula menurun beriringan hingga sekitar epoch ke-16 mulai menunjukkan stagnansi bahkan kenaikan gradual pada epoch-epoch berikutnya, mencapai nilai di atas 0.90 pada akhir pelatihan. Pola divergensi ini merupakan indikasi *overfitting*, di mana model semakin mengoptimalkan representasi pada data latih namun kehilangan kemampuan generalisasinya. Meskipun demikian, metrik retrieval Recall@1 tetap meningkat dari 13.53% hingga sekitar 45% dan MRR dari 0.2096 hingga kisaran 0.58, namun keduanya memasuki plateau yang relatif lebih awal dibandingkan model yang sama pada Flickr8K, yakni sejak sekitar epoch ke-20. Kondisi ini mengindikasikan bahwa meskipun loss divergen, representasi semantik yang telah terbentuk pada fase awal pelatihan cukup memadai untuk mempertahankan kualitas retrieval, namun potensi peningkatan lebih lanjut terhambat oleh *overfitting* pada ruang loss.

Implementasi Model CLIP dengan IndoBERT

Model CLIP dengan IndoBERT juga dilatih dengan menggunakan arsitektur yang sudah dijelaskan pada tahapan penelitian sebelumnya. Parameter yang digunakan dalam pelatihan model dijelaskan pada Tabel 4.13 berikut ini.

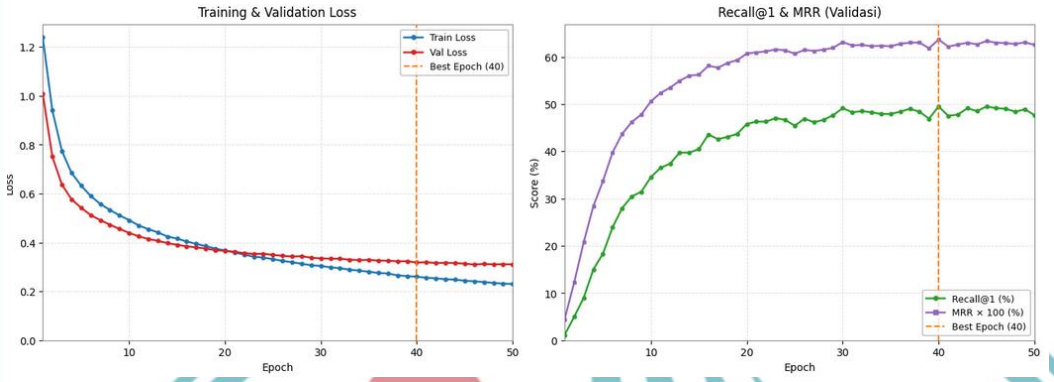
Tabel 4.13 Parameter Model CLIP dengan IndoBERT

Nama Parameter	Keterangan
Optimizer	AdamW
Weight Decay	0.01
Max Length	128
Batch Size	64
Learning Rate Model	2e-5
Learning Rate Proj.	1e-4
Temperature InfoNCE	0.07
Epochs	100
EarlyStop	Patience=10 dan restore_best_weight = True

Proses pelatihan model ini dimulai dengan mengkonfigurasi optimizer menggunakan AdamW. Optimizer AdamW dengan *weight decay* dipilih karena mampu memberikan proses optimasi yang stabil serta membantu mengurangi risiko overfitting selama pelatihan model. Nilai max length 128 digunakan mengikuti batas maksimum token yang didukung oleh arsitektur IndoBERT. Batch size sebesar 64 dipilih dengan mempertimbangkan keseimbangan antara efisiensi pelatihan dan kapasitas memori GPU yang tersedia.

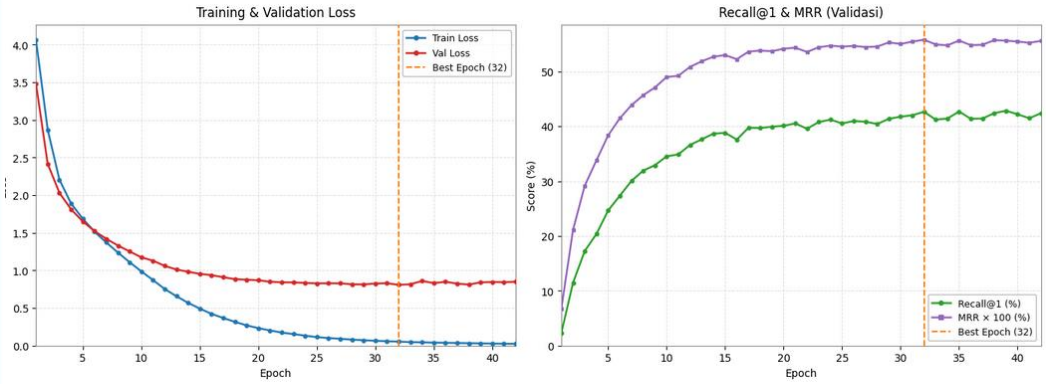
Pada konfigurasi ini, learning rate model ditetapkan sebesar 2e-5 untuk menyesuaikan proses fine-tuning pada encoder pretrained, sedangkan learning rate projection layer menggunakan nilai yang lebih tinggi, yaitu 1e-4 atau 5x dari learning rate model yang sudah ditetapkan. Perbedaan ini diterapkan karena projection layer dilatih dari awal sehingga membutuhkan penyesuaian bobot yang lebih cepat. Temperature InfoNCE sebesar 0.07 digunakan untuk mengatur sensitivitas perhitungan similarity antara *embedding* gambar dan teks pada *contrastive learning*. Jumlah epoch ditentukan sebanyak 100 untuk memberikan kesempatan model mempelajari pola hubungan multimodal secara lebih baik. Selain itu, diterapkan mekanisme Early Stopping dengan patience 10 dan

restore_best_weight = True untuk mengembalikan bobot model terbaik yang diperoleh selama proses training.



Gambar 4.21 Grafik Pelatihan Model CLIP IndoBERT pada Dataset Flickr8K

Gambar 4.21 menunjukkan pelatihan model CLIP dengan IndoBERT pada dataset Flickr8K selama 50 epoch menunjukkan karakteristik yang paling menonjol berupa gap antara *training loss* dan *validation loss* yang sangat tipis dan konsisten sepanjang pelatihan. *Training loss* turun dari 1.2402 menuju 0.2303, sementara *validation loss* yang dimulai dari 1.0093 turun menjadi 0.3105, dengan selisih keduanya yang hampir seragam di setiap epoch tanpa tanda-tanda pembesaran gap yang berarti. Kondisi gap yang tipis ini secara kuat mengindikasikan generalisasi model yang sangat baik, di mana IndoBERT sebagai encoder teks mampu menghasilkan representasi yang tidak hanya fit terhadap data latih tetapi juga robust terhadap data validasi. Metrik Recall@1 meningkat pesat pada 20 epoch pertama hingga mencapai kisaran 45–47%, kemudian melambat dan berfluktuasi tipis di sekitar 47–49% pada epoch selanjutnya. Hal serupa terjadi pada MRR yang naik dari 0.0442 menuju plateau di kisaran 0.62–0.63. Kombinasi gap loss yang tipis dengan plateau metrik yang relatif lebih awal ini menunjukkan bahwa model telah mencapai batas kapasitas representasinya pada skala dataset ini meski tanpa masalah *overfitting*.



Gambar 4.22 Pelatihan Model CLIP IndoBERT pada Dataset Flickr30K

Pelatihan model CLIP dengan IndoBERT pada dataset Flickr30K selama 42 epoch memperlihatkan fenomena *overfitting* yang paling mencolok di antara seluruh konfigurasi yang diuji. Seperti yang terlihat pada Gambar 4.22, *Training loss* turun secara agresif dan berkelanjutan dari 4.0598 hingga 0.0242, sementara *validation loss* yang semula menurun beriringan hingga sekitar epoch ke-20 kemudian memasuki fase plateau dan bahkan menunjukkan fluktuasi kecil yang tidak stabil pada separuh akhir pelatihan. Gap yang melebar drastis antara kedua kurva loss ini ditambah dengan *training loss* hampir menyentuh nol sementara *validation loss* tertahan di atas 0.80 merupakan sinyal *overfitting* yang kuat, mengindikasikan bahwa model mulai menghafal pola spesifik data latih alih-alih mempelajari representasi semantik yang general. Meskipun loss mengalami divergensi yang signifikan, metrik Recall@1 dan MRR justru tetap menunjukkan peningkatan bertahap hingga mencapai plateau masing-masing di kisaran 41–43% dan 0.55–0.56, sebelum kemudian landai pada separuh akhir pelatihan. Fenomena ini mengindikasikan representasi yang terbentuk pada ruang *embedding* masih memiliki struktur semantik yang cukup untuk mempertahankan performa retrieval, meskipun optimasi loss terus berlanjut secara sepihak pada data latih.

Implementasi Model CLIP dengan XLM-RoBERTa

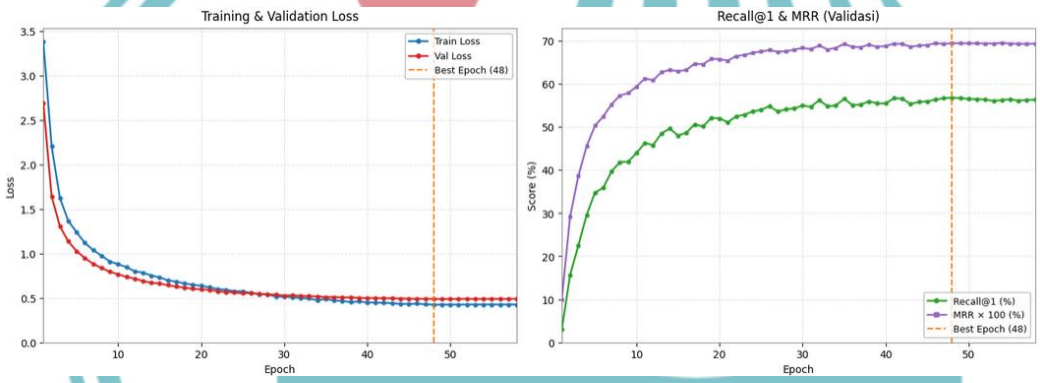
Model CLIP dengan XLM-RoBERTA juga dilatih dengan menggunakan arsitektur yang sudah dijelaskan pada tahapan penelitian sebelumnya. Parameter yang digunakan dalam pelatihan model sebagai berikut.

Tabel 4.14 Parameter Model CLIP dengan XLM-RoBERTA

Nama Parameter	Keterangan
Optimizer	AdamW
Weight Decay	0.01
Max Length	512
Batch Size	64
Learning Rate Model	1e-5
Learning Rate Proj.	5e-5
Temperature InfoNCE	0.07
Epochs	100
EarlyStop	Patience = 10 dan restore_best_weight = True

Parameter pada Tabel 4.14 digunakan pada proses fine-tuning model CLIP dengan XLM-RoBERTA yang dirancang untuk menangani representasi multimodal multibahasa. Optimizer AdamW digunakan karena mampu menjaga stabilitas proses pelatihan sekaligus mengurangi risiko overfitting melalui penerapan weight decay. Nilai max length 512 sesuai dengan batas token maksimum yang digunakan oleh arsitektur XLM-RoBERTa. Batch size sebesar 64 dipilih agar proses training tetap efisien dan dapat berjalan optimal sesuai kapasitas GPU yang tersedia. Learning rate model pada CLIP dengan XLM-RoBERTA ditetapkan sebesar 1e-5, lebih rendah dibandingkan konfigurasi CLIP dengan IndoBERT yang menggunakan 2e-5. Perbedaan ini disebabkan oleh karakteristik CLIP dengan XLM-RoBERTA yang telah melalui proses pretraining multibahasa dengan alignment teks dan gambar yang lebih kompleks. Learning rate yang lebih rendah juga dapat mencegah *catastrophic forgetting* ketika proses training karena dengan bobot learning rate yang kecil, pelatihan berlangsung lebih hati-hati sehingga kemampuan representasi multibahasa yang sudah dimiliki CLIP dengan XLM-RoBERTA tetap terjaga selama fine-tuning. Sementara itu, learning rate projection layer menggunakan nilai 5e-5 karena layer tersebut dilatih dari awal dan

