KEMENTERIAN RISET, TEKNOLOGI, DAN PENDIDIKAN TINGGI
UNIVERSITAS SUMATERA UTARA
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
Jalan Universitas No. 9A Kampus USU, Medan 20155
el/Fax: 061 8228048, e-mail: fasilkomti@usu.ac.id, laman: http://fasilkom-ti.usu.ac.id

FORM PENGAJUAN JUDUL

Nama : Andhika Mandalanta Saragih

NIM : 211401076

Judul diajukan oleh* : ☐ Dosen
☑ Mahasiswa

Bidang Ilmu (tulis dua bidang) : *Generative AI, Computer Vision*

Uji Kelayakan Judul** : ○ Diterima ○ Ditolak

Hasil Uji Kelayakan Judul :

Calon Dosen Pembimbing I:
Dr. T. Henny Febriana Harumy S.Kom., M.Kom
(Jika judul dari dosen maka dosen tersebut berhak menjadi pembimbing I)

Paraf Calon Pembimbing 1

Calon Dosen Pembimbing II:
Sri Melvani Hardi S.Kom., M.Kom

Paraf Calon Pembimbing 2

Medan, 17 Januari 2025
Ka. Laboratorium Penelitian,

* *Centang salah satu atau keduanya*
** *Pilih salah satu*

Dr. Jos Timanta Tarigan S.Kom., M.Sc.
NIP. 198501262015041001

KEMENTERIAN RISET, TEKNOLOGI, DAN PENDIDIKAN TINGGI
UNIVERSITAS SUMATERA UTARA
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
Jalan Universitas No. 9A Kampus USU, Medan 20155
el/Fax: 061 8228048, e-mail: fasilkomti@usu.ac.id, laman: http://fasilkom-ti.usu.ac.id

RINGKASAN JUDUL YANG DIAJUKAN

*Semua kolom dibawah ini diisi oleh mahasiswa yang sudah mendapat judul

Judul / Topik Skripsi	Optimasi Caption Otomatis: Studi Refinement Caption dari Model Vision-Language Menggunakan GPT-3.5
Latar Belakang	Dalam era digital saat ini, penggunaan media sosial dan platform berbasis gambar seperti Instagram, Pinterest, serta e-commerce semakin meningkat. Platform-platform ini memungkinkan pengguna untuk membagikan konten visual dengan jutaan pengguna lainnya (Appel et al., 2020). Namun, salah satu tantangan yang sering dihadapi pengguna adalah kesulitan dalam membuat caption yang sesuai dengan gambar yang diunggah (Ghandi et al., 2023). Caption yang menarik dan relevan dapat meningkatkan keterlibatan pengguna dan memberikan konteks yang lebih jelas terhadap gambar yang ditampilkan. Oleh karena itu, muncul kebutuhan akan sistem otomatis yang dapat menghasilkan caption berkualitas tinggi dengan lebih cepat dan efisien (Phukan & Panda, 2021). Berbagai penelitian telah dilakukan untuk mengembangkan sistem image captioning yang mengintegrasikan visi komputer dan pemrosesan bahasa alami (Natural Language Processing/NLP). Pendekatan awal yang populer adalah model yang menggabungkan jaringan konvolusional untuk ekstraksi fitur gambar dan jaringan rekuren untuk menghasilkan teks deskriptif. Model ini kemudian dikembangkan lebih lanjut dengan mekanisme atensi, memungkinkan fokus pada bagian gambar yang lebih relevan saat menghasilkan deskripsi. Sejak 2020, pendekatan berbasis Transformer telah menjadi dominan dalam bidang ini. Vision Transformers telah muncul sebagai alternatif yang kompetitif untuk jaringan konvolusi untuk tugas-tugas pemberian keterangan gambar, menunjukkan akurasi yang lebih tinggi dalam menghasilkan deskripsi yang terperinci (Fang et al., 2022). Selanjutnya, penelitian oleh (Dandwate et al., 2023) membandingkan kinerja model Transformer dan LSTM dengan mekanisme atensi dalam tugas image captioning, menunjukkan bahwa arsitektur Transformer dapat menghasilkan deskripsi gambar yang lebih akurat dan efisien. Selain itu, (Osman et al., 2024) mengusulkan model captioning berbasis konsep menggunakan arsitektur multi-encoder Transformer, yang secara signifikan meningkatkan kualitas deskripsi gambar dengan menangkap konteks gambar secara lebih efektif. Seiring dengan perkembangan teknologi, model berbasis Transformer mulai diterapkan dalam image captioning, seperti BLIP (Bootstrapped Language-Image Pretraining) (Li et al., 2022), yang menunjukkan peningkatan signifikan dalam pemahaman gambar dan teks secara bersamaan. BLIP memanfaatkan teknik pretraining untuk meningkatkan keterhubungan antara representasi visual dan bahasa,

