BAB II

TINJAUAN PUSTAKA

2.1 Penelitian Terdahulu

Penulis	Judul	Jurnal / Tempat Publikasi (Lengkap)	Metode	Hasil	Relevansi dengan Penelitian Ini
H. A. Alawwad, A. Zafar, A. Alhothali, U. Naseem, A. Alkhathlan, A. Jamal [1]	*Evaluating Multimodal Large Language Models on Educational Textbook Question Answering*	*Proceedings of the International Generative AI, Computing and Language Models Conference (GACLM), 2025*	Evaluasi MLLM pada tugas QA multimodal berbasis gambar dan teks buku pelajaran	Menunjukkan bahwa MLLM mampu memahami dan menjawab pertanyaan berbasis visual dengan akurasi tinggi	Memberikan dasar bahwa model multimodal relevan digunakan untuk konteks edukasi berbasis gambar di Tigaraksa
P. Xu, W. Shao, K. Zhang, P. Gao, S. Liu, M. Lei, F. Meng, S. Huang, Y. Qiao, P. Luo [2]	*LVLM-eHub: A Comprehensive Evaluation Benchmark for Large Vision-Language Models*	*arXiv Preprint (Cornell University Repository)*	Benchmark evaluasi untuk mengukur performa *Vision-Language Models* pada berbagai task	Memberikan standar perbandingan performa VLM modern	Mendukung pemilihan model vision (BLIP) dalam penelitian ini.

J. Yi, J. Yin, J. Xu, P. Bao, Y. Wang, W. Fan, H. Wang [3]	ImageRef-VL: Enabling Contextual Image Referencing in Vision-Language Models	*arXiv Preprint*	Mengembangkan model yang mampu memahami konteks referensi gambar dalam percakapan	Meningkatkan akurasi model dalam menghubungkan pertanyaan dengan isi visual	Relevan untuk memastikan chatbot memahami metadata gambar secara tepat.
G. Luo, Y. Zhou, T. Ren, S. Chen, X. Sun, R. Ji [4]	Cheap and Quick: Efficient Vision-Language Instruction Tuning for LLM	*arXiv Preprint*	Instruction tuning untuk meningkatkan kemampuan reasoning visual pada LLM	Menurunkan biaya pelatihan sambil meningkatkan kepatuhan model terhadap instruksi	Penting untuk desain prompt dan kontrol perilaku SLM pada chatbot.
R. Janssens, P. Wolfert, T. Demeester, T. Belpaeme [5]	Integrating Visual Context into Language Models for Situated Social Conversation Starters	*IEEE Transactions on Affective Computing*	Integrasi *visual-context* ke modul bahasa	Konteks visual meningkatkan relevansi respons, terutama pada percakapan sosial berbasis gambar	Mendukung gagasan chatbot yang memberikan penjelasan berdasarkan isi gambar.
L. Yan, L. Zhao, V. Echeverria, Y. Jin, R. Alfredo, X. Li, D. Gašević, R. Martinez-Maldonado [6]	VizChat: Enhancing Learning Analytics Dashboards Using Multimodal Generative AI Chatbots	*Lecture Notes in Computer Science (Artificial Intelligence in Education, Springer)*	Chatbot multimodal untuk analitik pembelajaran	Meningkatkan pemahaman siswa terhadap materi visual	Menjadi fondasi penggunaan chatbot multimodal dalam konteks belajar anak-anak.

E. Markin, V. Zuparova, A. Martyshkin [7]	Integration of Large Language Models and Computer Vision Algorithms in LMS	*Proceedings of the International Russian Smart Industry Conference (SmartIndustry Con), 2025*	Integrasi LLM + Computer Vision	Efektif untuk verifikasi otomatis tugas berbasis gambar dan analisis data pendidikan	Menguatkan integrasi Vision AI + bahasa sebagai pipeline analisis.
M. Y. Lee [8]	Building Multimodal AI Chatbots	*Undergraduate Bachelor Thesis (2023)*	Mengembangkan arsitektur chatbot multimodal	Menyediakan panduan arsitektural membangun chatbot multimodal end-to-end	Dijadikan referensi untuk pipeline chatbot multimodal kamu.
A. Mohammed [9]	Multimodal AI Architectures: Integrating Vision and Language for Enhanced Scene Understanding	*International Journal of Multidisciplinary Research in Science, Engineering and Technology*	Menggabungkan model vision dan bahasa untuk scene understanding	Memperkuat akurasi pemahaman konteks visual oleh model	Mendukung penggunaan BLIP + OCR + embedding dalam pipeline riset.
C. Chen, D. Han, C.-C. Chang [10]	MPCCT: Multimodal Vision-Language Learning Paradigm with Context-Based Compact Transformer	*Pattern Recognition (Elsevier)*	Compact transformer multimodal	Efisien tetapi tetap akurat untuk tugas vision-language	Menguatkan alasan menggunakan model ringan seperti SLM
J. Chan, Y. Li [11]	Enhancing Higher Education with Generative AI: A Multimodal Approach for Personalised Learning	*New Technology in Education and Training (LNEDT)*	Multimodal generative AI di pendidikan	Meningkatkan engagement & personalisasi	Mendukung penggunaan chatbot multimodal untuk pembelajaran di Tigaraksa