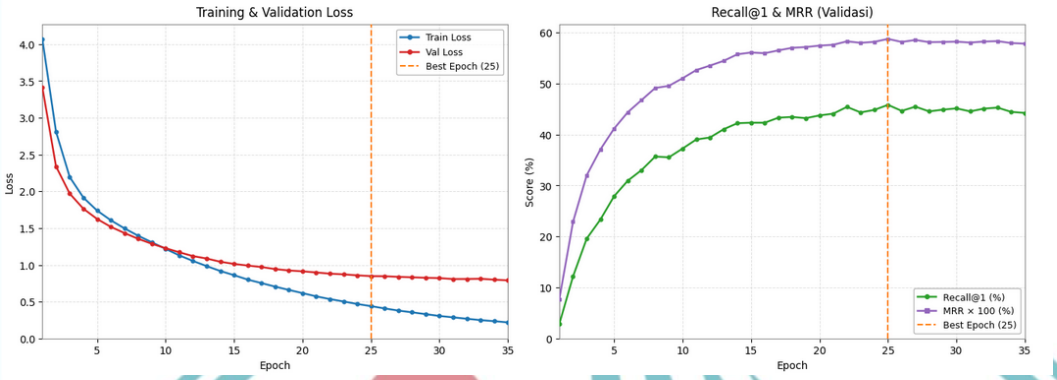
memerlukan penyesuaian parameter yang lebih cepat dibandingkan encoder utama pretrained. Seperti sebelumnya, Temperature InfoNCE sebesar 0.07 digunakan untuk mengontrol distribusi similarity *embedding* antara teks dan gambar sehingga proses *contrastive learning* dapat menghasilkan representasi multimodal yang lebih baik. Jumlah epoch ditetapkan sebanyak 100 agar model memiliki cukup iterasi dalam mempelajari pola hubungan antara data visual dan teks. Selain itu, mekanisme Early Stopping dengan patience 10 dan restore_best_weight = True diterapkan untuk menghentikan proses pelatihan ketika performa model tidak lagi mengalami peningkatan dalam beberapa epoch berturut-turut, sekaligus mengembalikan bobot terbaik yang diperoleh selama training.



Gambar 4.23 Pelatihan Model CLIP XLM-RoBERTa pada Dataset Flickr8K

Gambar 4.23 menunjukkan pelatihan model CLIP dengan XLM-RoBERTa pada dataset Flickr8K berlangsung selama 58 epoch dan menampilkan pola yang unik serta anomalistik pada fase akhir pelatihan. Pada separuh pertama, *training loss* menurun dengan wajar dari 3.3779 menuju kisaran 0.43–0.45 sekitar epoch ke-40, sementara *validation loss* turun lebih konsisten dari 2.6861 menuju 0.49. Namun mulai sekitar epoch ke-44 hingga akhir, *training loss* memperlihatkan yang stagnan di sekitar 0.4257–0.4264 secara berulang, sementara *validation loss* masih terus menurun secara perlahan. Pola stagnasi *training loss* ini mengindikasikan adanya kemungkinan *learning rate* yang terlalu kecil, menyebabkan batas kapasitas optimasi yang dicapai pada konfigurasi pelatihan yang digunakan. Terlepas dari anomali tersebut, metrik retrieval justru terus menunjukkan peningkatan; Recall@1 mencapai puncak di sekitar 56–57% dan MRR di kisaran 0.694–0.695 pada epoch-epoch akhir, yang mengindikasikan bahwa meskipun loss berhenti bergerak, ruang

embedding masih mengalami penyempurnaan marginal yang berdampak positif terhadap kualitas retrieval.



Gambar 4.24 Pelatihan Model CLIP XLM-RoBERTa pada Dataset Flickr30K

Gambar 4.24 menunjukkan pelatihan model CLIP dengan XLM-RoBERTa pada dataset Flickr30K selama 35 epoch menunjukkan pola konvergensi yang paling sehat dan stabil di antara seluruh konfigurasi berbasis encoder teks yang dievaluasi pada dataset tersebut. *Training loss* menurun secara mulus dari 4.0734 menuju 0.2186, sementara *validation loss* juga turun secara konsisten dari 3.4180 hingga 0.7899 tanpa menampakkan plateau prematur maupun divergensi yang signifikan sepanjang 35 epoch pelatihan. Gap antara kedua kurva loss memang terbuka seiring berjalannya pelatihan, namun tidak disertai pembalikan arah pada *validation loss* seperti yang terjadi pada konfigurasi CLIP IndoBERT di dataset yang sama. Metrik Recall@1 meningkat secara progresif dari 2.85% menuju kisaran 44–46% dan MRR dari 0.0771 menuju kisaran 0.58, dengan laju peningkatan yang paling tajam terjadi pada 15 epoch pertama sebelum kemudian melandai dan memasuki plateau yang relatif stabil. Perbandingan dengan CLIP dengan IndoBERT pada Flickr30K menunjukkan bahwa XLM-RoBERTa menghasilkan dinamika pelatihan yang lebih terkendali pada dataset berskala lebih besar, meskipun nilai absolut Recall@1 dan MRR yang dicapai keduanya berada pada rentang yang berdekatan, mengindikasikan bahwa skala dataset Flickr30K mampu memitigasi sebagian perbedaan kapasitas representasi antara kedua encoder teks tersebut.

A. Evaluasi Model

Evaluasi Metrik Ketiga Model

Pengukuran kemampuan model terhadap tugas *cross-modal retrieval* dilakukan menggunakan dataset test yang terpisah dari tahap training. Pemisahan ini dilakukan agar model dapat menebak di luar dataset yang dilatih. Metrik evaluasi model dinilai menggunakan metrik kuantitatif seperti Recall@K dan MRR. Pengukuran juga dilakukan pada kemampuan model untuk melakukan pencarian gambar berdasarkan text dan text berdasarkan gambar. Hasil evaluasi ketiga model tersebut dapat dijabarkan pada tabel 4.15 berikut ini.

Tabel 4.15 Hasil Evaluasi Pelatihan dengan Dataset Flickr8K Indonesia

Model	Metrik Evaluasi Model			
	Text to Image		Image to Text	
CLIP Standar	Recall@1	54.19%	Recall@1	51.97%
	Recall @5	82.51%	Recall @5	84.24%
	Recall @10	88.92%	Recall @10	92.12%
	MRR	0.6641	MRR	0.6579
CLIP dengan IndoBERT	Recall@1	40.15%	Recall@1	45.57%
	Recall @5	72.91%	Recall@5	79.80%
	Recall @10	84.98%	Recall@10	89.16%
	MRR	0.5515	MRR	0.6100
CLIP dengan XLM-RoBERTa	Recall@1	46.55%	Recall@1	52.96%
	Recall@5	83.50%	Recall@5	86.21%
	Recall@10	92.86%	Recall@10	92.86%
	MRR	0.6164	MRR	0.6654

Tabel 4.15 merangkum hasil evaluasi tugas image-text retrieval dua arah pada dataset Flickr8K Indonesia menggunakan tiga konfigurasi model. Secara umum, arsitektur dasar CLIP menunjukkan performa keseluruhan terbaik dengan Recall@1 (53,08%) dan MRR (0,6610) tertinggi, yang didukung oleh kurva pelatihan stabil dan minim overfitting. Sebagai perbandingan, konfigurasi CLIP dengan XLM-RoBERTa menghasilkan evaluasi yang sangat kompetitif karena model ini mampu mengungguli CLIP pada metrik Recall@5 (84,86%) dan Recall@10 (92,86%). Sebaliknya, performa terendah ditunjukkan oleh integrasi CLIP dengan IndoBERT (Recall@1: 42,86%; MRR: 0,5808). Kinerja suboptimal ini ditandai oleh kurva pelatihan yang mengalami plateau lebih awal, menunjukkan bahwa batas kapasitas representasi model tersebut tercapai lebih cepat untuk skala dataset ini.

Tabel 4.16 Hasil Evaluasi Pelatihan dengan Dataset Flickr30K Indonesia

Model	Metrik Evaluasi Model			
	Text to Image		Image to Text	
CLIP Standar	Recall@1	47.00%	Recall@1	50.60%
	Recall @5	76.70%	Recall @5	76.10%
	Recall @10	84.70%	Recall @10	85.30%
	MRR	0.6019	MRR	0.6209
CLIP dengan IndoBERT	Recall@1	40.00%	Recall@1	43.60%
	Recall @5	70.90%	Recall@5	72.30%
	Recall @10	80.90%	Recall@10	83.10%
	MRR	0.5406	MRR	0.5662
CLIP dengan XLM-RoBERTa	Recall@1	42.40%	Recall@1	49.90%
	Recall@5	73.60%	Recall@5	77.80%
	Recall@10	84.20%	Recall@10	86.70%
	MRR	0.5651	MRR	0.6210

Tabel 4.16 merangkum hasil evaluasi ketiga konfigurasi model pada dataset Flickr30K Indonesia. Berbeda dengan hasil pada dataset Flickr8K, selisih performa antar model pada dataset ini menjadi lebih kecil. Model CLIP tetap mencatatkan hasil tertinggi dengan rata-rata Recall@1 (48,80%) dan MRR (0,6114). Namun, nilainya kini bersaing sangat ketat dengan CLIP berbasis XLM-RoBERTa yang mencetak rata-rata Recall@1 sebesar 46,15%, bahkan sedikit lebih unggul pada metrik Recall@10 (85,45%). Kemampuan XLM-RoBERTa untuk menyusul ini sejalan dengan grafik proses pelatihannya yang paling stabil dibandingkan model lain. Sebaliknya, kombinasi CLIP dengan IndoBERT kembali menempati posisi terendah dengan rata-rata Recall@1 (41,80%) dan MRR (0,5534). Rendahnya hasil ini disebabkan oleh kendala *overfitting* yang cukup parah pada saat pelatihan, sehingga model tersebut tidak mampu belajar dengan baik meskipun jumlah datanya sudah diperbesar.

Analisis Hasil Evaluasi Metrik Ketiga Model

Hasil evaluasi pada kedua dataset secara konsisten menunjukkan bahwa kekuatan *embedding alignment* lebih menentukan performa retrieval dibandingkan spesialisasi linguistik *text encoder*. CLIP standar unggul bukan karena *text encoder* bawaannya lebih memahami Bahasa Indonesia, melainkan karena image encoder dan *text encoder*-nya telah dioptimalkan bersama dalam satu ruang vektor yang sama sejak pretraining pada 400 juta pasangan data, sehingga fine-tuning pada caption Bahasa Indonesia hanya memerlukan penyesuaian marginal. Sebaliknya, CLIP dengan IndoBERT dan CLIP dengan XLM-RoBERTa memulai dengan *text encoder* yang ruang *embedding*nya asing terhadap image encoder, sehingga seluruh beban pelatihan jatuh pada tugas membangun alignment baru melalui projection head dengan data yang jauh lebih terbatas.

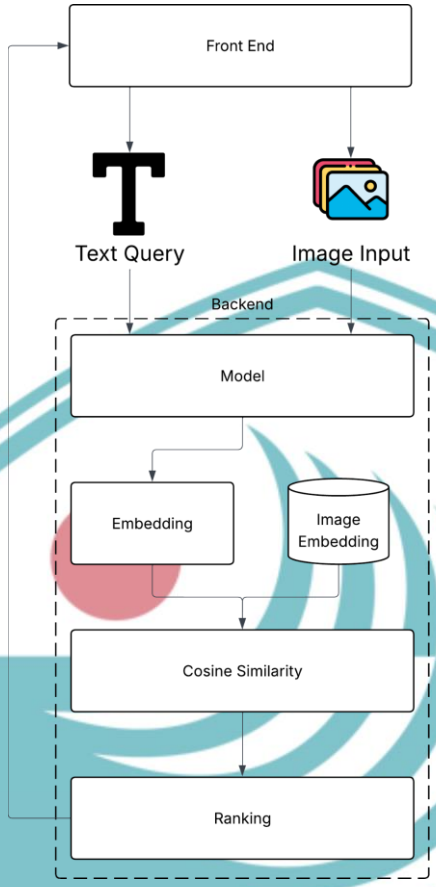
CLIP dengan IndoBERT mencatatkan performa terendah di kedua dataset meskipun merupakan satu-satunya model yang dirancang eksklusif untuk Bahasa Indonesia. Ini disebabkan oleh akumulasi hambatan: representasi IndoBERT yang dihasilkan melalui paradigma MLM secara fundamental berbeda dari cara CLIP merepresentasikan teks, sehingga projection head harus menjembatani jarak representasional yang lebih besar. Meskipun XLM-RoBERTa juga berbasis MLM,

telah terekspos pada data lintas bahasa dalam skala masif yang mencakup teks visual-deskriptif, sehingga representasinya lebih kompatibel dengan ruang *embedding* CLIP dibanding IndoBERT yang dilatih eksklusif pada korpus Bahasa Indonesia. Spesialisasi linguistik IndoBERT, meskipun nyata, tidak cukup untuk mengkompensasi ketidakcocokan ruang *embedding* ini pada skala data yang tersedia.