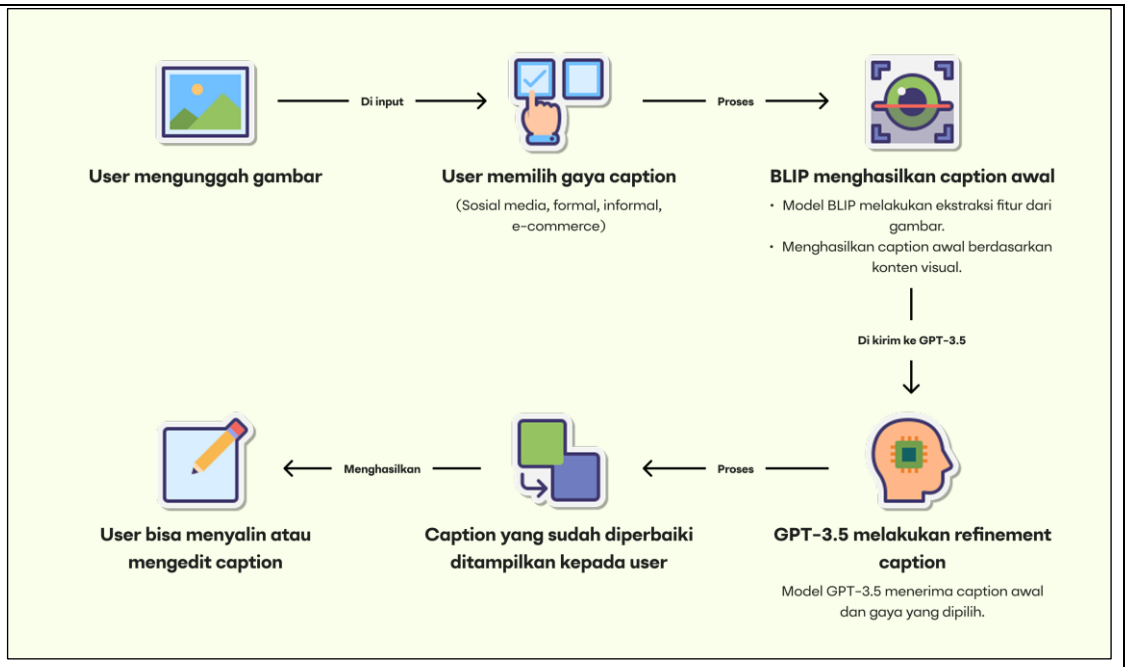
	sehingga mampu menghasilkan deskripsi yang lebih akurat dan relevan dibandingkan pendekatan sebelumnya. Selain BLIP, model berbasis Transformer lainnya seperti Flamingo telah menunjukkan kemampuan dalam memahami dan menghasilkan deskripsi gambar yang akurat. Flamingo adalah model bahasa visual yang dirancang untuk pembelajaran sedikit tembakan (few-shot learning), memungkinkan adaptasi cepat terhadap tugas baru dengan sedikit contoh. Flamingo dapat menangani berbagai tugas multimodal, termasuk captioning, dialog visual, klasifikasi, dan pertanyaan-pertanyaan visual. Kemampuan ini menjadikannya alat yang kuat dalam meningkatkan kualitas caption yang dihasilkan oleh model vision-language (Alayrac et al., 2022). Selain itu, evaluasi captioning juga menjadi tantangan besar dalam pengembangan model vision-language. Meskipun berbagai metrik seperti BLEU, METEOR, dan CIDEr digunakan untuk mengukur kesamaan semantik antara caption yang dihasilkan dan ground truth, namun metrik ini sering kali tidak dapat menangkap aspek keberagaman, koherensi, dan kualitas naratif dari deskripsi gambar. Caption yang menarik dan informatif tidak hanya bergantung pada kesamaan kata-kata, tetapi juga pada kemampuan untuk mengomunikasikan konteks gambar secara lebih mendalam dan relevan bagi audiens. Oleh karena itu, diperlukan metrik evaluasi yang lebih menyeluruh, yang tidak hanya mengukur kesamaan leksikal tetapi juga kualitas dan keterlibatan naratif dari caption yang dihasilkan. Penelitian terbaru mengusulkan penggunaan perceived quality atau kualitas yang diterima oleh manusia, serta model evaluasi berbasis deep learning yang mampu memahami dan menilai kualitas naratif secara lebih holistik (Luo et al., 2022). Seiring dengan kemajuan teknologi dalam model vision-language berbasis Transformer, pengembangan Large Language Models (LLM) seperti GPT-3.5 juga telah membawa revolusi dalam pemrosesan bahasa alami. GPT-3.5 memiliki kemampuan untuk memahami dan menghasilkan teks dengan kualitas tinggi berdasarkan konteks yang diberikan (Kaplan et al., 2020). Model ini dapat digunakan untuk menyempurnakan caption yang dihasilkan oleh model vision-language agar lebih natural, menarik, dan sesuai dengan gaya bahasa yang diinginkan pengguna. Meskipun berbagai penelitian telah mengeksplorasi penggunaan model vision-language dan LLM secara terpisah, studi yang menggabungkan kedua pendekatan ini dalam proses refinement caption otomatis masih terbatas. Oleh karena itu, penelitian ini bertujuan untuk mengeksplorasi bagaimana model vision-language seperti BLIP dapat digunakan untuk menghasilkan caption awal dari gambar, dan bagaimana GPT-3.5 dapat meningkatkan kualitas caption tersebut agar lebih akurat, natural, dan menarik bagi pengguna.