A. Bewersdorff et al. [13]	Taking the Next Step with Generative Artificial Intelligence: The Transformative Role of MLLM in Science Education	*Learning and Individual Differences (Elsevier)*	Studi implementasi MLLM	MLLM terbukti meningkatkan pemahaman konsep ilmiah	Menjadi bukti bahwa multimodal AI efektif untuk pendidikan
S. Ahmed, M. T. Younes, A. Moustafa, A. Allam, H. Moustafa [14]	MSA at ImageCLEF 2025 Multimodal Reasoning	*CLEF Working Notes (Conference and Labs of the Evaluation Forum)*	Ensemble Vision-Language Models	Meningkatkan akurasi reasoning dan pemahaman visual	Relevan untuk memperkuat reasoning berbasis visual
X. Xu et al. [15]	Development and Evaluation of a Retrieval-Augmented Large Language Model Framework for Enhancing Endodontic Education	*International Journal of Medical Informatics*	Mengintegrasikan RAG untuk reasoning pendidikan	RAG mengurangi halusinasi dan meningkatkan akurasi jawaban	Sangat relevan karena RAG adalah inti sistem chatbot dari penelitian ini
T. Gong, C. Lyu, S. Zhang, Y. Wang, M. Zheng, Q. Zhao, K. Liu, W. Zhang, P. Luo, K. Chen [18]	MultiModal-GPT: A Vision and Language Model for Dialogue with Humans	*arXiv Preprint*	Dialog multimodal berbasis gambar	Mampu menjawab berdasarkan input visual	Fondasi untuk chatbot percakapan berbasis metadata gambar

Berdasarkan Tabel 2.1, dapat disimpulkan bahwa pengembangan *multimodal Artificial Intelligence* telah mengalami kemajuan signifikan, mencakup kemampuan pemahaman visual, *image-based reasoning*, serta integrasi antara *vision model* dan *language model*. Sejumlah penelitian terdahulu telah mengeksplorasi kapabilitas *multimodal AI* dalam konteks pendidikan. Penelitian [1] mengevaluasi kemampuan model multimodal dalam menjawab pertanyaan berbasis buku pelajaran dan menunjukkan bahwa kualitas *reasoning* sangat bergantung pada kemampuan model dalam mengekstraksi serta merepresentasikan informasi visual secara akurat. Penelitian [2] memperkenalkan *large-scale multimodal benchmark* untuk mengukur performa *Vision-Language Models* pada berbagai tugas, seperti *visual grounding*, pemahaman objek, dan *Optical Character Recognition* (OCR), sehingga menjadi rujukan penting dalam evaluasi pemahaman visual model.

Selanjutnya, penelitian [3] mengembangkan model yang mendukung *multi-image referencing* dalam percakapan, yang memperkuat kemampuan AI dalam mengaitkan teks dengan konteks visual secara presisi. Pendekatan *multimodal fine-tuning* juga ditunjukkan dalam penelitian [4], yang menegaskan bahwa *instruction tuning* yang tepat dapat meningkatkan kualitas respons multimodal tanpa memerlukan sumber daya komputasi yang besar. Pada ranah percakapan berbasis visual, penelitian [5] menunjukkan bahwa integrasi konteks gambar dapat meningkatkan relevansi dialog, sementara penelitian [6] menghadirkan chatbot multimodal untuk *learning analytics dashboard* yang mampu memberikan penjelasan berbasis visualisasi data.

Integrasi *Large Language Model* (LLM) dan algoritma *computer vision* untuk otomasi penilaian akademik ditunjukkan dalam penelitian [7], yang memperkuat argumen bahwa *multimodal reasoning* memiliki peran penting dalam domain pendidikan. Penelitian [8] memberikan fondasi konseptual pembangunan chatbot multimodal, sedangkan penelitian [9] mengkaji arsitektur multimodal yang menggabungkan ekstraksi fitur visual dengan *reasoning* untuk *scene understanding*. Pendekatan *multimodal transformer* yang lebih efisien dijelaskan

dalam penelitian [10], sementara penelitian [11] membahas pemanfaatan *generative AI* dalam pembelajaran berbasis multimodal.

Dalam konteks pendidikan sains, penelitian [13] menunjukkan bahwa *multimodal LLM* dapat meningkatkan pemahaman konsep ilmiah melalui kombinasi teks dan visual. Penelitian [14] menekankan pentingnya *visual reasoning* melalui *ensemble vision-language models* dalam kompetisi *ImageCLEF*. Sementara itu, penelitian [15] mengusulkan kerangka *Retrieval-Augmented Generation* (RAG) untuk pendidikan kedokteran dan membuktikan bahwa integrasi *retrieval* mampu meningkatkan akurasi serta menurunkan risiko *hallucination*. Selain itu, penelitian [18] memperkenalkan *Multimodal-GPT*, yaitu model yang mampu menerima input visual sebagai dasar percakapan.

Meskipun demikian, tinjauan terhadap penelitian terdahulu menunjukkan bahwa sebagian besar studi masih berfokus pada pengembangan model, arsitektur, *benchmark*, atau penerapan pada pendidikan tingkat lanjut. Hingga saat ini, belum terdapat penelitian yang secara spesifik menargetkan konteks pembelajaran berbasis komunitas, khususnya lingkungan desa. Selain itu, belum ada penelitian yang memanfaatkan gambar edukatif lokal hasil *Text-to-Image* (T2I) sebagai basis pengetahuan utama, maupun membangun *visual metadata pipeline* secara komprehensif. Meskipun beberapa penelitian telah mengeksplorasi *multimodal reasoning* dan integrasi RAG, belum ada yang menggabungkan *BLIP* untuk *captioning*, OCR untuk ekstraksi teks dalam gambar, *Sentence Transformer* untuk *embedding*, *Supabase Vector* sebagai *knowledge store*, serta *Small Language Model* sebagai mesin *reasoning* dalam satu sistem chatbot multimodal yang ditujukan untuk pendidikan dasar di lingkungan desa.