Temuan paling signifikan adalah tren konvergensi CLIP dengan XLM-RoBERTa terhadap CLIP saat skala data diperbesar. Gap rata-rata MRR keduanya menyusut pada dataset Flickr30K. Penyusutan ini didorong oleh kombinasi dua faktor: CLIP mengalami degradasi lebih besar saat skala data meningkat, sementara CLIP dengan XLM-RoBERTa lebih stabil, mengindikasikan bahwa kemampuan generalisasi lintas bahasa XLM-RoBERTa lebih robust terhadap perubahan distribusi data.

4.3.3 Implementasi Model pada Aplikasi Website

Arsitektur sistem pencarian gambar yang berbasis teks yang dijelaskan pada Gambar 4.25 diimplementasikan pada aplikasi website, di mana sistem terbagi menjadi dua lapisan utama yaitu Front End dan Backend. Pada lapisan Front End, pengguna memberikan input berupa kueri dalam bahasa Indonesia, ataupun *input* gambar sebagai *Image Input*. Kedua input tersebut kemudian diteruskan ke lapisan Backend, dimana model memproses kueri menjadi vektor *embedding*, lalu dibandingkan dengan *Image Embedding* yang telah diolah pada saat sistem dinyalakan dan disimpan di dalam basis data secara lokal. Kemudian kedua *embedding* akan dihitung perbandingannya menggunakan perhitungan *cosine similarity*. Hasil perbandingan tersebut kemudian diurutkan melalui proses ranking dari nilai kemiripan tertinggi ke terendah, dan hasilnya dikembalikan ke lapisan Front End untuk ditampilkan kepada pengguna.

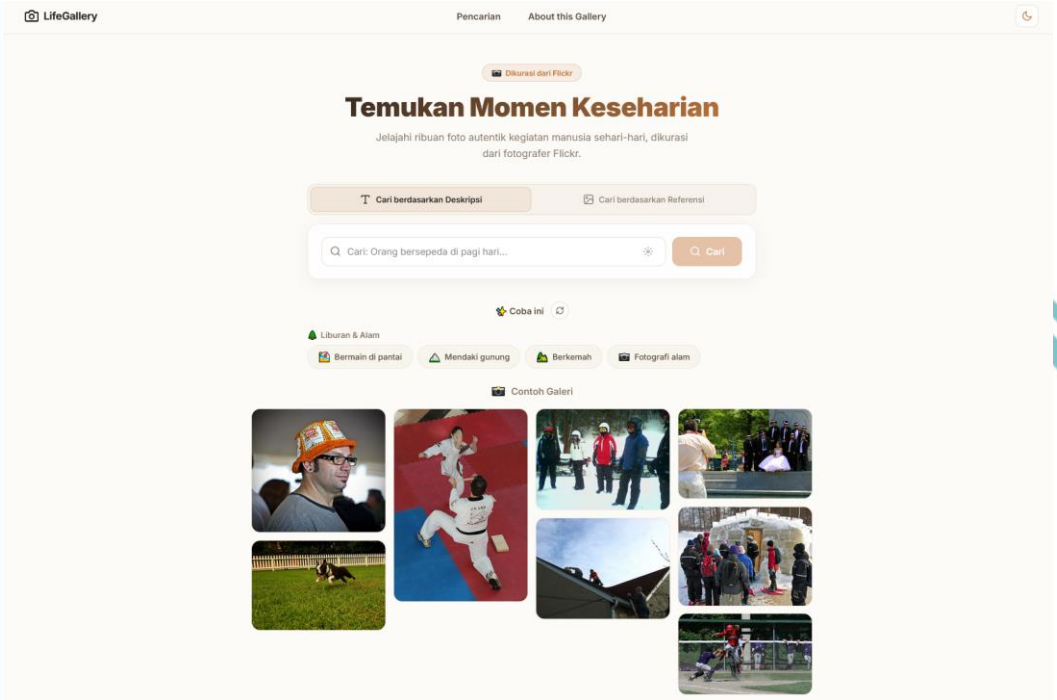


Gambar 4.25 Alur Implementasi Aplikasi Website

Implementasi teknis website terdapat dua komponen yaitu komponen front-end dan back-end. Seluruh implementasi yang teknis yang menjadi fungsionalitas aplikasi website dapat diakses pada platform GitHub dengan alamat https://github.com/Chifaaan/gallery-web

Tampilan Halaman Utama

Halaman utama aplikasi adalah halaman yang pertama kali pengguna lihat ketika mengakses aplikasi website galeri ini.



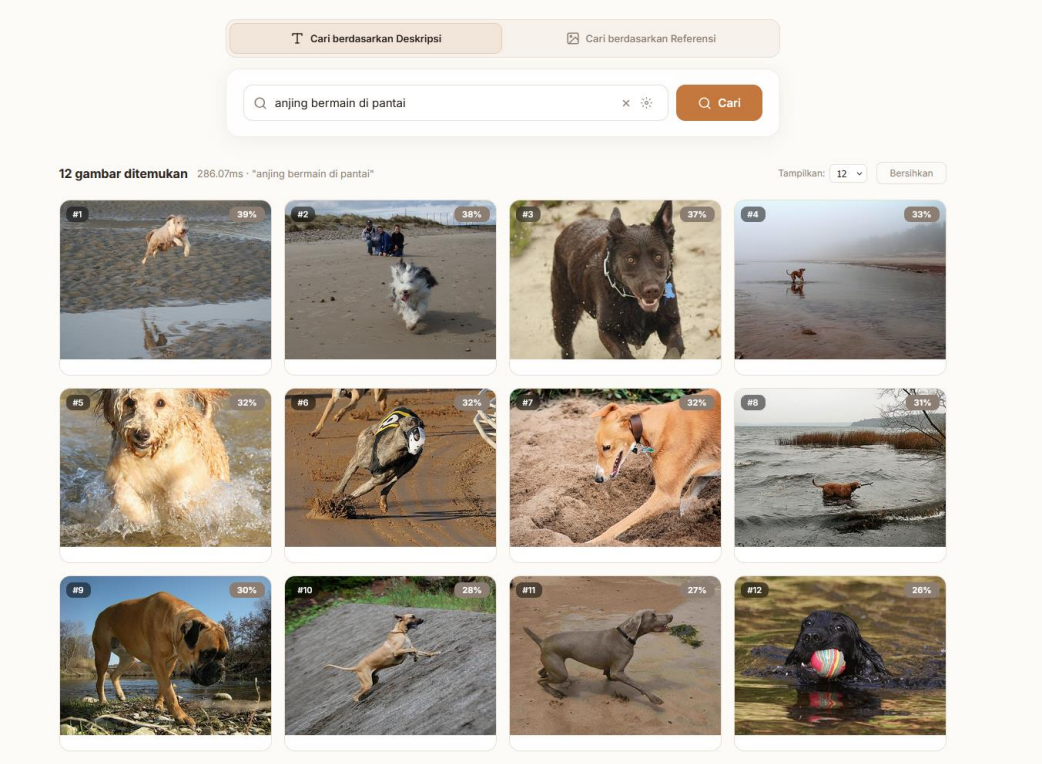
Gambar 4.26 Tampilan Halaman Utama

Gambar 4.26 secara umum memuat menu navigasi untuk berpindah antar halaman yang berada di atas, penjelasan singkat mengenai fungsi utama aplikasi, serta antarmuka berbentuk *tabs* yang memungkinkan pengguna memilih metode pencarian yang diinginkan sesuai dengan modalitas yang tersedia yaitu Teks dan Gambar. Selain itu, halaman ini juga menyediakan pengaturan jumlah hasil pencarian serta contoh kueri siap pakai sebagai panduan bagi pengguna yang baru pertama kali menggunakan sistem, sehingga pengguna dapat langsung memahami cara kerja aplikasi tanpa memerlukan instruksi tambahan.

Tampilan hasil pencarian gambar berdasarkan teks

Halaman hasil pencarian berbasis teks berfungsi menampilkan keluaran sistem setelah pengguna mengeksekusi pencarian menggunakan deskripsi teks. Halaman ini secara umum memuat antarmuka pencarian yang tetap tersedia secara persisten untuk memudahkan modifikasi kueri, informasi ringkas mengenai performa

eksekusi pencarian, serta kumpulan gambar hasil pencarian yang disusun secara terurut berdasarkan tingkat kemiripan terhadap kueri yang dimasukkan.

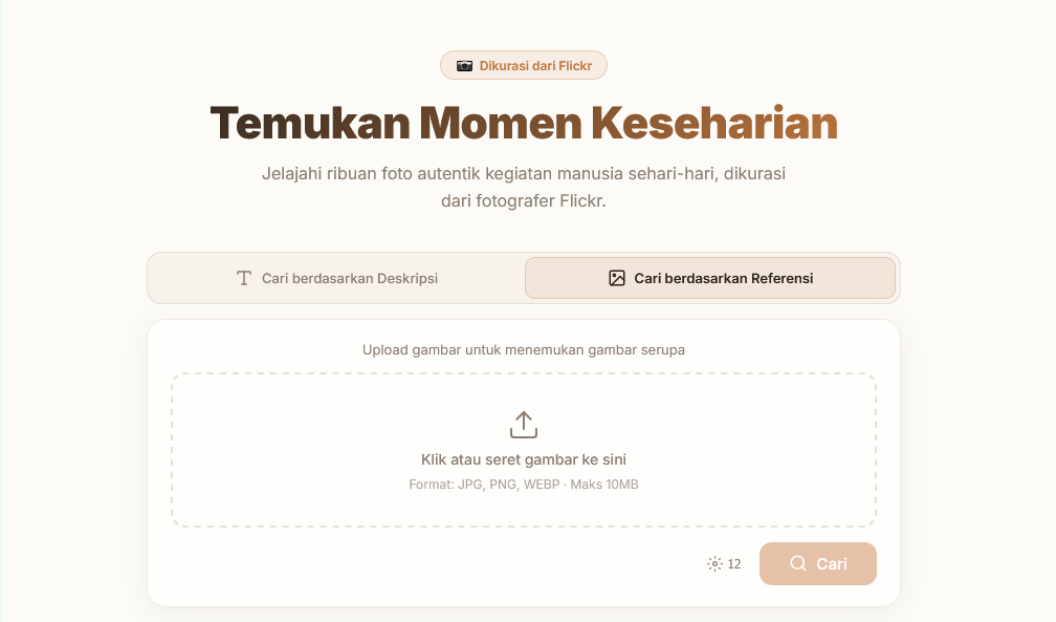


Gambar 4.27 Tampilan Halaman Hasil Pencarian Gambar dengan Teks

Gambar 4.27 menampilkan hasil pencarian pada aplikasi menampilkan keluaran sistem setelah pengguna mengeksekusi pencarian berbasis teks dengan kueri bahasa Indonesia. Di bawah antarmuka pencarian, sistem menampilkan informasi ringkas hasil pencarian berupa jumlah gambar yang ditemukan sebanyak 12 gambar sesuai dengan jumlah gambar yang ditetapkan, waktu eksekusi pencarian, serta kueri yang digunakan, dimana pada sisi kanan terdapat tombol "Bersihkan" untuk mereset hasil pencarian. Gambar-gambar hasil pencarian ditampilkan dalam tata letak kisi (*grid*) empat kolom, di mana setiap gambar dilengkapi dengan nomor urut peringkat pada sudut kiri atas, persentase skor kemiripan (*similarity score*) pada sudut kanan atas, serta teks deskripsi otomatis yang dihasilkan sistem di bagian bawah setiap gambar sebagai keterangan kontekstual, sehingga pengguna dapat mengevaluasi relevansi setiap hasil pencarian secara langsung.

