Terbit *online* pada laman : http://teknosi.fti.unand.ac.id/

Jurnal Nasional Teknologi dan Sistem Informasi

| ISSN (Print) 2460-3465 | ISSN (Online) 2476-8812 |

Literature Review

Tinjauan Literatur Sistematis: Teknik Optical Character Recognition Dan Information Retrieval Pada Sistem Tanya Jawab

Sekar Arum Chrysanti[a], *Roni Habibi*[b,*]

[a,b]*Program Studi D4 Teknik Informatika, Universitas Logistik dan Bisnis Internasional, Jalan Sariasih No. 54 Bandung 40151, Indonesia*

INFORMASI ARTIKEL

Sejarah Artikel:
Diterima Redaksi: 10 Juni 2026
Revisi Akhir: 20 Agustus 2026
Diterbitkan *Online*: 28 Agustus 2026

KATA KUNCI

Sistem Tanya Jawab Berbasis Gambar,
Optical Character Recognition,
Information Retrieval,
Multimodal Learning,
Tinjauan Literatur Sistematis

KORESPONDENSI

E-mail:
roni.habibi@ulbi.ac.id

A B S T R A C T

Perkembangan Sistem Tanya Jawab Berbasis Gambar (*Image-based Question Answering*) mengalami peningkatan pesat seiring kemajuan teknologi *Optical Character Recognition* (OCR), *information retrieval*, dan kecerdasan buatan multimodal. Namun, integrasi OCR, mekanisme *retrieval*, dan *multimodal reasoning* masih menghadapi berbagai tantangan konseptual dan metodologis, sementara sintesis penelitian yang mengkaji keterkaitan ketiganya masih terbatas. Penelitian ini bertujuan memetakan perkembangan, tren, tantangan, dan arah penelitian masa depan terkait teknik berbasis OCR dan *information retrieval* pada sistem *Image-based Question Answering*. Metode yang digunakan adalah *Systematic Literature Review* (SLR) dengan mengacu pada pedoman PRISMA. Sebanyak 61 artikel terindeks dianalisis menggunakan kerangka TCCM (*Theory, Context, Characteristics, and Methodology*). Hasil penelitian menunjukkan bahwa *Multimodal Deep Learning*, arsitektur Transformer, dan *Attention Mechanism* menjadi fondasi utama pengembangan sistem QA modern. Selain itu, integrasi *Knowledge Graph*, *Large Language Model* (LLM), dan *Retrieval-Augmented Generation* (RAG) semakin dominan karena mampu meningkatkan kemampuan *reasoning* dan pemahaman kontekstual. Terdapat temuan yang mengindikasikan pergeseran paradigma dari pendekatan berbasis *feature engineering* menuju sistem *knowledge-driven multimodal reasoning*. Temuan ini memberikan dasar bagi pengembangan sistem QA berbasis gambar yang lebih adaptif, andal, dan kontekstual. Meskipun telah berkembang dengan pesat dan cepat, namun pengembangan IQA juga memiliki tantangan utama meliputi *semantic gap*, kompleksitas *multimodal fusion*, *hallucination*, keterbatasan dataset multibahasa, dan rendahnya interpretabilitas sistem.

1. PENDAHULUAN

Dalam satu dekade terakhir, perkembangan sistem *Image-based Question Answering* (Image-QA) menunjukkan dinamika yang sangat pesat, terutama sejak kemajuan teknologi *computer vision*, *natural language processing* (NLP), *Optical Character Recognition* (OCR), serta kemunculan *multimodal large language models* (MLLMs). Pada tahap awal perkembangannya, penelitian umumnya masih berfokus pada kemampuan dasar sistem dalam mengenali teks dan objek visual melalui pendekatan OCR konvensional untuk digitalisasi dokumen maupun ekstraksi informasi visual (Lee et al., 2024; Banerjee et al., 2022) [1], [2]. Namun, seiring meningkatnya kebutuhan terhadap sistem cerdas yang tidak hanya mampu "melihat", tetapi juga memahami konteks visual secara lebih mendalam, arah penelitian mulai bergeser menuju integrasi *knowledge graph*, *deep learning*, dan *retrieval-based reasoning* guna menghasilkan jawaban yang lebih relevan dan kontekstual (Liu et al., 2021; Li et al., 2022) [3], [4]. Hal yang menarik adalah bahwa transformasi tersebut tidak berlangsung secara linear, melainkan berkembang melalui berbagai eksperimen lintas domain yang memperlihatkan semakin kompleksnya kebutuhan sistem QA modern. Kehadiran *Retrieval-Augmented Generation* (RAG) dan MLLMs, misalnya, telah membuka kemungkinan baru bagi sistem untuk melakukan penalaran lintas modalitas secara lebih adaptif pada berbagai konteks, mulai dari kesehatan, warisan budaya, maritim, pendidikan, hingga industri (Xu et al., 2024; Xie et al., 2024) [5], [6]. Dalam konteks ini, OCR tidak lagi diposisikan sekadar