Penelitian ini menempati posisi berbeda dengan berfokus pada pemanfaatan gambar edukatif yang dihasilkan melalui penelitian T2I sebelumnya sebagai sumber pengetahuan utama. Selain itu, penelitian ini menekankan pengembangan chatbot multimodal yang aman, minim *hallucination*, dan relevan bagi guru serta siswa sekolah dasar. Dengan demikian, penelitian ini mengisi kesenjangan berupa

kebutuhan sistem multimodal yang tidak hanya mampu memahami gambar, tetapi juga mampu memberikan penjelasan edukatif berbasis *visual metadata* yang diekstraksi secara sistematis melalui integrasi *Vision AI*, OCR, *vector embedding*, RAG, dan *Small Language Models*.

2.1.1 Kesenjangan Penelitian

Berdasarkan tinjauan terhadap penelitian terdahulu, terdapat sejumlah kesenjangan penelitian yang menjadi dasar pengembangan penelitian ini. Sebagian besar studi sebelumnya masih berfokus pada pengembangan model AI secara umum, evaluasi benchmark atau penerapan pada jenjang pendidikan menengah hingga lanjut. Hingga saat ini, masih sangat terbatas penelitian yang secara spesifik menargetkan konteks pembelajaran berbasis komunitas desa, khususnya untuk siswa sekolah dasar. Lingkungan belajar seperti Desa Wisata Tigaraksa memiliki karakteristik lokal, keterbatasan sumber daya serta kebutuhan pedagogis yang berbeda sehingga memerlukan pendekatan sistem pembelajaran berbasis AI yang lebih kontekstual dan sesuai dengan kondisi lapangan.

Selain itu, penelitian terdahulu umumnya memanfaatkan dataset visual publik atau data yang tidak merepresentasikan konteks lokal. Masih jarang ditemukan penelitian yang secara khusus menggunakan kumpulan gambar edukatif lokal hasil teknologi Text-to-Image (T2I) sebagai basis pengetahuan utama dalam sistem pembelajaran. Padahal, gambar T2I yang telah divalidasi secara lokal memiliki potensi besar sebagai aset pembelajaran kontekstual apabila diolah menjadi informasi terstruktur yang dapat dipahami oleh mesin.

Kesenjangan berikutnya terletak pada aspek integrasi teknologi multimodal. Sebagian besar penelitian hanya berfokus pada satu atau dua komponen seperti captioning gambar atau pemodelan bahasa, tanpa mengintegrasikan keseluruhan proses pemahaman visual dan penalaran dalam satu sistem yang utuh. Hingga saat ini masih terbatas penelitian yang menggabungkan Vision Model untuk pemahaman gambar, *Optical Character Recognition (OCR)* untuk ekstraksi teks, embedding untuk pencarian semantik, *vector database* sebagai basis pengetahuan,

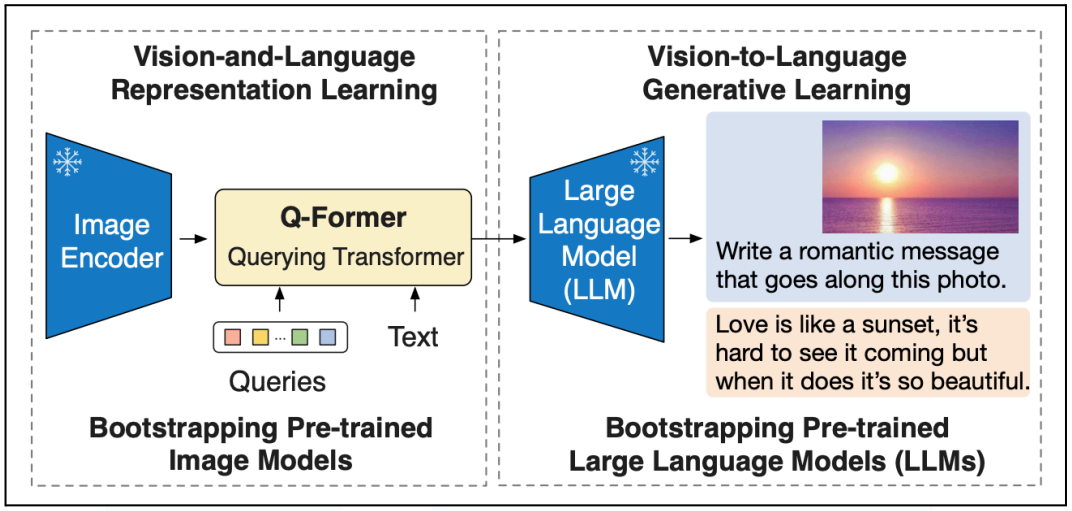
serta *Small Language Model* (SLM) untuk proses reasoning dalam satu aplikasi chatbot edukatif dasar.

Di sisi lain, penelitian sebelumnya umumnya berhenti pada tahap penyediaan visual edukatif statis. Kondisi ini menimbulkan kesenjangan dalam hal interpretasi dan interaksi, di mana gambar AI tidak memiliki penjelasan terstruktur dan pengguna tidak dapat berinteraksi atau mengajukan pertanyaan secara langsung berdasarkan isi visual. Akibatnya, pemanfaatan gambar dalam pembelajaran masih sangat bergantung pada interpretasi manual oleh guru dan belum mendukung proses belajar interaktif bagi siswa.

Terakhir, dalam konteks pendidikan anak-anak, aspek keandalan dan keamanan informasi menjadi perhatian utama. Banyak sistem AI generatif berpotensi menghasilkan jawaban yang tidak akurat atau mengalami halusinasi karena tidak memiliki mekanisme grounding yang kuat. Hingga saat ini, masih terbatas penelitian yang secara eksplisit mengintegrasikan pendekatan *Retrieval-Augmented Generation (RAG)* dengan sistem multimodal untuk memastikan bahwa setiap respons AI selalu merujuk pada fakta visual yang relevan. Penelitian ini hadir untuk mengisi kesenjangan tersebut dengan mengembangkan sistem chatbot multimodal yang mampu memahami, menjelaskan dan berinteraksi dengan gambar edukatif secara aman, faktual dan sesuai dengan kebutuhan pembelajaran siswa sekolah dasar di lingkungan desa.