Tampilan Halaman Pencarian Gambar berdasarkan Gambar

Halaman mode pencarian berbasis gambar pada aplikasi Smart Gallery menampilkan antarmuka yang aktif ketika pengguna memilih tab "Cari dengan Gambar" seperti gambar 4.28 berikut ini.

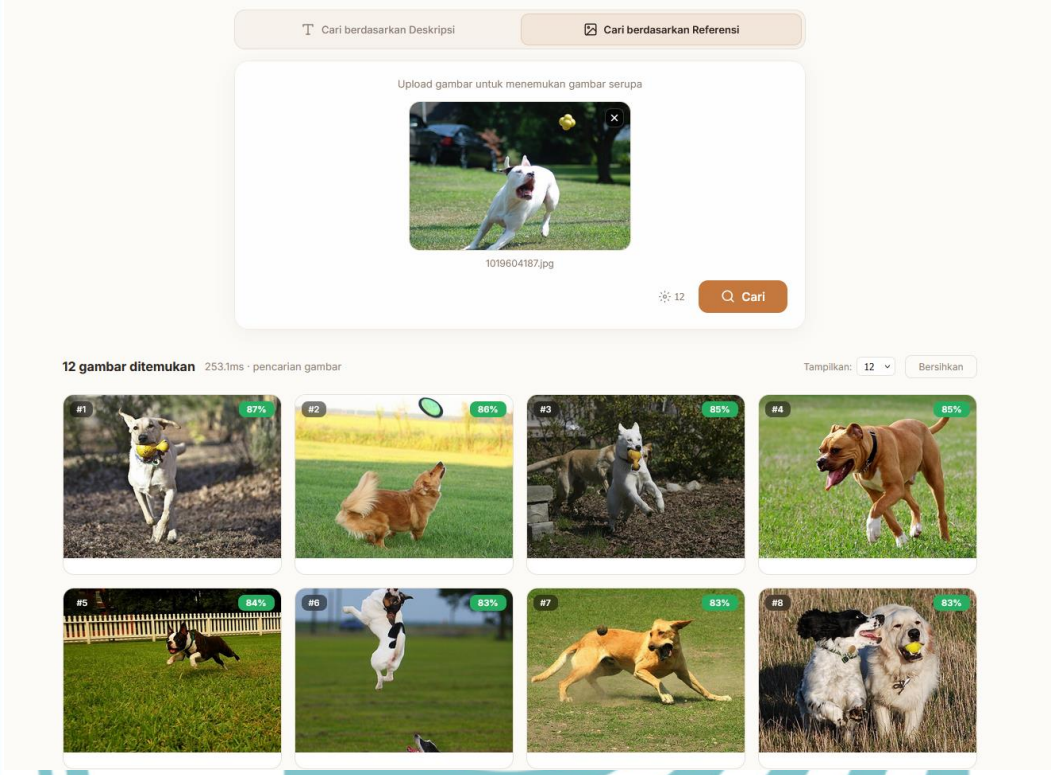


Gambar 4.28 Tampilan Halaman Pencarian dengan Gambar

Pada mode ini, kolom masukan teks digantikan oleh area unggah gambar (*upload area*) dilengkapi ikon unggah dan instruksi "Klik atau seret gambar ke sini" sebagai panduan interaksi bagi pengguna, serta keterangan format berkas yang didukung yaitu JPG, PNG, dan WEBP dengan batas ukuran maksimal 10MB. Serupa dengan mode pencarian teks, antarmuka ini tetap menyediakan menu tarik-turun (*dropdown*) untuk mengatur jumlah hasil pencarian dengan nilai default dua belas, serta tombol "Cari" untuk mengeksekusi pencarian setelah gambar referensi berhasil diunggah.

Tampilan hasil pencarian gambar berdasarkan gambar

Halaman hasil pencarian berbasis gambar pada aplikasi Smart Gallery menampilkan keluaran sistem setelah pengguna mengunggah gambar referensi berupa foto seperti yang ditampilkan pada Gambar 4.29.



Gambar 4.29 Tampilan Halaman Hasil Pencarian Gambar dengan Gambar

Pada bagian atas antarmuka, area unggah gambar kini menampilkan pratinjau (*preview*) gambar referensi yang telah diunggah beserta nama berkasnya, dilengkapi tombol hapus pada sudut kanan atas pratinjau untuk memungkinkan pengguna mengganti gambar referensi, serta menu tarik-turun jumlah hasil dan tombol "Cari" yang tetap tersedia untuk keperluan pengulangan atau modifikasi pencarian. Di bawah antarmuka pencarian, sistem menampilkan informasi ringkas yang mencakup jumlah gambar yang ditemukan sebanyak 12 gambar, waktu eksekusi pencarian, serta keterangan jenis pencarian "pencarian gambar", di mana pada sisi kanan terdapat tombol "Bersihkan" untuk mereset hasil. Gambar-gambar hasil pencarian ditampilkan dalam tata letak kisi empat kolom yang sama dengan pencarian teks, di mana setiap gambar dilengkapi nomor urut peringkat pada sudut kiri atas, persentase skor kemiripan (*similarity score*) pada sudut kanan atas, serta teks deskripsi metadata gambar di bagian bawah setiap gambar.

Tampilan Halaman About Gallery

Halaman "About this Gallery" pada aplikasi Smart Gallery yang terjelaskan pada Gambar 4.30 berfungsi sebagai halaman informasi yang memberikan pemahaman kepada pengguna mengenai fondasi teknis dan konfigurasi sistem yang mendasari cara kerja aplikasi.

Gambar 4.30 Tampilan Halaman About Gallery

Halaman ini secara umum memuat penjelasan tentang dataset gambar yang digunakan sebagai basis data pencarian beserta karakteristik dan cakupan kontennya, keterangan mengenai asal-usul dan relevansi dataset tersebut dalam konteks penelitian kecerdasan buatan, serta informasi mengenai bahasa yang diadopsi sistem dalam memproses dan menampilkan deskripsi gambar kepada pengguna. Kehadiran halaman ini bertujuan untuk membangun transparansi sistem kepada pengguna dengan menjelaskan sumber data, ruang lingkup konten, serta kapabilitas bahasa yang didukung oleh aplikasi, sehingga pengguna dapat memahami konteks dan keterbatasan sistem pencarian gambar yang digunakan.

4.4 Pengujian Sistem

Pengujian sistem ini terdiri dari dua tipe pengujian, yaitu black box testing dan System Usability Scale (SUS). Black box testing yang dilakukan secara independen. Sedangkan SUS dilakukan oleh pengguna akhir.

4.4.1 Deskripsi Pengujian

A. **Pengujian Model berdasarkan Kueri:** dilakukan dengan cara mengamati relevansi hasil *retrieval* yang dikembalikan oleh ketiga konfigurasi model yaitu, CLIP standar, CLIP dengan IndoBERT, dan CLIP dengan XLM-RoBERTa terhadap kueri teks Bahasa Indonesia yang diberikan. Pengamatan difokuskan pada kemampuan masing-masing model dalam memahami semantik kueri dan mengembalikan gambar yang secara visual dan kontekstual sesuai.

B. **Pengujian Black Box**: Pengujian *black box* dilakukan untuk memverifikasi fungsionalitas aplikasi Smart Gallery dari sisi pengguna tanpa mempertimbangkan detail implementasi internal sistem. Pengujian ini berfokus pada validasi fitur-fitur utama aplikasi, meliputi fungsi pencarian gambar berbasis teks (*text query*), pencarian berbasis gambar (*image query*), tampilan hasil pencarian, serta penanganan kondisi masukan yang tidak valid. Setiap fitur diuji berdasarkan kesesuaian antara masukan yang diberikan dengan keluaran yang diharapkan sesuai spesifikasi sistem.

C. **Pengujian UAT**: Pengujian *User Acceptance Testing* (UAT) dilakukan untuk mengevaluasi tingkat penerimaan dan kemudahan penggunaan aplikasi Smart Gallery dari perspektif pengguna umum non-teknis. Pengujian ini melibatkan sejumlah responden yang diminta untuk menggunakan aplikasi secara langsung dan memberikan penilaian terhadap aspek kemudahan penggunaan (*usability*), relevansi hasil pencarian, serta kepuasan pengguna secara keseluruhan. Hasil UAT digunakan sebagai bukti bahwa sistem yang dibangun telah memenuhi kebutuhan dan ekspektasi pengguna akhir yang menjadi target aplikasi.

4.4.2 Prosedur Pengujian

Pengujian Model berdasarkan Kueri

pengujian model berdasarkan kueri dilakukan terhadap model yang diteliti yaitu, CLIP standar, CLIP dengan IndoBERT, dan CLIP dengan XLM-RoBERTa menggunakan 8 kueri berbahasa Indonesia yang terdiri dari 4 kueri bebas dan 4 kueri berbasis caption asli dataset Flickr30K Indonesia dengan 1.000 gambar pada split pengujian. Setiap kueri menghasilkan satu gambar teratas per model yang ditampilkan secara berdampingan untuk memudahkan perbandingan visual. Gambar dianggap relevan apabila objek utama dan konteks yang disebutkan dalam kueri dapat diidentifikasi secara visual pada gambar hasil. Penilaian relevansi dilakukan secara manual oleh peneliti berdasarkan kesesuaian visual antara gambar hasil dan kueri yang diberikan.

Pengujian Black Box

Pengujian *black box* merupakan metode pengujian perangkat lunak yang berfokus pada evaluasi fungsionalitas sistem berdasarkan input dan output tanpa memperhatikan struktur internal atau kode program di dalamnya. Pengujian *Black Box* pada aplikasi website pencarian gambar ini dilakukan secara mandiri dengan membandingkan ekspektasi yang diharapkan dengan hasil aktual pada setiap fitur. Pengujian ini dilakukan pada setiap fungsionalitas aplikasi dan halaman yang ada.

Tabel 4.17 Prosedur Pengujian Black Box

Kode Tes	Nama Test	Langkah Uji	Input	Ekspektasi Output
TC-01	Tampilan Halaman utama	Buka website melalui browser	-	Halaman Home muncul dengan lengkap
TC-02	Navigasi mode pencarian	Klik pencarian berdasarkan teks dan gambar	-	Sistem memberikan input sesuai dengan modalitasnya
TC-03	Pencarian dengan file valid	Pilih file .png, .jpeg dan klik Search	File berformat valid (.png,.jpeg)	Sistem menampilkan hasil pencarian gambar
TC-04	Pencarian dengan file tidak valid	Pilih file .txt atau .pdf lalu klik Search	FIle tidak valid	Sistem menolak input user dan menampilkan format file yang valid
TC-05	Tampilan Hasil Pencarian Gambar dengan teks	Memasukkan kueri teks pada search bar	Kueri teks	Sistem menampilkan hasil pencarian gambar
TC-06	Tampilan Halaman About Gallery	Klik menu “About this Gallery”	-	Halaman Tentang galeri tampil dengan informasi tentang galeri
TC-07	Mengubah Jumlah Gambar yang ditampilkan	Klik dropdown jumlah gambar yang ditampilkan	-	Gambar yang tampil sesuai dengan jumlah gambar yang ditetapkan

Pengujian System Usability Scale (SUS)

Pengujian System Usability Scale (SUS) dilakukan untuk memastikan bahwa sistem yang dikembangkan telah memenuhi kebutuhan dan harapan user. SUS terdiri dari sepuluh pernyataan yang menggunakan skala Likert dengan rentang

nilai 1 hingga 5, di mana nilai 1 menyatakan "Sangat Tidak Setuju" dan nilai 5 menyatakan "Sangat Setuju". SUS terdiri dari 10 pertanyaan yang harus dijawab oleh responden menyangkut *usability* penggunaan aplikasi website seperti Tabel berikut ini

Tabel 4.18 Pertanyaan SUS

Kode	Pertanyaan
SUS1	Saya berpikir akan menggunakan web ini lagi
SUS2	Saya merasa website ini rumit untuk digunakan
SUS3	Saya merasa website ini mudah untuk digunakan
SUS4	Saya membutuhkan orang lain atau teknisi untuk menggunakan website ini
SUS5	Saya merasa fitur-fitur yang tersedia berjalan dengan semestinya
SUS6	Saya merasa ada banyak hal yang tidak konsisten (tidak teratur) pada website ini
SUS7	Saya rasa orang lain dapat memahami cara menggunakan website ini dengan cepat
SUS8	Saya merasa website ini membingungkan
SUS9	Saya merasa tidak memiliki hambatan ketika menggunakan website ini
SUS10	Saya perlu membiasakan diri terlebih dahulu untuk menggunakan website ini

4.4.3 Data Hasil Pengujian

Data Hasil Pengujian Model berdasarkan Kueri

Data hasil pengujian dari tiga variasi model yang dievaluasi, yaitu CLIP, CLIP-IndoBERT, dan CLIP-XLM-RoBERTa dapat dilihat dari hasil pencarian kueri pada Tabel 4.19. Berdasarkan prosedur, pengujian dilakukan menggunakan 8 kueri berbahasa Indonesia yang mencakup 4 kueri bebas dan 4 kueri berbasis *caption*.

