

**OPTIMASI DESKRIPSI GAMBAR BERBASIS VISION
TRANSFORMER DAN GENERATIVE PRETRAINED
TRANSFORMER-2 UNTUK PRODUK PAKAIAN**

SKRIPSI SARJANA INFORMATIKA

Oleh

Yogie Fadila Rahman
227064516090



**PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNOLOGI KOMUNIKASI DAN
INFORMATIKA
UNIVERSITAS NASIONAL
2026**

OPTIMASI DESKRIPSI GAMBAR BERBASIS VISION TRANSFORMER DAN GENERATIVE PRETRAINED TRANSFORMER-2 UNTUK PRODUK PAKAIAN

SKRIPSI SARJANA INFORMATIKA

Karya ilmiah sebagai salah satu syarat untuk memperoleh gelar
Sarjana Komputer dari Fakultas Teknologi Komunikasi dan Informatika

Oleh

Yogie Fadila Rahman
227064516090



**PROGRAM STUDI INFORMATIKA
FAKULTAS TEKNOLOGI KOMUNIKASI DAN
INFORMATIKA
UNIVERSITAS NASIONAL
2026**

HALAMAN PENGESAHAN

TUGAS AKHIR

OPTIMASI DESKRIPSI GAMBAR BERBASIS VISION TRANSFORMER
DAN GENERATIVE PRETRAINED TRANSFORMER-2 UNTUK PRODUK
PAKAIAN



(Dr. Fauziah, S.Kom., MMSI.)

PERNYATAAN KEASLIAN TUGAS AKHIR

Saya menyatakan dengan sesungguhnya bahwa Tugas Akhir dengan judul :

OPTIMASI DESKRIPSI GAMBAR BERBASIS VISION TRANSFORMER DAN GENERATIVE PRETRAINED TRANSFORMER-2 UNTUK PRODUK PAKAIAN

Yang dibuat untuk melengkapi salah satu persyaratan menjadi Sarjana Komputer pada Program Studi Informatika Fakultas Teknologi Komunikasi dan Informatika Universitas Nasional, sebagaimana yang saya ketahui adalah bukan merupakan tiruan atau publikasi dari Tugas Akhir yang pernah diajukan atau dipakai untuk mendapatkan gelar di lingkungan Universitas Nasional maupun perguruan tinggi atau instansi lainnya, kecuali pada bagian – bagian tertentu yang menjadi sumber informasi atau acuan yang dicantumkan sebagaimana mestinya.

Jakarta, 25 Februari 2026


METERAI
TEMPEL
PANGKAT 215237785

Yogie Fadila Rahman

227064516090

LEMBAR PERSETUJUAN REVIEW AKHIR

Tugas Akhir dengan judul :

**OPTIMASI DESKRIPSI GAMBAR BERBASIS
VISION TRANSFORMER DAN GENERATIVE
PRETRAINED TRANSFORMER-2 UNTUK
PRODUK PAKAIAN**

Dibuat untuk melengkapi salah satu persyaratan menjadi Sarjana Komputer pada Program Studi Informatika, Fakultas Teknologi Komunikasi dan Informatika Universitas Nasional. Tugas Akhir ini diujikan pada Sidang Review Akhir Semester Ganjil 2025-2026 pada tanggal 25 Februari Tahun 2026

Dosen Pembimbing 1

Dr. Fauziah, S.Kom., MMSI.

NIDN 0104090784

Ketua Program Studi

Ratih Titi Komala Sari, ST., MM., MMSI

NIDN 0301038302

UNIVERSITAS NASIONAL

LEMBAR PERSETUJUAN JUDUL YANG TIDAK ATAU YANG DIREVISI

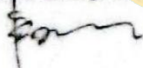
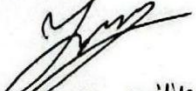
Nama : Yogie Fadila Rahman
 NPM : 227064516090
 Fakultas/Akademi : Fakultas Teknologi Komunikasi dan Informatika
 Program Studi : Informatika
 Tanggal Sidang : 25 Februari 2026

JUDUL DALAM BAHASA INDONESIA :

OPTIMASI DESKRIPSI GAMBAR BERBASIS VISION TRANSFORMER DAN
 GENERATIVE PRETRAINED TRANSFORMER-2 UNTUK PRODUK PAKAIAN

JUDUL DALAM BAHASA INGGRIS :