2.2 Tinjauan Teori

2.2.1 Vision Language Model (VLM)



Gambar 2.1 Arsitektur Vision Language Model

Vision-Language Model (VLM) merupakan sistem *Artificial Intelligence* yang dirancang untuk memproses dan memahami informasi visual serta teks secara simultan. Model ini bekerja dengan mengekstraksi fitur dari gambar menggunakan *visual encoder*, seperti *Vision Transformer*, kemudian mengintegrasikannya dengan representasi teks yang dihasilkan oleh *language model*. Integrasi kedua representasi ini memungkinkan VLM untuk membangun pemahaman multimodal yang selaras antara visual dan bahasa. Penelitian [12] memperkenalkan pendekatan *Image as a Foreign Language* melalui metode *BEiT*, di mana gambar direpresentasikan sebagai token diskret sehingga dapat diproses dengan mekanisme yang serupa dengan teks. Pendekatan ini menjadi fondasi bagi pengembangan berbagai model multimodal modern yang mampu memahami, mendeskripsikan, serta menjawab pertanyaan berdasarkan konten gambar secara kontekstual.

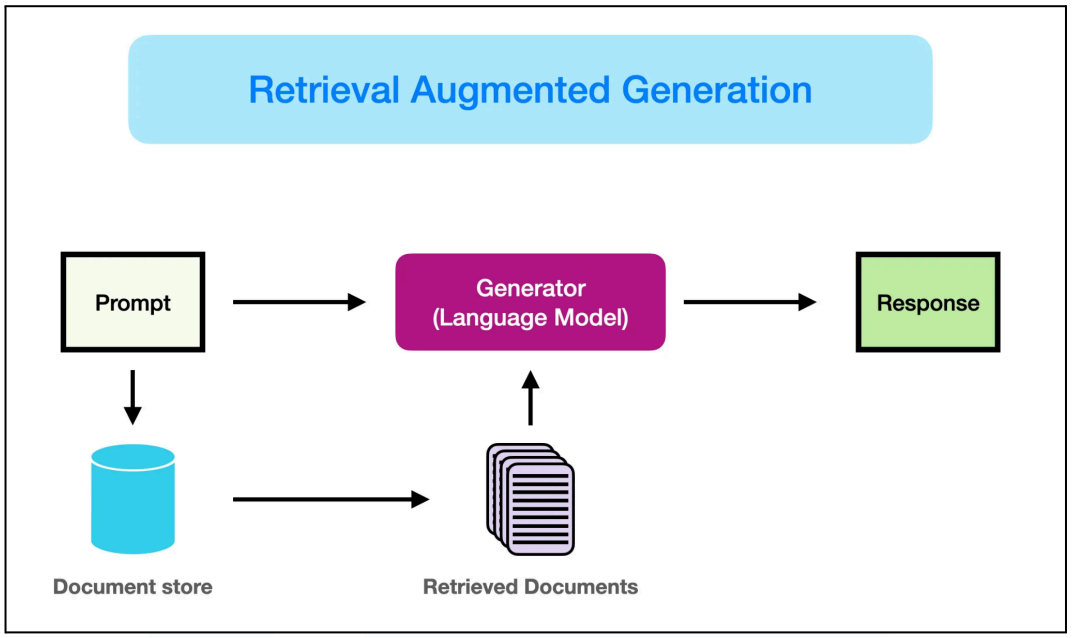
Dalam konteks pendidikan, VLM membuka peluang baru dalam pengembangan sistem pembelajaran interaktif berbasis visual. Melalui kemampuan pemahaman multimodal, VLM memungkinkan AI untuk memberikan penjelasan visual, melakukan analisis objek, serta menjalankan *reasoning* terhadap konten gambar, sehingga mendukung proses belajar yang lebih kontekstual, informatif, dan sesuai dengan kebutuhan peserta didik.

2.2.2 Multimodal Learning

Multimodal learning merujuk pada kemampuan sistem *Artificial Intelligence* untuk memproses dan mengintegrasikan lebih dari satu jenis data, seperti teks, gambar, video, dan audio, dalam satu kerangka pemahaman terpadu. Pendekatan ini memungkinkan model membangun representasi yang lebih kaya dibandingkan pemrosesan satu modalitas secara terpisah. Penelitian [16] menunjukkan bahwa penggabungan representasi visual dan bahasa mampu meningkatkan kualitas *reasoning* model, khususnya pada tugas-tugas yang melibatkan *product discovery* dan interpretasi visual yang kompleks. Lebih lanjut, penelitian [17] menegaskan peran krusial *multimodal learning* dalam domain medis, di mana integrasi data visual endoskopi dengan teks klinis menghasilkan penjelasan yang lebih komprehensif serta mendukung pengambilan keputusan klinis secara lebih akurat. Temuan ini menunjukkan bahwa pemahaman berbasis multimodal tidak hanya meningkatkan akurasi model, tetapi juga memperkaya konteks penalaran yang dihasilkan.

Konsep *multimodal learning* tersebut menjadi landasan utama dalam penelitian ini, khususnya dalam pengembangan chatbot berbasis *Vision AI*. Dengan memproses informasi visual dan tekstual secara terpadu dalam satu *pipeline*, chatbot mampu menghasilkan jawaban yang lebih presisi dan kontekstual. Pendekatan ini memastikan bahwa pemahaman sistem tidak bergantung pada satu jenis data semata, melainkan merupakan hasil integrasi menyeluruh antar modalitas, sehingga respons yang dihasilkan relevan, informatif, dan sesuai dengan kebutuhan pembelajaran.