Relevansi pada tabel tersebut ditentukan melalui penilaian manual oleh peneliti dengan mencocokkan kesesuaian visual antara gambar hasil dan kueri input.

Tabel 4.19 Hasil Image Retrieval berdasarkan Kueri Teks

Query	Jenis Kueri	CLIP	CLIP + IndoBERT	CLIP + XLM RoBERTa
Seekor anjing berlari di halaman	Bebas			
Anak berpakaian biru sedang bermain	Bebas			
Anak bermain air	Bebas			
kelompok orang berkumpul di luar ruangan	Bebas			
Seseorang berseragam karate menendang papan yang dipegang instrukturnya	Caption			
Seekor anjing berwarna coklat kemerahan berjalan di air sambil membawa tongkat di mulutnya	Caption			

Query	Jenis Kueri	CLIP	CLIP + IndoBERT	CLIP + XLM RoBERTa
Sekelompok anak memegang mangkuk dan sumpit saat mereka berdiri mengelilingi barbekyu	Caption			
Seorang anak berkemeja biru memainkan terompet di marching band.	Caption			

Hasil pengujian yang ditunjukkan pada Tabel 4.19 menunjukkan variasi performa yang berbeda antar model. Pada kueri bebas, ketiga model umumnya mampu menangkap objek utama dengan baik, namun CLIP dengan XLM-RoBERTa konsisten unggul dalam menangkap kombinasi atribut dan aksi secara bersamaan, seperti pada kueri *"anak berpakaian biru sedang bermain"* di mana hanya model ini yang berhasil menampilkan anak bermain dengan atribut warna yang tepat. Pada kueri berbasis caption, CLIP standar dan CLIP dengan XLM-RoBERTa mendominasi dengan menghasilkan gambar sesuai ground truth pada sebagian besar kueri, kecuali pada kueri dengan detail aksi spesifik seperti *"anjing membawa tongkat di air"* di mana CLIP dengan XLM-RoBERTa gagal menangkap aksi dan hanya memahami konteks lokasi. Ketiga model menunjukkan performa terbaik secara konsisten pada kueri yang mengandung konteks sosial yang jelas, seperti *"kelompok orang berkumpul di luar ruangan"* dan *"anak memegang sumpit di barbekyu"*, di mana ketiganya menghasilkan gambar yang sama dan sesuai.

Data Hasil Pengujian Black Box

Data hasil pengujian Black Box dikumpulkan secara independen sesuai yang telah disebutkan sebelumnya. Pengujian ini dilakukan dengan membandingkan ekspektasi dan hasil percobaan yang dilakukan. Jika hasil percobaan sesuai dengan harapan, maka skenario fitur sudah berhasil. Namun, jika hasilnya berbeda dengan

harapan maka skenario fitur belum dinyatakan berhasil sehingga memerlukan perbaikan. Hasil pengujiannya dapat dilihat pada Tabel 4.20 di bawah ini

Tabel 4.20 Hasil Pengujian Black Box

Kode Tes	Nama Test	Langkah Uji	Input	Ekspektasi Output	Hasil
TC-01	Tampilan Halaman utama	Buka website melalui browser	-	Halaman Home muncul dengan lengkap	Valid
TC-02	Navigasi mode pencarian	Klik pencarian berdasarkan teks dan gambar	-	Sistem memberikan input sesuai dengan modalitasnya	Valid
TC-03	Pencarian dengan file valid	Pilih file .png, .jpeg dan klik Search	File berformat valid (.png,.jpeg)	Sistem menampilkan hasil pencarian gambar	Valid
TC-04	Pencarian dengan file tidak valid	Pilih file .txt atau .pdf lalu klik Search	FIle tidak valid	Sistem menolak input user dan menampilkan format file yang valid	Valid
TC-05	Tampilan Hasil Pencarian Gambar dengan teks	Memasukkan kueri teks pada search bar	Kueri teks	Sistem menampilkan hasil pencarian gambar	Valid
TC-06	Tampilan Halaman About Gallery	Klik menu “About this Gallery”	-	Halaman Tentang galeri tampil dengan informasi tentang galeri	Valid
TC-07	Mengubah Jumlah Gambar yang ditampilkan	Klik dropdown jumlah gambar yang ditampilkan	-	Gambar yang tampil sesuai dengan jumlah gambar yang ditetapkan	Valid

Data Hasil Pengujian System Usability Scale (SUS)

Hasil pengujian aplikasi website kepada pengguna dengan metode survey menggunakan System Usability Scale dapat dilihat pada Tabel 4.21 di bawah ini

Tabel 4.21 Hasil Pengujian SUS

NO	Skor Responden									
	SUS 1	SUS 2	SUS 3	SUS 4	SUS 5	SUS 6	SUS 7	SUS 8	SUS 9	SUS 10
1	3	1	5	1	5	1	5	1	5	1
2	5	2	5	1	5	2	4	2	4	1
3	5	1	5	1	5	1	5	1	5	1
4	5	1	5	1	5	2	5	1	5	1
5	5	1	5	1	5	3	5	1	5	1
6	4	1	5	1	5	1	5	1	5	1
7	5	1	5	1	3	3	5	3	4	1
8	4	1	5	1	5	2	5	1	5	1
9	3	1	5	1	5	1	5	1	5	1
10	5	2	4	2	4	4	5	1	4	2
11	3	2	4	1	5	4	5	1	5	1

Hasil kusioner pada tabel di atas harus diolah menggunakan persamaan SUS yang sudah ditetapkan sesuai dengan urutan ganjil dan genap. Nilai SUS tersebut didapatkan dengan menggunakan perhitungan berdasarkan Tabel 4.22 berikut ini:

Tabel 4.22 Rumus Perhitungan SUS

Jenis Pernyataan	Rumus Skor Individual
Ganjil (Positif)	Skor = Nilai Jawaban – 1
Genap (Negatif)	Skor = 5 – Nilai Jawaban
Total Skor	Skor Ganjil + Skor Genap
Skor SUS	Total Skor x 2.5

Setelah mendapatkan Skor SUS, tahap terakhir adalah mengkonversi Persentase Nilai SUS menjadi *Rating* dengan perhitungan Tabel 4.23 di bawah ini

Tabel 4.23 *Rating* nilai SUS

SUS Score	*Grade*	*Rating*
>80,3	A	*Excellent*
68-80,2	B	*Good*
67	C	*Okay*
51-66	D	*Poor*
<51	E	*Awful*

Berdasarkan persamaan di atas maka skor hasil perhitungan SUS masing-masing responden dapat dilihat pada Tabel 4.24 di bawah ini

Tabel 4.24 Skor Perhitungan SUS

NO	**Skor Hasil Hitung SUS**										Jumlah	Nilai (**Jumlah x 2.5**)
	SUS 1	**SUS 2**	**SUS 3**	**SUS 4**	**SUS 5**	**SUS 6**	**SUS 7**	**SUS 8**	**SUS 9**	**SUS 10**		
1	2	4	4	4	4	4	4	4	4	4	38	95
2	4	3	4	4	4	3	3	3	3	4	35	87.5
3	4	4	4	4	4	4	4	4	4	4	40	100
4	4	4	4	4	4	3	4	4	4	4	39	97.5
5	4	4	4	4	4	2	4	4	4	4	38	95
6	3	4	4	4	4	4	4	4	4	4	39	97.5
7	4	4	4	4	2	2	4	2	3	4	33	82.5
8	3	4	4	4	4	3	4	4	4	4	38	95
9	2	4	4	4	4	4	4	4	4	4	38	95
10	4	3	3	3	3	1	4	4	3	3	31	77.5
11	2	3	3	4	4	1	4	4	4	4	33	82.5
Skor rata-rata (Hasil Akhir)												91.36
Keterangan Hasil												*Excellent*

Hasil perhitungan pada Tabel 4.24 menunjukkan nilai rata-rata SUS mencapai skor 91.36, yang secara klasifikasi termasuk dalam rating *Excellent* dengan *grade* "A". Data tersebut menegaskan bahwa performa aplikasi pencarian gambar menggunakan model CLIP telah berfungsi secara optimal dan dinilai sangat layak untuk digunakan pengguna.

4.4.4 Analisis Data

Analisis Hasil Evaluasi Pengujian Model berdasarkan Kueri

Hasil pengujian model berdasarkan kueri secara umum konsisten dengan temuan evaluasi kuantitatif yang telah dilakukan sebelumnya. CLIP standar yang memperoleh nilai MRR tertinggi pada evaluasi kuantitatif terbukti kompetitif pada sebagian besar kueri, namun terdapat kelemahan yang tidak tertangkap oleh metrik MRR, yaitu ketidakmampuan menangkap kombinasi atribut visual dan aksi secara bersamaan pada kueri singkat seperti *"anak berpakaian biru sedang bermain"*, di mana model hanya mengasosiasikan warna tanpa memahami konteks aktivitas. CLIP dengan XLM-RoBERTa yang memperoleh MRR tertinggi kedua pada Flickr30K juga menunjukkan konsistensi serupa, dengan kemampuan menangkap atribut warna dan aksi secara bersamaan yang lebih baik dibandingkan dua model lainnya, meskipun model ini gagal pada kueri dengan detail aksi yang sangat spesifik seperti *"anjing membawa tongkat di air"* di mana hanya konteks lokasi yang berhasil ditangkap. Temuan ini memperkuat hasil evaluasi kuantitatif dan menyatakan bahwa performa model pada kueri pendek maupun kueri dengan detail visual yang kompleks perlu dipertimbangkan secara bersama dalam menentukan model yang paling sesuai untuk diimplementasikan pada sistem pencarian gambar berbahasa Indonesia.

Analisis Hasil Pengujian Black Box

Berdasarkan pengujian black box, seluruh skenario fitur pencarian gambar berhasil dijalankan dengan hasil pengamatan yang sepenuhnya sesuai harapan. Pencapaian tingkat kelayakan 100% ini menunjukkan bahwa aplikasi website sudah sangat siap dan layak untuk digunakan oleh pengguna. Meskipun demikian, nilai sempurna ini

terbatas pada skenario yang telah dirancang sebelumnya di dalam tabel pengujian. Oleh karena itu, langkah antisipasi tetap diperlukan karena pengguna nyata sehari-hari masih berpotensi menghadapi situasi atau skenario tidak terduga di luar perencanaan pengujian.

Analisis Hasil Pengujian SUS

Pengujian usability sistem pencarian gambar menggunakan System Usability Scale (SUS) terhadap 11 responden menghasilkan rata-rata skor sebesar 91,36. Berdasarkan skala rating yang dijelaskan pada tabel 4.22, skor ini berada pada kategori *"Excellent"* dengan peringkat Grade A yang merepresentasikan tingkat usability tertinggi dalam skala normatif SUS. Hal ini mengindikasikan bahwa sistem pencarian gambar yang dikembangkan dapat diterima dan dioperasikan dengan baik oleh pengguna.