KEMENTERIAN RISET, TEKNOLOGI, DAN PENDIDIKAN TINGGI
UNIVERSITAS SUMATERA UTARA
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
Jalan Universitas No. 9A Kampus USU, Medan 20155
el/Fax: 061 8228048, e-mail: fasilkomti@usu.ac.id, laman: http://fasilkom-ti.usu.ac.id

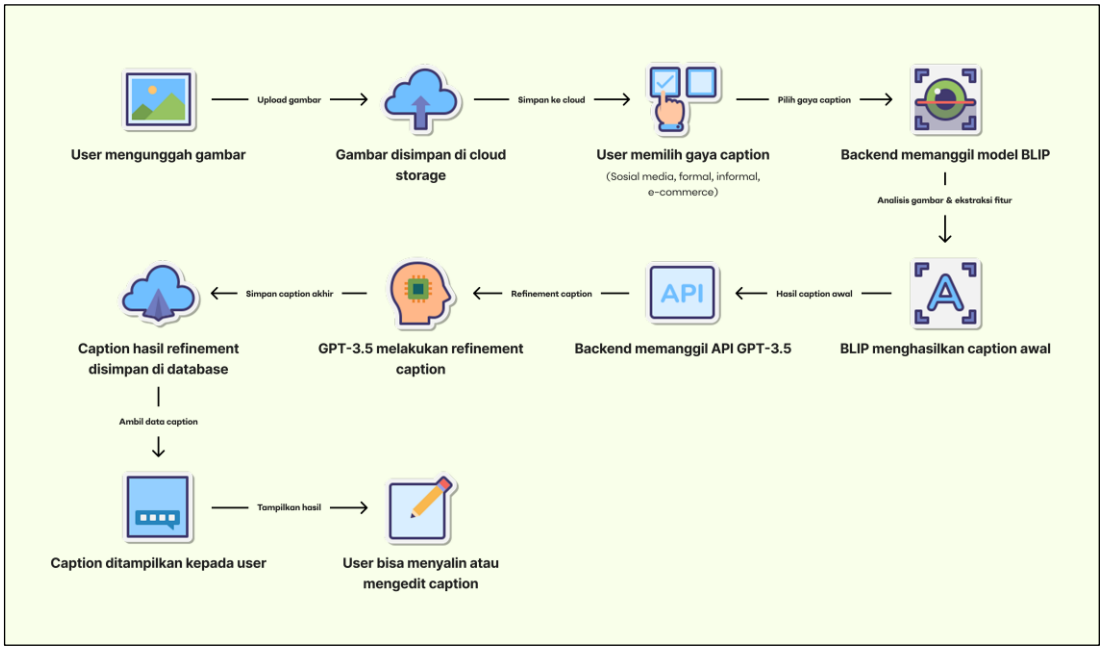
Rumusan Masalah	Dalam era digital, sistem image captioning berperan penting dalam menghasilkan deskripsi otomatis untuk gambar, namun model vision-language seperti BLIP masih memiliki keterbatasan dalam menghasilkan caption yang benar-benar natural, menarik, dan sesuai konteks. Di sisi lain, model bahasa besar seperti GPT-3.5 menawarkan potensi dalam menyempurnakan caption melalui proses refinement, tetapi efektivitasnya dalam meningkatkan kualitas caption dari model vision-language masih perlu dieksplorasi lebih lanjut. Oleh karena itu, penelitian ini berfokus pada optimalisasi caption otomatis dengan menggabungkan BLIP untuk menghasilkan caption awal dan GPT-3.5 untuk menyempurnakannya, serta mengevaluasi peningkatannya menggunakan metrik teks seperti BLEU, METEOR, ROUGE, dan CIDEr, serta umpan balik subjektif dari pengguna.