2.2.3 Retrieval Augmented Generation (RAG)



Gambar 2.1 Ilustrasi Retrieval Augmented Generation

Retrieval-Augmented Generation (RAG) merupakan pendekatan yang mengombinasikan proses *information retrieval* dengan kemampuan generatif dari *language model*. Dalam mekanisme ini, sistem terlebih dahulu melakukan pencarian terhadap dokumen atau metadata yang paling relevan, kemudian menggunakan hasil *retrieval* tersebut sebagai konteks eksplisit dalam proses generasi jawaban. Pendekatan ini memungkinkan model menghasilkan respons yang lebih akurat dan terkontrol, karena tidak sepenuhnya bergantung pada pengetahuan internal model.Penelitian [22] melalui *RoRA-VLM* menunjukkan bahwa integrasi *retrieval* mampu meningkatkan kejelasan *vision–language reasoning*, khususnya dalam menafsirkan konten visual. Temuan tersebut diperkuat oleh penelitian [23] yang membuktikan bahwa mekanisme *retrieval* meningkatkan akurasi *Visual Question Answering* (VQA) dengan menyediakan *grounding information* yang eksplisit kepada model. Selain itu, penelitian multimodal seperti *RAVEN* [24] dan *GraphRAG* [25] menegaskan pentingnya *retrieval* dalam memahami hubungan antar-gambar serta merepresentasikan struktur pengetahuan yang lebih kompleks dan terhubung. Konsep-konsep tersebut menjadi landasan utama dalam penelitian ini untuk memastikan bahwa

chatbot multimodal yang dikembangkan tidak menghasilkan *hallucination* dan selalu merujuk pada metadata visual yang valid. Dengan mengintegrasikan RAG dalam *pipeline* multimodal, sistem mampu memanfaatkan informasi visual yang telah diekstraksi secara sistematis sebagai konteks *reasoning*, sehingga respons yang dihasilkan bersifat faktual, relevan, dan sesuai dengan tujuan pembelajaran.

Embedding dan *vector search* melengkapi fungsi *Retrieval-Augmented Generation* (RAG) dengan menyediakan mekanisme representasi semantik yang konsisten dan terukur. *Embedding* merupakan representasi numerik dari data, baik teks maupun gambar, yang dipetakan ke dalam ruang vektor berdimensi tinggi sehingga hubungan semantik antar data dapat dihitung secara matematis. Dalam penelitian ini, *Sentence Transformer* digunakan untuk mengonversi metadata visual—seperti *caption*, daftar objek, dan teks hasil *Optical Character Recognition* (OCR) ke dalam bentuk vektor, sehingga tingkat kemiripan antar informasi dapat dihitung menggunakan *cosine similarity*.

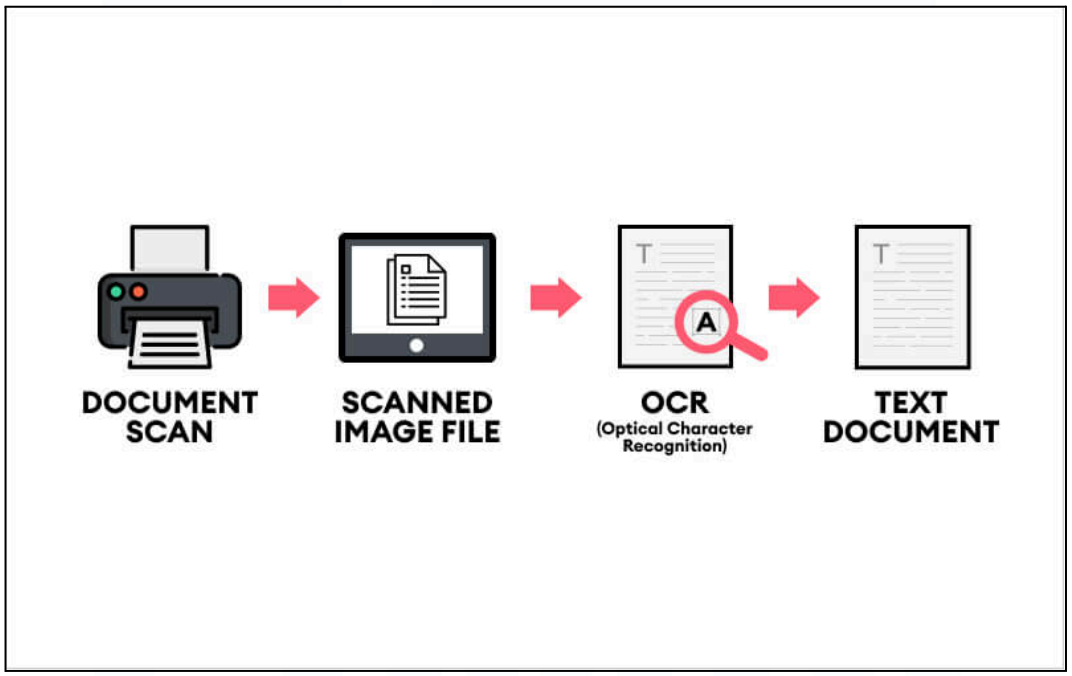
Penelitian [26] menegaskan bahwa *multimodal embedding* memungkinkan integrasi lintas sumber informasi, termasuk teks, gambar, tabel, hingga video, dalam satu *pipeline* terpadu. Pendekatan ini sangat relevan dalam sistem pembelajaran berbasis visual, karena memungkinkan pencarian informasi yang tidak hanya berbasis kata kunci, tetapi juga berdasarkan kesamaan makna. Pada penelitian ini, hasil *embedding* disimpan dalam *Supabase* yang mendukung *vector extension*, sehingga proses *similarity search* dapat dilakukan secara cepat dan efisien.