Berdasarkan skor SUS yang sudah dihitung, keunggulan utama sistem ini terletak pada kemudahan penggunaan dan aksesnya. Pengguna merasa sistem ini sangat mudah dipelajari dan bisa digunakan sendiri tanpa perlu bantuan orang teknis. Hal ini dibuktikan oleh tingginya nilai rata-rata pada pertanyaan Q7 (cepat dipelajari), Q4 (tidak butuh bantuan orang teknis), dan Q10 (tidak perlu belajar banyak hal sebelum menggunakan). Sebaliknya, item dengan skor konversi terendah terdapat pada Q6 mengenai konsistensi sistem yang menunjukkan bahwa sebagian pengguna merasakan perilaku sistem yang tidak seragam atau inkonsisten. Perbedaan performa model adalah hal yang berpotensi menjadi pemicu pengalaman pengguna yang tidak seragam, sehingga dapat menjelaskan persepsi inkonsistensi yang tertangkap pada Q6. Meskipun temuan ini tidak berdampak signifikan terhadap skor keseluruhan, hasil tersebut memberikan arah yang konkret untuk perbaikan sistem pada pengembangan berikutnya.

BAB 5
PENUTUP

5.1 Kesimpulan

Penelitian ini telah berhasil membandingkan dan mengimplementasikan pasangan model CLIP terbaik pada sistem pencarian gambar dengan teks bahasa Indonesia. Penelitian ini membuktikan adanya perbedaan performa yang signifikan di antara ketiga konfigurasi *text encoder* untuk tugas pencarian gambar dengan kueri berbahasa Indonesia dengan menggunakan dataset Flickr8K dan Flickr30K. Penelitian ini menghasilkan ranking yang konsisten yaitu dengan CLIP standar memberikan performa yang terbaik dibandingkan model CLIP dengan XLM-RoBERTa dan model CLIP dengan IndoBERT. Fitur yang berhasil dikembangkan model CLIP standar mencangkup:

1) Pencarian gambar berbasis Teks: Sistem dapat menemukan gambar yang relevan secara semantik hanya melalui deskripsi teks berbahasa Indonesia sehari-hari.
2) Pencarian gambar berbasis Gambar: Sistem dapat menemukan gambar-gambar serupa di dalam galeri berdasarkan kemiripan fitur visual dari gambar referensi yang diunggah.

Hasil eksperimen secara keseluruhan mengindikasikan bahwa keberhasilan suatu model dalam *contrastive cross-modal retrieval* ditentukan oleh tiga faktor utama:

1) Keselarasan *embedding space*: model yang ruang representasinya sudah terselaraskan secara lintas modalitas sejak tahap *pre-training* terbukti lebih unggul dibandingkan model yang menggabungkan dua *embedding space* berbeda melalui *projection layer*.
2) Kapasitas transfer lintas bahasa: model dengan kemampuan generalisasi lintas bahasa yang lebih kuat menunjukkan adaptasi yang lebih efektif terhadap kueri Bahasa Indonesia dibandingkan model monolingual. Hal ini mengindikasikan bahwa representasi multilingual lebih kompatibel dengan *visual embedding space* yang bersifat *language-agnostic*.

3) Skala dan kesesuaian data pelatihan: berkat gap performa antara CLIP standar dan model berbasis BERT yang menyempit secara konsisten dari Flickr8K ke Flickr30K mengindikasikan bahwa ketersediaan data yang lebih besar berpotensi menggeser keunggulan dari faktor arsitektur menuju faktor adaptasi linguistik.

Ketiga faktor ini secara kolektif menjelaskan mengapa keselarasan *embedding space* lebih dominan dibandingkan kompetensi linguistik spesifik bahasa dalam kondisi data yang terbatas.

Penelitian ini juga berhasil mengimplementasikan model *text encoder* terbaik, yaitu CLIP standar, ke dalam sistem Smart Gallery untuk mendukung pencarian gambar multimodal berbahasa Indonesia. Implementasi ini dikategorikan sebagai berhasil karena dapat diterima dan dioperasikan dengan baik oleh pengguna serta tergolong kategori *Excellent* sesuai dengan hasil pengujian SUS terhadap 11 responden.

5.2 Saran

Berdasarkan hasil penelitian ini, terdapat peluang pengenmbangan lebih lanjut yang dapat dilakukan di masa depan sebagai berikut:

1) Mengeksplorasi pendekatan *Knowledge Distillation* sebagai alternatif *contrastive learning* untuk menjembatani kesenjangan *embedding space* lintas arsitektur, serta melakukan perbandingan langsung dengan pendekatan M-CLIP dan CLIP-Adapter untuk memperoleh gambaran komparatif yang lebih komprehensif.
2) Pengembangan selanjutnya disarankan untuk melakukan eksplorasi *hyperparameter tuning* yang lebih sistematis, seperti variasi *learning rate*, *batch size*, dan *temperature* pada InfoNCE *loss*, guna mengidentifikasi konfigurasi optimal yang berpotensi meningkatkan performa masing-masing model secara lebih menyeluruh.
3) Mengembangkan atau memanfaatkan dataset yang lebih merepresentasikan konten visual spesifik Indonesia, guna meningkatkan relevansi sistem dalam skenario penggunaan yang lebih kontekstual.

DAFTAR PUSTAKA

dos Santos, G. O., Moreira, D. A. B., Ferreira, A. I., Silva, J., Pereira, L., Bueno, P., Sousa, T., Maia, H., Da Silva, N., Colombini, E., Pedrini, H., & Avila, S. (2023). CAPIVARA: Cost-efficient approach for improving multilingual CLIP performance on low-resource languages. *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, 184–207. https://doi.org/10.18653/v1/2023.mrl-1.15

Chen, G., Hou, L., Chen, Y., Dai, W., Shang, L., Jiang, X., Liu, Q., Pan, J., & Wang, W. (2023). mCLIP: Multilingual CLIP via cross-lingual transfer. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13028–13043. https://doi.org/10.18653/v1/2023.acl-long.728

Goring, S. (2025). CLIPSE – A Minimalistic CLIP-Based Image Search Engine for Research. 1–5. https://doi.org/https://doi.org/10.48550/arXiv.2504.17643

Aske, K. (2023). (Mis) Matching Metadata : Improving Accessibility in Digital Visual Archives through the EyCon Project (Mis) Matching Metadata : Improving Accessibility in Digital Visual Archives through the EyCon Project. 16(4). https://doi.org/10.1145/3594726

Mohammadi, M. (2023). Image-Text Connection : Exploring the Expansion of the Diversity within Joint Feature Space Similarity Scores. IEEE Access, PP, 1. https://doi.org/10.1109/ACCESS.2023.3327339

Gastaldi, J. L., Terilla, J., Malagutti, L., DuSell, B., Vieira, T., & Cotterell, R. (2025). *The foundations of tokenization: Statistical and computational concerns*. Proceedings of the International Conference on Learning Representations (ICLR 2025). https://arxiv.org/abs/2407.11606

Yi, W., Cai, X., Ma, H., Fu, Z., & Zhan, Y. (2025). *A library-oriented large language model approach to cross-lingual and cross-modal document retrieval*. *Electronics, 14*(15), 3145. https://doi.org/10.3390/electronics14153145

Gomes, G., Zerva, C., & Martins, B. (2024). Evaluation of Multilingual Image Captioning: How far can we get with CLIP models ? https://doi.org/https://doi.org/10.48550/arXiv.2502.06600

Han, Z., & Azman, A. B. I. N. (2024). Cross-Modal Retrieval : A Review of Methodologies , Datasets , and Future Perspectives. 12(August). https://doi.org/https://doi.org/10.1109/ACCESS.2024.3444817

Faisal, M. R., Nugrahadi, D. T., & Budiman, I. (2024). 3D word *embedding* vector feature extraction and hybrid CNN- LSTM for natural disaster reports identification. 22(5), 1196–1208. https://doi.org/10.12928/TELKOMNIKA.v22i5.26091

Wang, T., Li, F., Zhu, L., Li, J., Zhang, Z., & Shen, H. T. (2024). Cross-Modal Retrieval : A Systematic Review of Methods and Future Directions. 1–35. https://doi.org/https://doi.org/10.1109/JPROC.2024.3525147

Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2022). *Efficient transformers: A survey*. *ACM Computing Surveys, 55*(6), Article 109, 1–38. https://doi.org/10.1145/3530811

Radford, A., Wook, J., Chris, K., Aditya, H., Gabriel, R., Sandhini, G., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. https://doi.org/10.48550/arXiv.2103.00020

Hwang, T., Seo, H., Jung, J., & Jung, S. (2025). *Exploring selective layer freezing strategies in transformer fine-tuning: NLI classifiers with sub-3B parameter models*. *Applied Sciences, 15*(19), 10434. https://doi.org/10.3390/app151910434

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). an Image Is Worth 16X16 Words: Transformers for Image Recognition At Scale. ICLR 2021 - 9th International Conference on Learning Representations. https://doi.org/https://doi.org/10.48550/arXiv.2010.11929

Huang, Z., Dong, Q., Gong, S., & Zhu, X. (2023). *Model-aware contrastive learning: Towards escaping the dilemmas of InfoNCE*. Proceedings of the 40th International Conference on Machine Learning (ICML). https://arxiv.org/abs/2306.13298

Abasova, J., Tanuska, P., & Rydzi, S. (2021). Big data—knowledge discovery in production industry data storages—implementation of best practices. Applied Sciences (Switzerland), 11(16). https://doi.org/10.3390/app11167648

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). *Unsupervised cross-lingual representation learning at scale*. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

Mohammadi, M., Eftekhari, M., & Hassani, A. (2023). Image-text connection: Exploring the expansion of the diversity within joint feature space similarity scores. *IEEE Access*. https://doi.org/10.1109/ACCESS.2023.3327339

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). *IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP*. arXiv. https://arxiv.org/abs/2011.00677

Liang, W., Zhang, Y., Kwon, Y., Yeung, S., & Zou, J. (2022). Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, *35*, 17612–17625. https://doi.org/10.48550/arXiv.2203.02053

Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith, S., Roelofs, R., Gontijo-Lopes, R., Hajishirzi, H., Farhadi, A., Namkoong, H., & Schmidt, L. (2022). *Robust fine-tuning of zero-shot models*. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7959–7971. DOI: https://doi.org/10.1109/CVPR52688.2022.00780

Pambudi, A., & Abidin, Z. (2023). PENERAPAN CRISP - DM MENGGUNAKAN MLR K - FOLD PADA DATA SAHAM PT . TELKOM INDONESIA (PERSERO) TBK (TLKM) (STUDI KASUS : BURSA EFEK. 4(1), 1–14. https://doi.org/https://doi.org/10.33365/jdmsi.v4i1.2462

Plotnikova, V., Dumas, M., & Milani, F. P. (2022). Applying the CRISP-DM data mining process in the financial services industry: Elicitation of adaptation requirements. Data and Knowledge Engineering, 139(October 2021), 1–17. https://doi.org/10.1016/j.datak.2022.102013

Martınez-Plumed, F., Contreras-ochando, L., Kull, M., & Lachiche, N. (2021). CRISP-DM Twenty Years Later : From Data Mining Processes to Data Science Trajectories. 4347(c). https://doi.org/10.1109/TKDE.2019.2962680

Kong, J. T. H., Juwono, F. H., Ngu, I. Y., Nugraha, I. G. D., Maraden, Y., & Wong, W. K. (2023). A Mixed Malay – English Language COVID-19 Twitter Dataset : A Sentiment Analysis. https://doi.org/https://doi.org/10.3390/bdcc7020061

Maina, N. K., Muketha, G. M., & Wambugu, G. M. (2023). *A new complexity metric for UML sequence diagrams*. *International Journal of Software Engineering & Applications, 14*(1), 9–22. https://doi.org/10.5121/ijsea.2023.14102

Lapo, J. R., & Cumbicus-Pineda, O. M. (2024). Detection of recyclable solid waste using convolutional neural networks and PyTorch. IEEE Latin America Transactions , 22(6), 475–483. https://doi.org/10.1109/TLA.2024.10534304

Ramzi, E., & Audebert, N. (2023). Optimization of Rank Losses for Image Retrieval. 1–20. https://doi.org/https://doi.org/10.48550/arXiv.2309.08250

Yas, Q. M., Alazzawi, A., Yas, Q. M., & Rahmatullah, B. (2023). A Comprehensive Review of Software Development Life Cycle methodologies : Pros , Cons , and Future Directions A Comprehensive Review of Software Development Life Cycle methodologies : Pros , Cons , and Future Directions. 4(4). https://doi.org/https://doi.org/10.52866/ijcsm.2023.04.04.014