IMAGE DESCRIPTION OPTIMIZATION BASED ON VISION TRANSFORMER
 AND GENERATIVE PRETRAINED TRANSFORMER-2 FOR APPAREL
 PRODUCTS

TANDA TANGAN DAN TANGGAL		
Pembimbing 1	Ka Prodi	Mahasiswa
TGL : 5/2 2026.	TGL : 3/3 2026	TGL : 3/3 2026
 Dr. Fadiah, S. Kom, Mm	 Suhono, S. Kom, Mm	 Yogie Fadila R MM., MMSI

ABSTRAK

Perkembangan pesat industri e-commerce menuntut ketersediaan deskripsi produk yang akurat, konsisten, dan informatif guna membantu konsumen dalam memahami serta membandingkan produk. Namun, penyusunan deskripsi produk secara manual memerlukan waktu dan biaya yang besar serta berpotensi menimbulkan ketidakkonsistenan gaya penulisan. Oleh karena itu, penelitian ini bertujuan untuk mengembangkan dan mengoptimalkan sistem image captioning berbasis Vision Transformer (ViT) sebagai encoder dan Generative Pre-trained Transformer-2 (GPT-2) sebagai decoder untuk menghasilkan deskripsi produk fesyen secara otomatis dalam Bahasa Indonesia. Penelitian ini menggunakan pendekatan eksperimen dengan memanfaatkan dataset H&M Fashion Caption yang terdiri dari citra produk fesyen. Model dibangun menggunakan arsitektur VisionEncoderDecoderModel, di mana ViT berperan dalam mengekstraksi fitur visual global dari gambar produk, sementara GPT-2 menghasilkan teks deskripsi secara autoregresif. Untuk meningkatkan relevansi dan kesesuaian deskripsi terhadap gambar, diterapkan proses reranking berbasis CLIP, serta translasi otomatis dari Bahasa Inggris ke Bahasa Indonesia menggunakan MarianMT. Evaluasi dilakukan dengan beberapa skema pembagian data, yaitu 80:10:10, 70:15:15, dan 80:20, untuk mengamati pengaruh proporsi data terhadap performa model. Kinerja model dievaluasi menggunakan metrik ROUGE-L untuk mengukur kesesuaian struktur dan urutan kata antara caption prediksi dan caption referensi. Hasil penelitian menunjukkan bahwa konfigurasi pembagian data 70% data pelatihan, 15% data validasi, dan 15% data pengujian menghasilkan performa terbaik dengan nilai ROUGE-L sebesar 0,3460. Hasil ini menunjukkan bahwa integrasi ViT dan GPT-2 mampu menghasilkan deskripsi produk yang relevan, informatif, dan cukup natural dalam Bahasa Indonesia, sehingga berpotensi diterapkan sebagai solusi otomatisasi pembuatan deskripsi produk pada platform e-commerce.

Kata kunci: image captioning, Vision Transformer, GPT-2, deskripsi produk, e-commerce, Bahasa Indonesia, ROUGE-L.

ABSTRACT

The rapid growth of the e-commerce industry demands the availability of accurate, consistent, and informative product descriptions to assist consumers in understanding and comparing products. However, manual product description creation requires significant time and cost and is prone to inconsistencies in writing style. Therefore, this study aims to develop and optimize an image captioning system based on Vision Transformer (ViT) as the encoder and Generative Pre-trained Transformer-2 (GPT-2) as the decoder to automatically generate fashion product descriptions in the Indonesian language. This study employs an experimental approach using the H&M Fashion Caption dataset, which consists of fashion product images. The model is constructed using the VisionEncoderDecoderModel architecture, where ViT is responsible for extracting global visual features from product images, while GPT-2 generates textual descriptions in an autoregressive manner. To improve the relevance and alignment between images and generated descriptions, a CLIP-based reranking process is applied, followed by automatic translation from English to Indonesian using MarianMT. Evaluation is conducted using several data split configurations, namely 80:10:10, 70:15:15, and 80:20, to analyze the impact of data proportion on model performance. Model performance is evaluated using the ROUGE-L metric to measure structural and sequential similarity between predicted captions and reference captions. The experimental results indicate that the 70% training, 15% validation, and 15% testing data split achieves the best performance, with a ROUGE-L score of 0.3460. These results demonstrate that the integration of ViT and GPT-2 is capable of producing relevant, informative, and reasonably natural Indonesian product descriptions, indicating its potential application as an automated product description solution for e-commerce platforms.