Supabase Vector berperan sebagai *knowledge store* utama yang menyatukan metadata visual dan representasi vektornya dalam satu struktur yang terorganisasi. Dengan dukungan *vector indexing*, sistem mampu melakukan pencarian kemiripan untuk menemukan gambar maupun metadata yang paling relevan dengan pertanyaan pengguna. Integrasi antara *Supabase Vector*, metadata visual, *Sentence Transformer*, dan *pipeline* RAG menjadikan basis data ini sebagai pusat

pengetahuan yang mendukung proses *reasoning* chatbot multimodal secara konsisten.

Konsep ini sangat sesuai untuk implementasi pada konteks pembelajaran di desa wisata atau sekolah dasar, karena memungkinkan pembangunan sistem AI yang ringan, modular, dan tetap mudah dikembangkan sesuai kebutuhan pembelajaran di masa mendatang.

2.2.4 Optical Character Recognition (OCR)

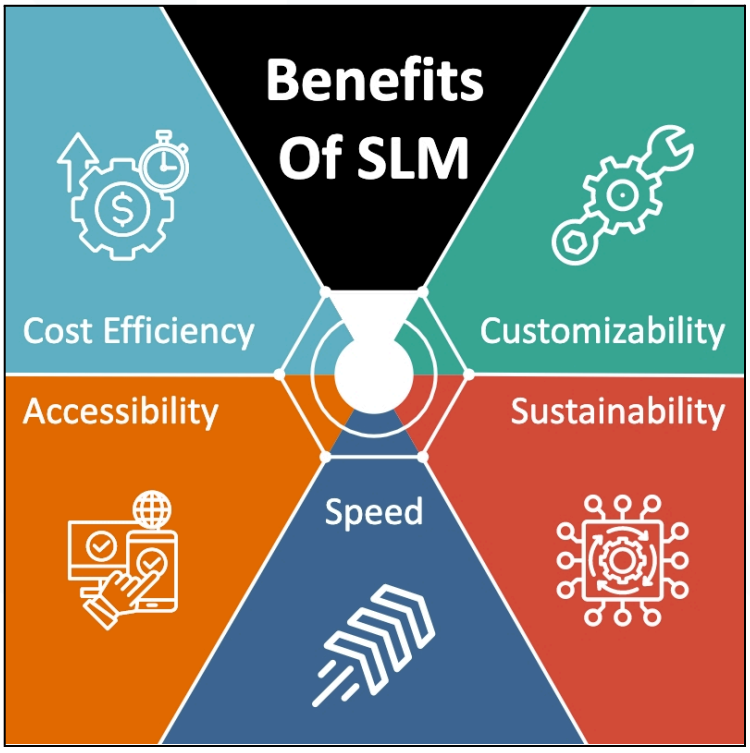


Gambar 2.3 Optical Character Recognition (OCR)

Optical Character Recognition (OCR) merupakan metode untuk mengekstraksi teks yang terdapat di dalam gambar dan mengubahnya menjadi representasi tekstual yang dapat diproses oleh sistem *Artificial Intelligence*. Meskipun tidak seluruh penelitian terdahulu secara eksplisit berfokus pada OCR, teknologi ini memiliki peran penting dalam sistem multimodal, khususnya ketika informasi visual mengandung elemen tekstual yang bermakna.

OCR memungkinkan teks yang muncul pada gambar seperti label, tanda, nama objek, atau tulisan sederhana untuk diekstraksi dan dijadikan *metadata* tambahan dalam proses *reasoning*. Dengan demikian, pemahaman sistem tidak hanya bergantung pada deskripsi visual yang dihasilkan oleh *Vision-Language Models*, tetapi juga mencakup informasi tekstual eksplisit yang tertanam di dalam gambar. Dalam penelitian ini, OCR diintegrasikan dengan *caption* yang dihasilkan oleh *Vision-Language Models* serta *vector embedding* untuk membentuk representasi visual yang lebih lengkap. Integrasi ini memungkinkan chatbot multimodal memiliki pemahaman yang lebih komprehensif terhadap gambar edukatif, terutama pada konteks pembelajaran dasar yang sering menggunakan visual dengan label atau tulisan sederhana. Pendekatan ini memperkuat akurasi *reasoning* dan relevansi respons chatbot terhadap pertanyaan pengguna.

2.2.5 Small Language Model



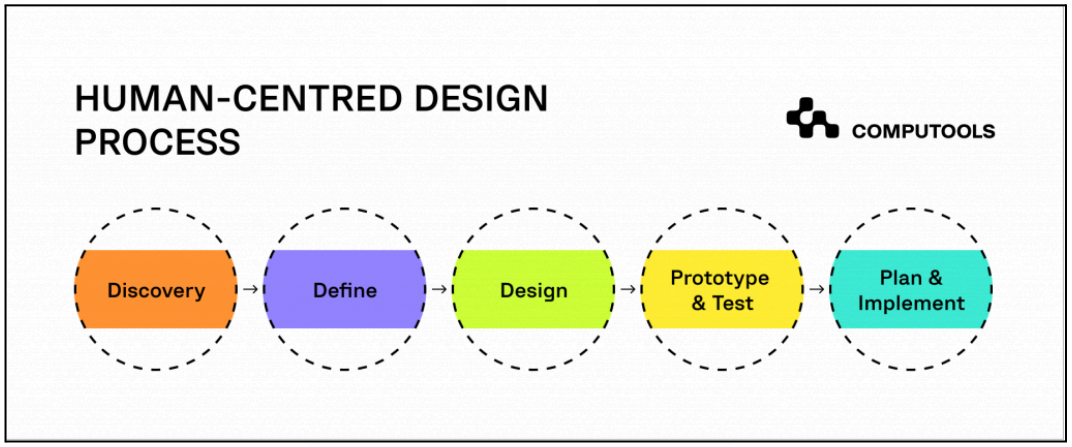
Gambar 2.4 Manfaat Small Language Models