Ghumatkar, R. S., & Date, A. (2023). Software Development Life Cycle (SDLC). November. https://doi.org/https://doi.org/10.22214/ijraset.2023.56554

Huang, X., Fayoumi, A., Winter, E., & Najdawi, A. (2025). Design Requirements of Breast Cancer Symptom- Management Apps. 1–24. https://doi.org/https://doi.org/10.3390/informatics12030072

Zhou, D.-W., Cai, Z.-W., Ye, H.-J., Zhan, D.-C., & Liu, Z. (2024). *Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need*. *International Journal of Computer Vision*. https://doi.org/10.1007/s11263-024-02220-5

Nguyen, T.-T., Huynh, V.-T., Nguyen, Q.-T., Nguyen, H.-P., Bao, L. L., Minh, T. H., Anh, M. N., Tien, T. N., Thuan, P. N., Phong, H. N., Thai, B. H., Nguyen, V.-T., Nguyen, D.-V., Pham, P.-H., Le-Hoang, M.-H., Le, N.-K., Nguyen, M.-C., Ho, M.-Q., Tran, N.-L., ... Tran, M.-T. (2025). *SHREC 2025: Retrieval of optimal objects for multi-modal enhanced language and spatial assistance (ROOMELSA). Computers & Graphics, 132*, 104400. https://doi.org/10.1016/j.cag.2025.104400

Fumero, M., Pegoraro, M., Maiorca, V., Locatello, F., & Rodolà, E. (2024). *Latent functional maps: A spectral framework for representation alignment*. In *Advances in Neural Information Processing Systems* (Vol. 37). https://doi.org/10.52202/079017-2115

Rusialdi, F. M. (n.d.). Rancang Bangun Aplikasi Jasa Titip Lokal Berbasis Web Dengan Integrasi Pembayaran Midtrans. Multinetics.

Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., & Beyer, L. (2022). *LiT: Zero-shot transfer with locked-image text tuning*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 18123–18133). https://doi.org/10.1109/CVPR52688.2022.01762

Unas, B. M., & Jurgelaitis, M. (2024). Transforming Sketches of UML Use Case Diagrams to Models. 12(October). https://doi.org/10.1109/ACCESS.2024.3514455

Bianchi, F., Attanasio, G., Pisoni, R., Terragni, S., Sarti, G., & Lakshmi, S. (2021). *Contrastive language-image pre-training for the Italian language*. arXiv. https://arxiv.org/abs/2108.08688

Alyami, A., Pileggi, S. F., Sohaib, O., & Hawryszkiewycz, I. (2023). Seamless transformation from use case to sequence diagrams. 1–26. https://doi.org/10.7717/peerj-cs.1444

Gong, Y., Cosma, G., & Finke, A. (2023). Neural-Based Cross-Modal Search And Retrieval Of Artwork. https://doi.org/https://doi.org/10.1109/SSCI52147.2023.10371948

DAFTAR RIWAYAT HIDUP

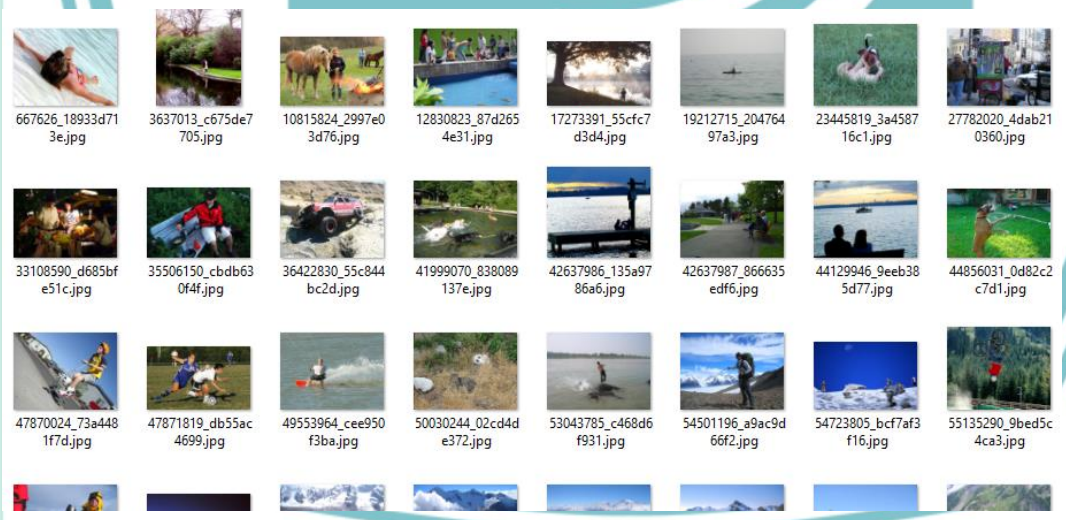
Muhammad Nur Irfan

Lahir di Jakarta pada tanggal 26 Juni 2004, sebagai anak pertama dari dua bersaudara. Lulus dari SD Angkasa 7 pada tahun 2016, kemudian melanjutkan pendidikan di SMP Negeri 80 Jakarta dan lulus pada tahun 2019, serta menempuh pendidikan di SMA Negeri 67 Jakarta pada tahun 2022. Saat ini menempuh Pendidikan Sarjana Terapan pada Program Studi Teknik Informatika di Politeknik Negeri Jakarta pada tahun 2022.

LAMPIRAN

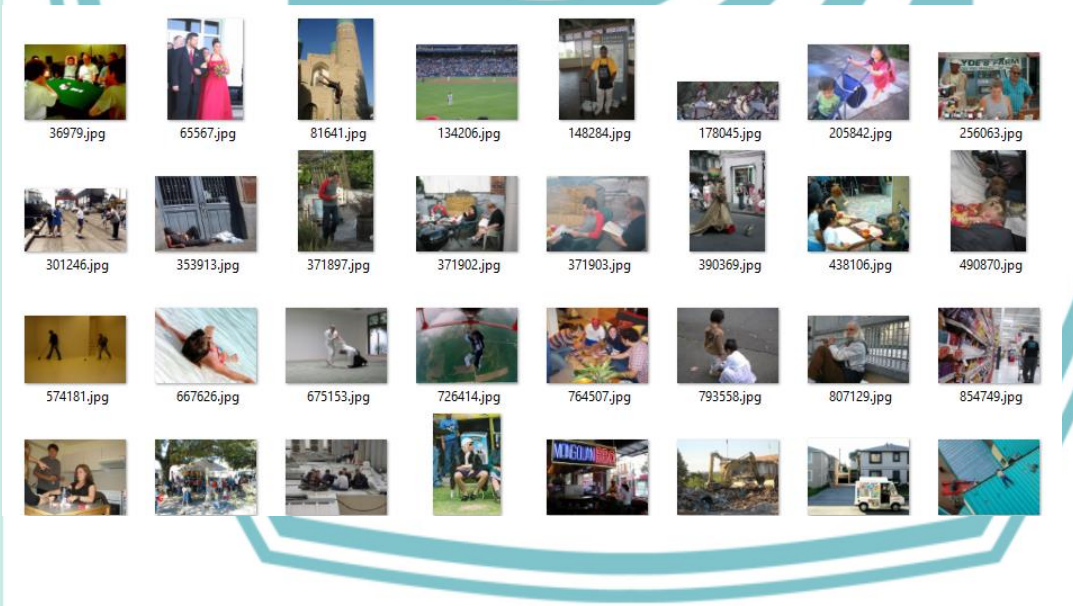
Lampiran 1 Contoh Dataset Flickr8K Indonesia

file_name,text
1000268201_693b08cb0e.jpg,Seorang anak berpakaian merah muda sedang menaiki tangga di jalan masuk.
1000268201_693b08cb0e.jpg,Seorang gadis memasuki sebuah bangunan kayu.
1000268201_693b08cb0e.jpg,Seorang gadis kecil naik ke rumah bermain kayu.
1000268201_693b08cb0e.jpg,Seorang gadis kecil menaiki tangga menuju rumah bermainnya.
1000268201_693b08cb0e.jpg,Seorang gadis kecil berpakaian merah muda masuk ke kabin kayu.
1001773457_577c3a7d70.jpg,Seekor anjing hitam dan anjing tutul sedang berkelahi
1001773457_577c3a7d70.jpg,Seekor anjing hitam dan anjing tiga warna bermain satu sama lain di jalan.
1001773457_577c3a7d70.jpg,Seekor anjing hitam dan seekor anjing putih dengan bintik-bintik coklat saling menatap di jalan.
1001773457_577c3a7d70.jpg,Dua anjing dari ras berbeda saling memandang di jalan.
1001773457_577c3a7d70.jpg,Dua anjing di trotoar bergerak menuju satu sama lain.
1002674143_1b742ab4b8.jpg,Seorang gadis kecil berlumuran cat duduk di depan lukisan pelangi dengan tangannya di dalam mangkuk.
1002674143_1b742ab4b8.jpg,Seorang gadis kecil sedang duduk di depan lukisan pelangi besar.
1002674143_1b742ab4b8.jpg,Seorang gadis kecil di rerumputan bermain dengan cat jari di depan kanvas putih dengan pelangi di atasnya.
1002674143_1b742ab4b8.jpg,Ada seorang gadis berkuncir duduk di depan lukisan pelangi.
1002674143_1b742ab4b8.jpg,Gadis muda dengan kuncir melukis di luar di rumput.
1003163366_44323f5815.jpg,Seorang pria berbaring di bangku sementara anjingnya duduk di sampingnya.
1003163366_44323f5815.jpg,Seorang lelaki berbaring di bangku tempat seekor anjing putih juga diikat.
1003163366_44323f5815.jpg,seorang pria tidur di bangku di luar dengan seekor anjing putih dan hitam duduk di sebelahnya.
1003163366_44323f5815.jpg,Seorang pria bertelanjang dada berbaring di bangku taman bersama anjingnya.
1003163366_44323f5815.jpg,pria berbaring di bangku sambil memegang tali anjing yang duduk di tanah
1007129816_e794419615.jpg,Seorang pria bertopi oranye menatap sesuatu.
1007129816_e794419615.jpg,Seorang pria memakai topi oranye dan kacamata.
1007129816_e794419615.jpg,Seorang pria dengan alat pengukur dan kacamata mengenakan topi Blitz.
1007129816_e794419615.jpg,Seorang pria berkacamata mengenakan topi rajutan kaleng bir.
1007129816_e794419615.jpg,Pria bertindik telinga itu berkacamata dan bertopi oranye.
1007320043_627395c3d8.jpg,Seorang anak sedang bermain di jaring tali.