Keywords: image captioning, Vision Transformer, GPT-2, product description, e-commerce, Indonesian language, ROUGE-L.

DAFTAR ISI

KATA PENGANTAR	i
ABSTRAK	iii
ABSTRACT	iv
DAFTAR ISI	v
DAFTAR TABEL	viii
DAFTAR GAMBAR	ix
DAFTAR RUMUS	x
BAB I	1
PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Identifikasi Masalah	3
1.3 Tujuan Penelitian	3
1.4 Manfaat Penelitian	4
1.5 Batasan Masalah	4
1.6 Matakuliah yang Mendasari	5
BAB II	6
TINJAUAN PUSTAKA	6
2.1 Penelitian Terdahulu	6
2.2 Research Gap	8
2.3 Landasan Teori	10
2.3.1 Image Captioning	10
2.3.2 Vision Transformer (ViT)	11
2.3.3 Generative Pre-trained Transformer 2 (GPT-2)	12
2.3.4 MarianMT (Neural Machine Translation)	14
2.3.5 Evaluasi Metrik pada Captioning	15
2.4 Kajian Pustaka	15
2.4.1 Penelitian Terkait Image Captioning	16
2.4.2 Perkembangan Model Berbasis Transformer	16
2.4.3 Integrasi CLIP dan GPT-2 dalam Sistem Captioning	16
2.4.4 Penerapan Model Bilingual dan MarianMT	17

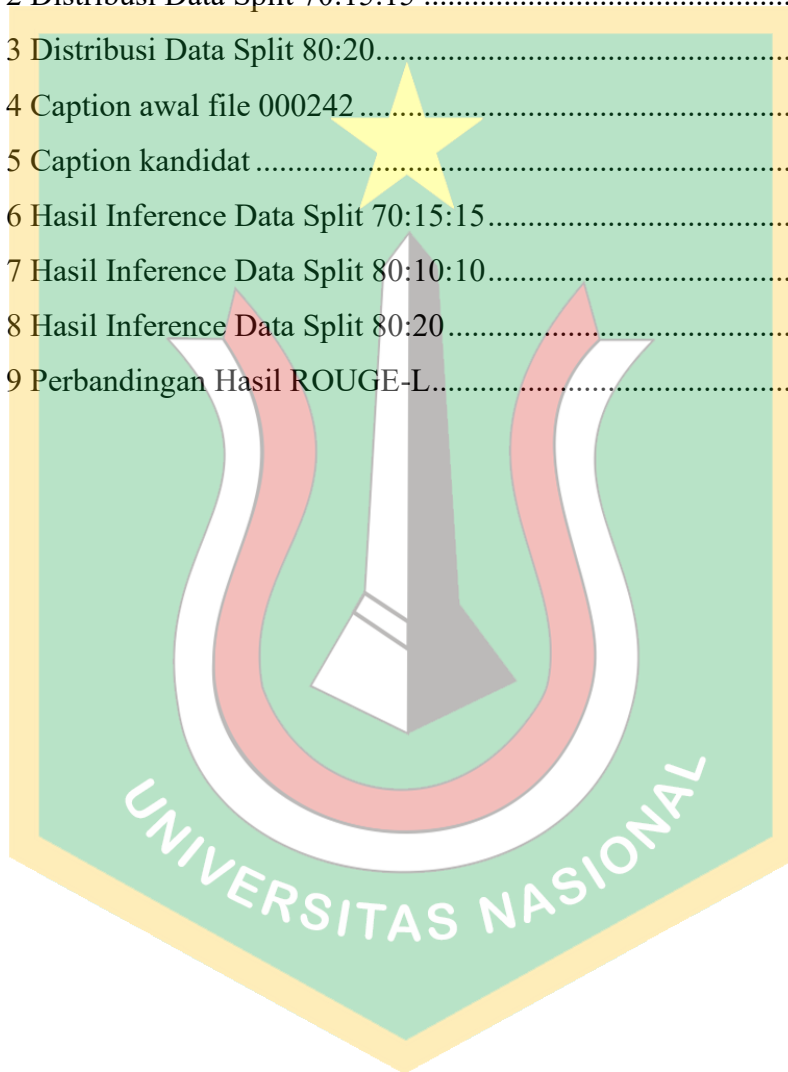
2.4.5	Optimalisasi dan Evaluasi Sistem Captioning	17
2.4.6	Relevansi Penelitian	18
BAB III		19
METODE PENELITIAN		19
3.1	Jenis Penelitian	19
3.2	Fokus Penelitian	19
3.3	Sumber Data	20
3.4	Alur Penelitian	21
3.4.1	Identifikasi Masalah	22
3.4.2	Studi Literatur	22
3.4.3	Perumusan Model Penelitian	22
3.4.4	Pengumpulan Data	23
3.4.5	Preprocessing Data	23
3.4.6	Pelatihan Model	23
3.4.7	Teknik Regularisasi dan Pencegahan Overfitting	24
3.4.8	Eksperimen dan Evaluasi Model	25
3.4.9	Analisis Hasil	26
3.4.10	Kesimpulan dan Saran	26
3.5	Waktu Penelitian	27
3.6	Teknik Pengumpulan Data	27
3.7	Peralatan Penelitian	28
3.7.1	Perangkat Keras	28
3.7.2	Perangkat Lunak	28
BAB IV		31
HASIL DAN PEMBAHASAN		31
4.1	Eksplorasi Dataset	31
4.2	Preprocessing Data	32
4.2.1	Preprocessing Image	33
4.2.2	Captioning Image	34
4.3	Pelatihan Model	37
4.3.1	Grafik Training Loss dan Validation Loss	38
4.3.2	Analisis Overfitting dan Generalisasi Model	40