Small Language Model adalah model bahasa berukuran kecil hingga menengah (1–8B parameter) yang dirancang untuk efisiensi komputasi tanpa kehilangan

kualitas reasoning dasar. Penelitian-penelitian modern menunjukkan bahwa SLM dapat menghasilkan jawaban yang tetap relevan selama didukung oleh konteks yang kuat melalui RAG. Hal ini menjadi dasar pemilihan SLM dalam penelitian ini dan memastikan sistem dapat berjalan cepat, hemat memori, dan cocok untuk deployment tingkat desa atau sekolah.

2.3 Algoritma dan Framework

2.3.1 Human Centric Design (HCD)



Gambar 2.5 Proses Human Centered Design

Penelitian ini menggunakan pendekatan *Human-Centered Design (HCD)* sebagai prinsip dasar pengembangan sistem, di mana desain diarahkan berdasarkan kebutuhan guru dan siswa sebagai pengguna utama. Pendekatan ini sejalan dengan rekomendasi penelitian *VizChat* yang menekankan bahwa integrasi AI dalam lingkungan belajar harus mempertimbangkan kenyamanan pengguna, aksesibilitas dan kemudahan interpretasi informasi visual [6]. Selain itu, penelitian di bidang pendidikan multimodal menunjukkan bahwa sistem AI yang baik adalah sistem yang menyesuaikan kompleksitas output dengan kemampuan pengguna, bukan sebaliknya [11], [13]. Karena itu, seluruh fitur chatbot mulai dari deskripsi visual hingga jawaban dirancang agar ramah anak dan relevan bagi guru di lingkungan desa.

2.3.2 Vision–Language Processing Framework (BLIP + OCR)

Komponen vision dalam penelitian ini menggabungkan dua teknik utama: BLIP dan OCR. BLIP digunakan untuk mengekstraksi *caption* dan deskripsi visual dari gambar edukatif. Model ini terbukti efektif untuk menghasilkan representasi cross-modal yang akurat dalam berbagai penelitian multimodal sebelumnya [3], [18]. BLIP mampu mengenali objek, hubungan dan konteks visual sehingga metadata yang dihasilkan dapat dijadikan dasar reasoning oleh sistem RAG. OCR digunakan untuk mengidentifikasi teks yang muncul pada gambar, misalnya tulisan pada papan edukasi, label tanaman atau teks pada infografis. Keberadaan OCR penting karena beberapa gambar edukatif desa mengandung informasi teks yang perlu diambil sebagai bagian dari pengetahuan gambar. Penelitian multimodal menunjukkan bahwa penggabungan teks OCR dengan caption visual meningkatkan akurasi reasoning dan mengurangi risiko halusinasi [19].

Dengan menggabungkan *BLIP* dan *OCR,* penelitian ini memastikan bahwa metadata visual mencakup caption, daftar objek dan teks OCR sehingga pengetahuan gambar menjadi lebih komprehensif.

2.3.3 Multimodal Reasoning Framework (Embedding → RAG → SLM)

Framework multimodal dalam penelitian ini mengintegrasikan tiga komponen utama, yaitu *Sentence Transformer*, *Retrieval-Augmented Generation* (RAG), dan *Small Language Model* (SLM), ke dalam satu alur *reasoning* terpadu. Ketiga komponen tersebut bekerja secara berurutan untuk memastikan bahwa pemahaman visual yang diperoleh sistem dapat diterjemahkan menjadi jawaban yang faktual, terkontrol, dan tidak mengalami *hallucination*.

Tahap pertama dimulai dengan *Sentence Transformer*, yang berfungsi mengubah metadata visual—meliputi *caption* hasil BLIP, daftar objek visual, serta teks hasil *Optical Character Recognition* (OCR) menjadi representasi *embedding* dalam ruang vektor. Representasi ini menjadi fondasi utama dalam proses *similarity search*, yaitu mekanisme pencarian metadata yang paling relevan berdasarkan pertanyaan pengguna. Penelitian *LVLM-eHub* menegaskan bahwa kualitas

representasi *embedding* memiliki pengaruh signifikan terhadap performa *multimodal reasoning*, terutama ketika metadata visual digunakan sebagai konteks tambahan [2]. Temuan serupa juga ditunjukkan dalam studi *multimodal retrieval* seperti *RAVEN* dan *RoRA*, yang menekankan bahwa *embedding* yang kuat mampu meningkatkan efisiensi dan akurasi proses pencarian konteks visual [22], [24]. Pada penelitian ini, seluruh *embedding* disimpan dalam *Supabase* yang mendukung *vector indexing*, sehingga proses *retrieval* dapat dilakukan secara cepat dan presisi.

Tahap kedua adalah *Retrieval-Augmented Generation* (RAG), yang berperan untuk memastikan bahwa jawaban chatbot tidak dihasilkan berdasarkan tebakan model, melainkan berlandaskan metadata visual yang relevan dan tervalidasi. Berbagai penelitian, mulai dari studi awal RAG [15] hingga pengembangan framework seperti *VisRAG* [37] dan *Typed-RAG* [39], menunjukkan bahwa RAG mampu meningkatkan faktualitas, akurasi, dan keamanan sistem generatif. Dalam *pipeline* penelitian ini, RAG menjalankan tiga fungsi utama, yaitu sebagai *retriever* yang mengambil metadata visual paling relevan, sebagai *grounding mechanism* yang membatasi ruang jawaban agar tetap sesuai fakta gambar, serta sebagai *hallucination guard* untuk menurunkan risiko model mengarang konten yang tidak terdapat pada visual. Pendekatan ini menjadi sangat krusial dalam konteks pendidikan, di mana kebenaran informasi dan keamanan konten merupakan prioritas utama bagi guru dan siswa [21], [33].