Metodologi	1. Pengumpulan Data
	  o Menggunakan dataset gambar dan caption seperti Flickr30k sebagai data latih dan uji.
	2. Implementasi Model
	  o Menggunakan BLIP sebagai model vision-language untuk menghasilkan caption awal dari gambar.
	  o Caption awal tersebut kemudian diolah oleh GPT-3.5 untuk proses refinement, agar lebih natural dan sesuai konteks.
	3. Evaluasi Hasil
	  o Melakukan perbandingan antara caption sebelum dan sesudah refinement menggunakan metrik evaluasi teks seperti:
	  ▪ BLEU (Bilingual Evaluation Understudy)
	  ▪ METEOR (Metric for Evaluation of Translation with Explicit ORdering)
	  ▪ ROUGE (Recall-Oriented Understudy for Gisting Evaluation)
	  ▪ CIDEr (Consensus-based Image Description Evaluation)
	  o Melakukan uji subjektif kepada pengguna untuk menilai keterbacaan dan kualitas caption.

KEMENTERIAN RISET, TEKNOLOGI, DAN PENDIDIKAN TINGGI
UNIVERSITAS SUMATERA UTARA
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
Jalan Universitas No. 9A Kampus USU, Medan 20155
el/Fax: 061 8228048, e-mail: fasilkomti@usu.ac.id, laman: http://fasilkom-ti.usu.ac.id



Gambar 1. Alur bagaimana BLIP dan GPT-3.5 digunakan untuk caption refinement



Gambar 2. Pipeline Proses Implementasi Sistem Caption Refinement

KEMENTERIAN RISET, TEKNOLOGI, DAN PENDIDIKAN TINGGI
UNIVERSITAS SUMATERA UTARA
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
Jalan Universitas No. 9A Kampus USU, Medan 20155
el/Fax: 061 8228048, e-mail: fasilkomti@usu.ac.id, laman: http://fasilkom-ti.usu.ac.id

Referensi

Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Han, S. C. T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., … Simonyan, K. (2022). Flamingo: a Visual Language Model for Few-Shot Learning. *Advances in Neural Information Processing Systems*, *35*(NeurIPS).

Appel, G., Grewal, L., Hadi, R., & Stephen, A. T. (2020). The future of social media in marketing. *Journal of the Academy of Marketing Science*, *48*(1), 79–95. https://doi.org/10.1007/s11747-019-00695-1

Dandwate, P., Shahane, C., Jagtap, V., & Karande, S. C. (2023). Comparative Study of Transformer and LSTM Network with Attention Mechanism on Image Captioning. *Lecture Notes in Networks and Systems*, *720 LNNS*, 527–539. https://doi.org/10.1007/978-981-99-3761-5_47

Fang, Z., Wang, J., Hu, X., Liang, L., Gan, Z., Wang, L., Yang, Y., & Liu, Z. (2022). Injecting Semantic Concepts into End-to-End Image Captioning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2022-June*, 17988–17998. https://doi.org/10.1109/CVPR52688.2022.01748

Ghandi, T., Pourreza, H., & Mahyar, H. (2023). Deep Learning Approaches on Image Captioning: A Review. *ACM Computing Surveys*, *56*(3). https://doi.org/10.1145/3617592

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling Laws for Neural Language Models*. http://arxiv.org/abs/2001.08361

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *Proceedings of Machine Learning Research*, *162*(2), 12888–12900.

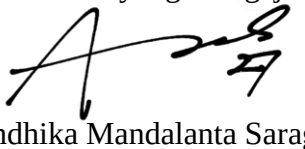
Luo, G., Cheng, L., Jing, C., Zhao, C., & Song, G. (2022). A thorough review of models, evaluation metrics, and datasets on image captioning. *IET Image Processing*, *16*(2), 311–332. https://doi.org/10.1049/ipr2.12367

Osman, A. A. E., Shalaby, M. A. W., Soliman, M. M., & Elsayed, K. M. (2024). Novel concept-based image captioning models using LSTM and multi-encoder transformer architecture. *Scientific Reports*, *14*(1), 1–15. https://doi.org/10.1038/s41598-024-69664-1

Phukan, B. B., & Panda, A. R. (2021). *An Efficient Technique for Image Captioning Using Deep Neural Network*. 481–491. https://doi.org/10.1007/978-981-16-1056-1_38

Medan, 17 Februari 2025

Mahasiswa yang mengajukan,



Andhika Mandalanta Saragih

NIM. 211401076