4.4	Hasil dan Evaluasi Metrik.....	41
4.4.1	Hasil Inference	41
4.4.2	Evaluasi Metrik dengan ROUGE-L	42
4.5	Implementasi Streamlit	44
4.5.1	Tampilan Utama Aplikasi	45
4.5.2	Tampilan Proses Captioning Gambar	45
BAB V		47
KESIMPULAN DAN SARAN		47
5.1	Kesimpulan	47
5.2	Saran.....	48
DAFTAR PUSTAKA		49
LAMPIRAN		51



DAFTAR TABEL

Tabel 2. 1 Penelitian Terdahulu	6
Tabel 3. 1 Tabel Waktu Penelitian.....	27
Tabel 3. 2 Split Data.....	28
Tabel 4. 1 Distribusi Data Split 80:10:10	31
Tabel 4. 2 Distribusi Data Split 70:15:15	32
Tabel 4. 3 Distribusi Data Split 80:20.....	32
Tabel 4. 4 Caption awal file 000242	35
Tabel 4. 5 Caption kandidat	36
Tabel 4. 6 Hasil Inference Data Split 70:15:15.....	42
Tabel 4. 7 Hasil Inference Data Split 80:10:10.....	42
Tabel 4. 8 Hasil Inference Data Split 80:20.....	42
Tabel 4. 9 Perbandingan Hasil ROUGE-L.....	43



DAFTAR GAMBAR

Gambar 2. 1 ViT Patch Embedding	12
Gambar 2. 2 Transformer Encoder.....	14
Gambar 3. 1 Dataset.....	21
Gambar 3. 2 Diagram Alur Penelitian.....	21
Gambar 4. 1 Gambaran Dataset	32
Gambar 4. 2 Data sesudah resize	33
Gambar 4. 3 Data sebelum resize.....	33
Gambar 4. 4 Gambar sebelum cleaning	34
Gambar 4. 5 Gambar sesudah cleaning.....	34
Gambar 4. 6 Caption English.....	35
Gambar 4. 7 Contoh file 000242.....	35
Gambar 4. 8 Contoh file 002645.....	36
Gambar 4. 9 Caption ID (MarianMT).....	37
Gambar 4. 10 Caption ID (mBART50).....	37
Gambar 4. 11 Caption ID (mT5).....	37
Gambar 4. 12 Diagram Garis Split 80/10/10	39
Gambar 4. 13 Diagram Garis Split 70/15/15	39
Gambar 4. 14 Diagram Garis Split 80/20.....	40
Gambar 4. 15 Contoh Gambar Dataset	43
Gambar 4. 16 Tampilan Utama Streamlit	45
Gambar 4. 17 Tampilan Saat Proses Generate Teks	46
Gambar 4. 18 Generate Teks Kemeja Hitam	46
Gambar 4. 19 Generate Teks Kemeja Putih.....	46

DAFTAR RUMUS

(2.1)	11
(2.2)	12
(2.3)	12
(2.4)	12
(2.5)	13
(2.6)	13
(2.7)	13
(2.8)	13
(2.9)	15
(2.10)	15