https://doi.org/10.25077/TEKNOSI.v12i2.2026.284-299

sebagai teknologi ekstraksi teks, tetapi mulai memainkan peran penting sebagai penghubung antara representasi visual dan proses reasoning berbasis pengetahuan.

Tren penelitian terkini juga memperlihatkan bahwa fokus pengembangan sistem QA berbasis gambar tidak lagi terbatas pada peningkatan akurasi pengenalan objek atau teks semata. Sebaliknya, perhatian peneliti kini mulai bergeser pada bagaimana sistem mampu melakukan penalaran semantik, memahami hubungan antarkonteks, serta mengintegrasikan sumber pengetahuan eksternal secara dinamis. Model berbasis *Transformer* dan *multimodal learning* bahkan telah mendominasi berbagai benchmark VQA dengan performa yang secara signifikan melampaui pendekatan tradisional (Kim et al., 2025) [7]. Namun demikian, di balik peningkatan performa tersebut, muncul tantangan baru yang tidak kalah kompleks, khususnya terkait *hallucination*, *multilingual bias*, keterbatasan reasoning, dan rendahnya transparansi jawaban pada domain spesifik.

Dalam kondisi tersebut, integrasi OCR dengan teknik *information retrieval* menjadi semakin relevan untuk dikaji. Banyak skenario QA berbasis gambar, khususnya pada domain dokumen, medis, maupun industri, menuntut sistem untuk melakukan ekstraksi teks visual, pencarian informasi yang relevan, serta penalaran berbasis konteks secara simultan. Oleh sebab itu, hubungan antara OCR, retrieval mechanism, dan multimodal reasoning menjadi salah satu isu penting yang mulai mendapat perhatian luas dalam penelitian mutakhir. Sayangnya, meskipun jumlah publikasi meningkat secara signifikan, sintesis komprehensif mengenai keterkaitan ketiga aspek tersebut masih relatif terbatas.

Meskipun sejumlah studi *Systematic Literature Review* (SLR) telah dilakukan sebelumnya, sebagian besar masih memiliki fokus yang cenderung parsial. Roy et al. (2023) [8], misalnya, lebih berfokus pada *Community Question Answering* (CQA), sedangkan Sun et al. (2023) [9] menitikberatkan pada klasifikasi pertanyaan di bidang geosains. Di sisi lain, Kim et al. (2025) [7] berhasil memetakan perkembangan VQA modern, tetapi pendekatan yang digunakan masih cenderung deskriptif dan belum mengeksplorasi keterkaitan antara ekstraksi teks visual, mekanisme retrieval, serta tantangan kontemporer seperti *hallucination* dan bias lintas bahasa.

Secara khusus, celah penelitian terlihat pada belum adanya SLR yang secara sistematis membahas evolusi integrasi OCR dan teknik *information retrieval* dalam sistem *Image-based Question Answering*. Padahal, perkembangan terbaru menunjukkan bahwa kombinasi OCR, retrieval mechanism, dan MLLMs mulai menjadi fondasi penting dalam pengembangan sistem QA multimodal yang lebih adaptif dan kontekstual. Selain itu, sebagian besar review terdahulu masih berfokus pada klasifikasi model atau domain aplikasi tertentu, sehingga belum mampu memberikan gambaran konseptual mengenai hubungan antara kompleksitas modalitas, integrasi pengetahuan, dan konteks implementasi sistem.