Tahap ketiga adalah *Small Language Model* (SLM) yang berfungsi sebagai mesin *reasoning* untuk merangkai jawaban akhir. SLM dipilih karena memiliki keunggulan dalam hal efisiensi komputasi, kecepatan respons, serta keamanan penggunaan pada lingkungan lokal dengan keterbatasan infrastruktur. Penelitian terkait *teacher-centric SLM* dan model edukatif lokal menunjukkan bahwa SLM mampu memberikan performa *reasoning* yang kompetitif apabila didukung oleh konteks yang kuat melalui RAG [34], [35]. Selain itu, karya seperti *Multimodal-GPT* memperlihatkan bahwa model dengan jumlah parameter relatif kecil tetap mampu menghasilkan jawaban yang terarah dan konsisten selama

grounding visualnya memadai [18]. Dalam penelitian ini, SLM bertugas menyusun jawaban berdasarkan metadata visual hasil *retrieval*, sehingga output yang dihasilkan bersifat faktual, jelas, mudah dipahami, dan sesuai dengan tingkat literasi anak-anak di Desa Wisata Tigaraksa.

2.3.4 Integrated Multimodal Intelligence Pipeline

Framework akhir dalam penelitian ini merupakan integrasi seluruh komponen ke dalam satu alur sistem terpadu, yaitu *Vision → Embedding → Retrieval (RAG) → Reasoning (SLM) → User Interface (UI)*. Arsitektur ini mengadopsi pendekatan *modular pipeline* yang umum digunakan pada sistem multimodal modern, seperti *ImageRef-VL* [3] dan *LVLM-eHub* [2], yang menegaskan bahwa performa optimal tidak dicapai melalui satu model berukuran besar, melainkan melalui rangkaian komponen dengan fungsi spesifik yang saling melengkapi.

Pada tahap awal, *vision model* seperti *BLIP* dan *Optical Character Recognition* (OCR) mengekstraksi metadata visual dari gambar lokal, meliputi *caption*, objek, dan teks yang muncul pada gambar. Metadata ini kemudian diubah menjadi *embedding* vektor menggunakan *Sentence Transformer*, sehingga dapat direpresentasikan secara semantik dan dicari melalui mekanisme *similarity search*. Selanjutnya, modul *Retrieval-Augmented Generation* (RAG) bertugas memilih metadata visual yang paling relevan terhadap pertanyaan pengguna, memastikan proses *grounding* berlangsung secara akurat serta mencegah terjadinya *hallucination*.

Pada tahap *reasoning*, *Small Language Model* (SLM) memanfaatkan metadata hasil *retrieval* untuk menghasilkan jawaban yang faktual, aman, dan sesuai dengan tingkat pemahaman pengguna, khususnya siswa sekolah dasar. Seluruh proses tersebut kemudian disajikan melalui *User Interface* (UI) yang dirancang khusus untuk kebutuhan guru dan siswa, sehingga interaksi dengan sistem bersifat intuitif dan mendukung aktivitas pembelajaran.

Pendekatan integratif ini memastikan bahwa sistem mampu memahami dan menjelaskan gambar edukatif lokal secara konsisten, akurat, dan relevan. Dengan demikian, framework yang diusulkan selaras dengan tujuan penelitian, yaitu mendukung pembelajaran berbasis visual di Desa Wisata Tigaraksa melalui pemanfaatan chatbot multimodal yang aman dan kontekstual.

2.4 Software yang digunakan

2.4.1 Python

Python digunakan sebagai bahasa pemrograman utama dalam penelitian ini. Bahasa ini dipilih karena memiliki ekosistem pustaka yang sangat lengkap untuk *machine learning, computer vision, natural language processing (NLP),* dan integrasi API. Library seperti *Transformers*, *Sentence-Transformers*, *PyTesseract*, dan *Supabase-Py* memungkinkan implementasi seluruh pipeline mulai dari *vision processing (BLIP & OCR), embedding Sentence Transformer,* hingga integrasi dengan database Supabase. Python juga mendukung prototyping cepat dan mudah dipindahkan ke lingkungan produksi.

2.4.2 Visual Studio Code

Visual Studio Code digunakan sebagai editor utama untuk menulis dan mengelola kode. VS Code menawarkan fleksibilitas tinggi melalui berbagai ekstensi seperti *Python Linter, Jupyter Notebook Extension, dan Git Integration*. Fitur-fitur ini membantu mempercepat proses debugging, mengatur struktur proyek, serta mempermudah kerja kolaboratif dengan sistem kontrol versi GitHub. VS Code juga mendukung manajemen environment dan konfigurasi API, sehingga cocok untuk pengembangan aplikasi berbasis AI.

2.4.3 Google Collab

Google Collab digunakan sebagai platform eksperimen model karena menyediakan akses GPU gratis yang sangat membantu dalam menjalankan model vision dan embedding dengan lebih cepat. Notebook interaktif Colab

memudahkan proses eksplorasi, evaluasi performa, visualisasi hasil, dan penyimpanan catatan eksperimen. Selain itu, Colab terintegrasi langsung dengan *Google Drive* sehingga mempermudah pengelolaan dataset serta dokumentasi proses pelatihan dan pengujian.