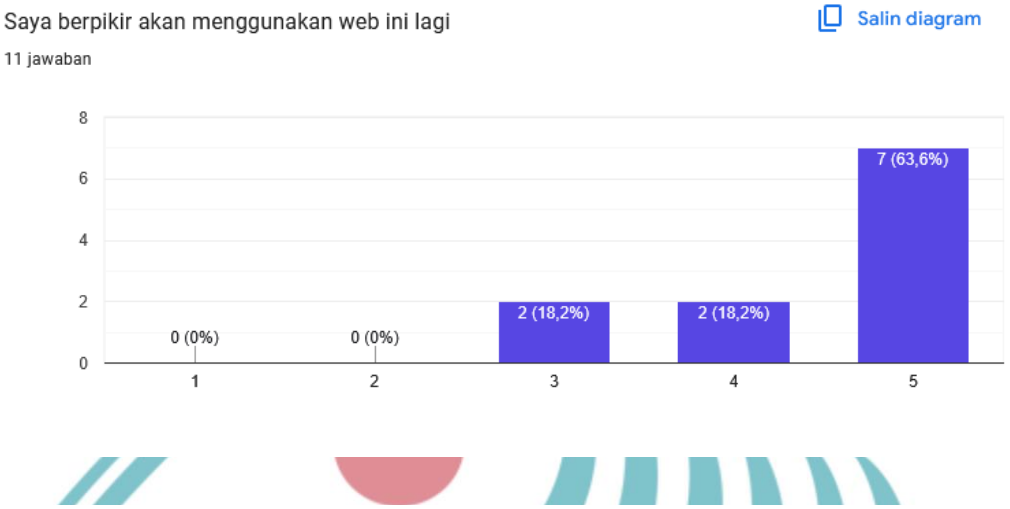


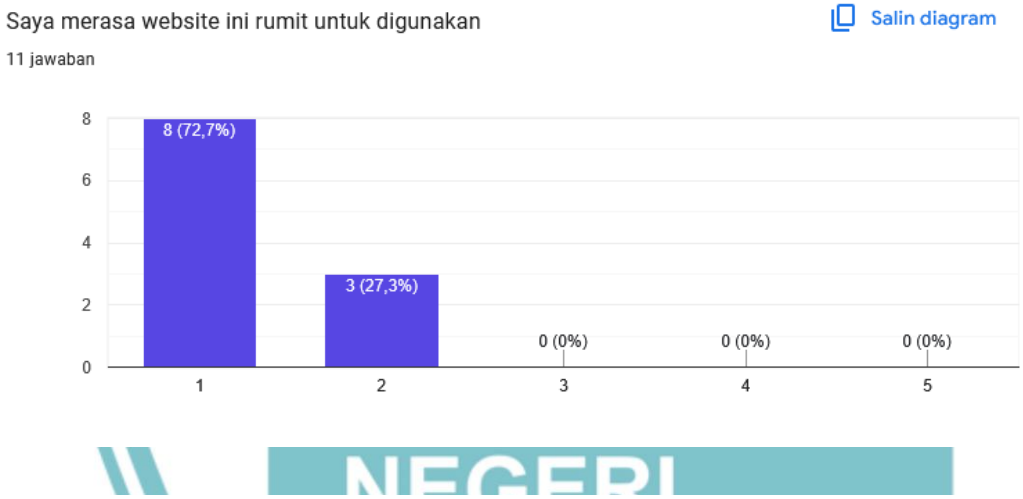
Lampiran 2 Contoh Dataset Flickr30K Indonesia

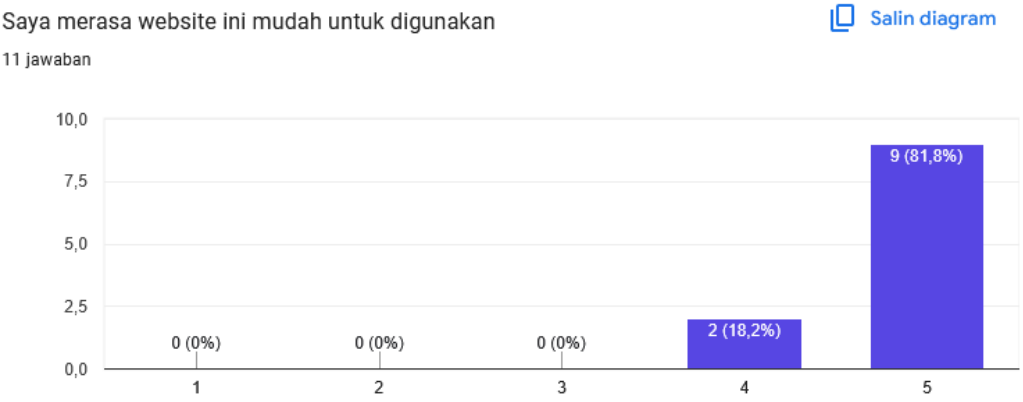
file_name,text
2199250692.jpg,Seorang pria menarik gerobak di atas roda dimuat dengan jaring dan berada di depan Ocean Blue Company.
2199250692.jpg,Seseorang berjalan dengan alat artistik di depan Ocean Blue Company.
2199250692.jpg,Pria bertopi dan sandal menarik gerobak di belakangnya.
2199250692.jpg,Seorang pria menarik sekitar beberapa hal aneh pada roda.
2199250692.jpg,Orang menarik gerobak di jalan.
2199710546.jpg,Seorang DJ berdiri di bawah lampu terang dengan iklan.
2199710546.jpg,Pria berpakaian kuning terlihat DJ Smirnoff di pesta.
2199710546.jpg,Seorang pria bermain untuk kerumunan dengan sorotan pada dirinya.
2199710546.jpg,Ini adalah seorang pria yang DJ-ing di stan DJ.
2199710546.jpg,Seorang DJ di panggung melakukan.
2199721922.jpg,Seorang wanita dengan celana hijau di pantai sambil bertepuk tangan sambil mengamati sebuah band.
2199721922.jpg,Seorang wanita mengenakan celana pendek hijau dan putih dan atas adalah di pantai bertepuk tangan.
2199721922.jpg,Seorang wanita muda dengan celana pendek hijau dan bikini menari di atas pasir.
2199721922.jpg,Seorang gadis dengan pakaian renang menonton acara DJ di pantai.
2199721922.jpg,Wanita ini sedang bersenang-senang di disko techno.
2199793371.jpg,Seekor anjing coklat dan putih bermain dengan bola sepak bola di tanah bersalju.
2199793371.jpg,pitbull coklat ini membawa bola sepak bola di mulutnya.
2199793371.jpg,pitbull kembali mengunyah bola sepak bola di salju.
2199793371.jpg,Besar coklat anjing bermain dengan bola sepak bola di salju
2199793371.jpg,Seekor anjing bermain sepak bola di salju.
2199932695.jpg,Seorang anak kecil yang mengenakan warna bendera Amerika menyekop salju dari tangga depan.
2199932695.jpg,Seorang anak dengan jaket Amerika Serikat menyekop salju dari tangga di depan rumah.
2199932695.jpg,Seorang anak mengenakan jaket yang mengatakan Amerika Serikat menyekop salju.
2199932695.jpg,Seorang anak muda mengenakan jaket USA sekop salju.
2199932695.jpg,Seorang anak kecil bermain dengan salju di tangga.
2200070339.jpg,Seorang anak memakai piyama bergaris merah dan putih berdiri dan memegang tangannya.



Lampiran 3 Pertanyaan Kuesioner SUS





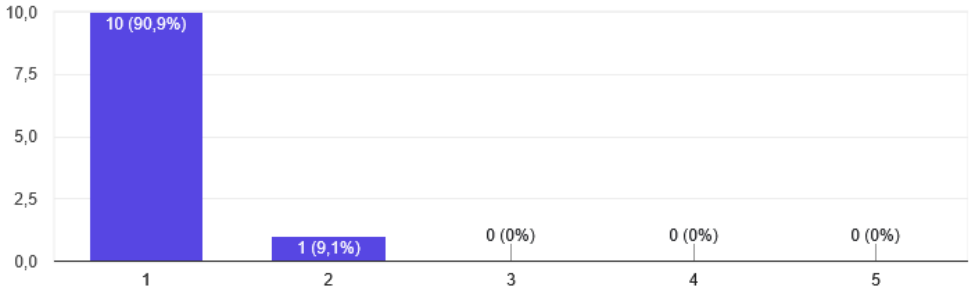


(Lanjutan)

Saya membutuhkan orang lain atau teknisi untuk menggunakan website ini

Salin diagram

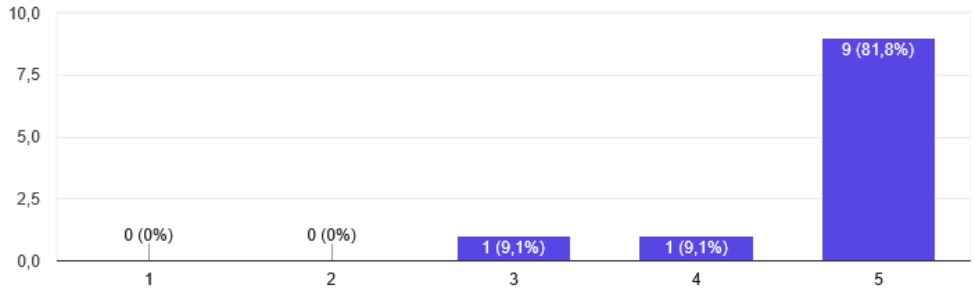
11 jawaban



Saya merasa fitur-fitur yang tersedia berjalan dengan semestinya

Salin diagram

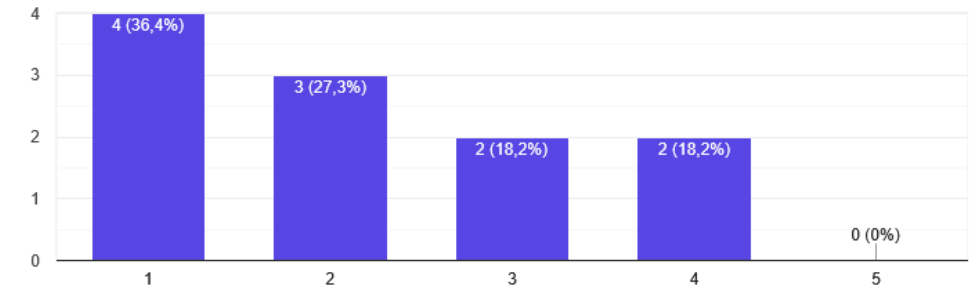
11 jawaban



Saya merasa ada banyak hal yang tidak konsisten (tidak teratur) pada website ini

Salin diagram

11 jawaban

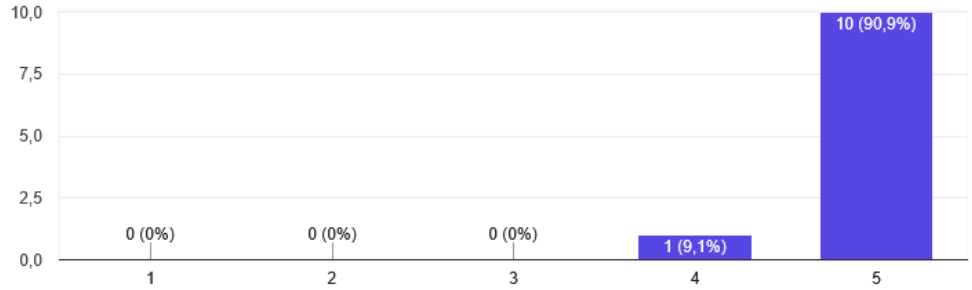


(Lanjutan)

Saya rasa orang lain dapat memahami cara menggunakan website ini dengan cepat

Salin diagram

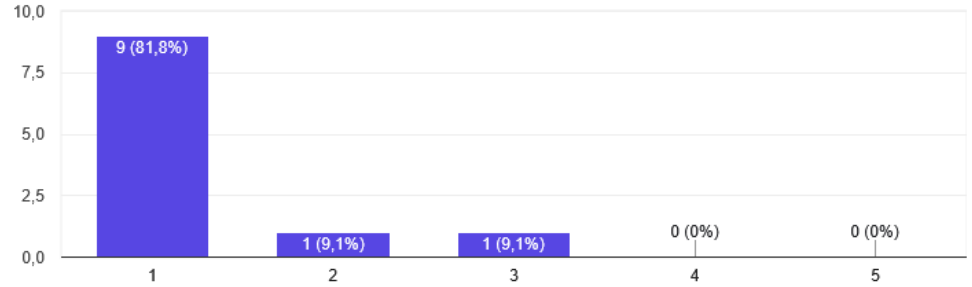
11 jawaban



Saya merasa website pencarian gambar ini membingungkan

Salin diagram

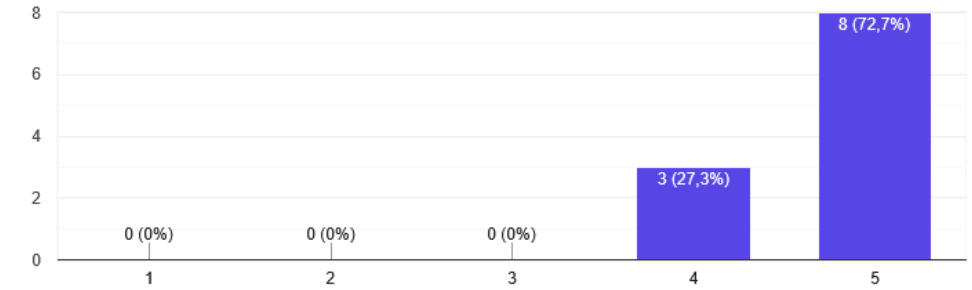
11 jawaban



Saya merasa tidak memiliki hambatan ketika menggunakan website ini

Salin diagram

11 jawaban



(Lanjutan)