Berdasarkan gap tersebut, penelitian ini menawarkan kebaruan melalui sintesis sistematis yang memetakan evolusi integrasi OCR dan teknik *information retrieval* pada sistem *Image-based Question Answering*. Berbeda dengan SLR sebelumnya yang cenderung terfragmentasi berdasarkan domain atau pendekatan model tertentu, penelitian ini mengusulkan kerangka klasifikasi berbasis tiga dimensi, yaitu: (1) *knowledge integration* yang mencakup integrasi pengetahuan internal dan eksternal, (2) *modality complexity* yang meliputi unimodal, multimodal, dan *cross-modal reasoning*, serta (3) *deployment context* yang membedakan domain generik dan domain spesifik. Lebih lanjut, penelitian ini juga berupaya mengidentifikasi bagaimana integrasi OCR, retrieval mechanism, dan MLLMs berkontribusi terhadap peningkatan kemampuan reasoning sekaligus mengurangi *knowledge hallucination* pada sistem QA berbasis gambar.

Kontribusi ilmiah penelitian ini meliputi beberapa aspek utama. Pertama, penelitian ini menyusun peta perkembangan penelitian OCR-based dan retrieval-based QA selama periode 2015–2025. Kedua, penelitian ini mengidentifikasi tren metodologi, arsitektur model, dataset, serta metrik evaluasi yang paling dominan digunakan dalam pengembangan sistem QA berbasis gambar. Ketiga, penelitian ini mengembangkan taksonomi konseptual yang menghubungkan OCR, retrieval mechanism, dan multimodal reasoning dalam satu kerangka analisis yang lebih integratif. Terakhir, penelitian ini merumuskan agenda penelitian masa depan yang berkaitan dengan efisiensi model, adaptasi lintas domain, multilingual QA, dan mitigasi *hallucination* pada sistem multimodal.

Berdasarkan tujuan tersebut, penelitian ini merumuskan beberapa pertanyaan penelitian, yaitu: (1) Bagaimana tren perkembangan publikasi penelitian terkait OCR-based dan retrieval-based techniques pada *Image-based Question Answering* selama periode 2015–2025? (2) Metode, arsitektur model, dan teknik retrieval apa yang paling dominan digunakan dalam sistem QA berbasis gambar? (3) Bagaimana peran OCR, *knowledge retrieval*, dan multimodal reasoning dalam meningkatkan performa sistem *Image-based Question Answering*?

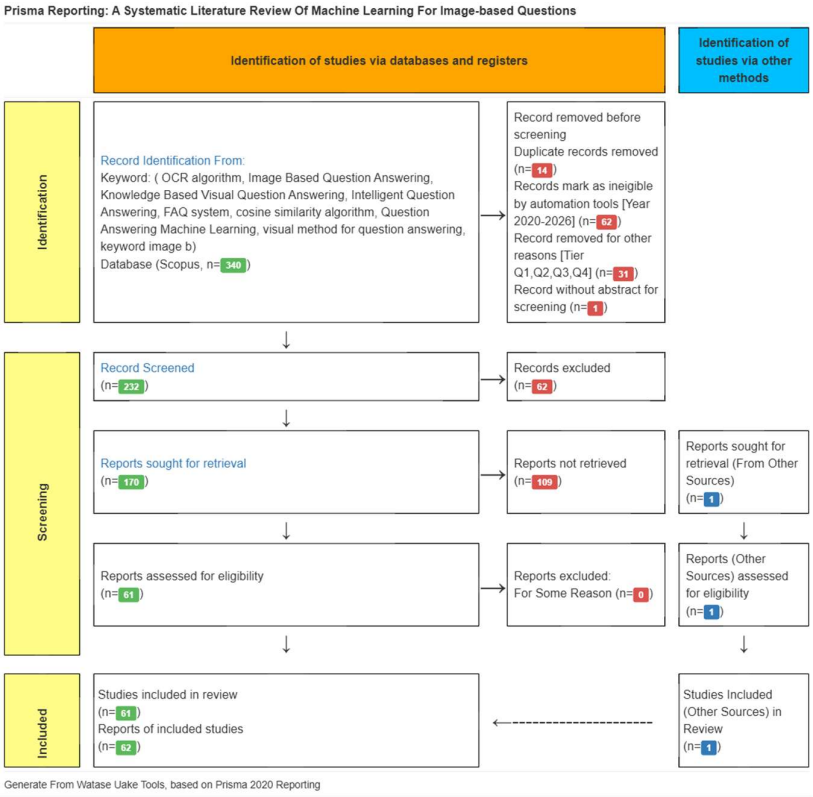
Untuk memberikan alur pembahasan yang sistematis, artikel ini disusun ke dalam beberapa bagian utama. Bagian pertama menjelaskan latar belakang, urgensi penelitian, serta identifikasi gap penelitian. Bagian kedua menguraikan metode SLR yang digunakan berdasarkan pedoman PRISMA. Selanjutnya, bagian ketiga menyajikan hasil analisis literatur dan pemetaan tren penelitian. Bagian keempat membahas sintesis temuan, implikasi teoretis maupun praktis, serta peluang penelitian di masa depan. Terakhir, bagian kelima menyajikan kesimpulan penelitian dan rekomendasi agenda riset lanjutan.

2. METODE

Penelitian ini menggunakan pendekatan *Systematic Literature Review* (SLR) untuk mengidentifikasi, mengevaluasi, dan mensintesis secara sistematis berbagai penelitian terkait penerapan OCR dan teknik *information retrieval* pada sistem *Image-based Question Answering*. Proses tinjauan dilakukan dengan mengacu pada pedoman *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA) yang diperkenalkan oleh Moher et al. (2009) [10]. Penggunaan pedoman PRISMA dipilih karena mampu meningkatkan transparansi, konsistensi, serta reproduktibilitas proses seleksi artikel dalam penelitian berbasis tinjauan sistematis, sebagaimana

direkomendasikan oleh Panic et al. (2013) [11], Siddaway et al. (2019) [12], dan ter Huurne et al. (2017) [13]. Alur lengkap tahapan PRISMA yang digunakan dalam penelitian ini dapat dilihat pada Gambar 1.

Proses penelitian diawali pada tahap *identification* dengan melakukan penelusuran artikel melalui basis data Scopus. Pemilihan Scopus didasarkan pada pertimbangan bahwa basis data tersebut memiliki cakupan literatur ilmiah yang luas dan menerapkan proses indeksasi yang ketat sehingga kualitas artikel yang terpublikasi relatif lebih terjamin melalui mekanisme *peer-review* yang kredibel (Lasda Bergman, 2012; Rocha et al., 2020) [14], [15]. Selain itu, penggunaan Scopus juga bertujuan untuk meminimalkan potensi duplikasi, inkonsistensi metadata, serta keberadaan artikel dari jurnal predator yang sering ditemukan pada basis data terbuka seperti Google Scholar (Hariningsih et al., 2024) [16]. Untuk memperluas cakupan identifikasi, penelitian ini juga memanfaatkan Sistem Watase Uake (Wahyudi, 2024) [16] sebagai sumber tambahan dalam menemukan artikel yang relevan.



Gambar 1. Laporan PRISMA

Penelusuran literatur dilakukan menggunakan beberapa kombinasi kata kunci yang berkaitan langsung dengan topik penelitian, antara lain *OCR algorithm*, *Image Based Question Answering*, *Knowledge Based Visual Question Answering*, *Intelligent Question Answering*, *FAQ system*, *cosine similarity algorithm*, *Question Answering Machine Learning*, *visual method for question answering*, *keyword image based*, dan *Multimodal Visual Question Answering*. Penggunaan berbagai kata kunci tersebut bertujuan untuk memperluas cakupan pencarian sekaligus memastikan bahwa artikel yang diperoleh memiliki keterkaitan dengan aspek OCR, retrieval mechanism, maupun multimodal reasoning pada sistem QA berbasis gambar.

Hasil penelusuran awal pada tahap *identification* menghasilkan sebanyak 340 artikel. Sebelum memasuki tahap berikutnya, dilakukan proses pembersihan data (*data cleaning*) untuk menghilangkan artikel yang tidak memenuhi kriteria awal penelitian. Sebanyak 14 artikel dihapus karena teridentifikasi sebagai duplikasi, 62 artikel dieliminasi karena berada di luar rentang tahun publikasi 2020–2026, 31 artikel dikeluarkan karena tidak memenuhi kriteria kualitas jurnal berdasarkan kuartil yang ditetapkan, dan 1 artikel dihapus karena tidak memiliki abstrak. Dengan demikian, total artikel yang dieliminasi pada tahap awal sebanyak 108 artikel sehingga tersisa 232 artikel yang dilanjutkan ke tahap *screening*.

Pada tahap *screening*, peneliti melakukan pemeriksaan terhadap judul dan abstrak dari 232 artikel yang telah lolos proses identifikasi awal. Tahap ini bertujuan untuk memastikan kesesuaian topik artikel dengan fokus penelitian mengenai OCR-based dan retrieval-based techniques pada *Image-baseQuestion Answering*. Dari hasil proses screening, sebanyak 62 artikel dikeluarkan karena tidak memiliki relevansi yang memadai dengan topik

penelitian. Selanjutnya, sebanyak 170 artikel yang dianggap potensial dicari versi *full-text*-nya untuk dilakukan evaluasi lebih lanjut. Namun, terdapat 109 artikel yang tidak berhasil diperoleh karena keterbatasan akses atau ketersediaan dokumen. Di samping itu, penelitian ini juga menambahkan 1 artikel dari sumber lain melalui Sistem Watase Uake sehingga total artikel yang memasuki tahap evaluasi kelayakan berjumlah 62 artikel.

Tahap berikutnya adalah *eligibility*, yaitu proses evaluasi secara menyeluruh terhadap artikel *full-text* yang berhasil diperoleh. Pada tahap ini, setiap artikel dianalisis berdasarkan kesesuaian topik, relevansi metodologi, kontribusi penelitian, serta keterkaitannya dengan fokus kajian OCR, *information retrieval*, dan *Image-based Question Answering*. Berbeda dengan tahap screening yang hanya berfokus pada judul dan abstrak, tahap eligibility memungkinkan peneliti melakukan penilaian yang lebih mendalam terhadap substansi penelitian. Hasil evaluasi menunjukkan bahwa seluruh artikel yang berhasil melewati tahap ini memenuhi kriteria kelayakan sehingga tidak terdapat artikel yang dieliminasi lebih lanjut.

Tahap akhir dalam protokol PRISMA adalah *inclusion*, yaitu penetapan artikel yang digunakan sebagai sumber utama analisis dalam penelitian ini. Berdasarkan keseluruhan proses seleksi, sebanyak 62 artikel dinyatakan memenuhi kriteria inklusi, yang terdiri atas 61 artikel hasil pencarian melalui Scopus dan 1 artikel tambahan dari sumber lain. Artikel-artikel tersebut kemudian dianalisis menggunakan pendekatan analisis tematik untuk mengidentifikasi pola penelitian, tren metodologi, perkembangan teknologi, serta tantangan yang muncul dalam pengembangan sistem QA berbasis gambar.

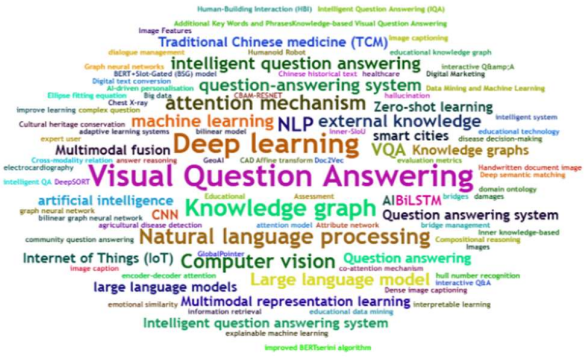
Proses analisis dilakukan dengan membaca dan mengevaluasi setiap artikel secara mendalam untuk memperoleh sintesis yang komprehensif terkait perkembangan OCR, retrieval mechanism, multimodal learning, serta integrasi *large language models* pada sistem QA modern. Selain itu, penelitian ini juga memanfaatkan Sistem Watase Uake (Wahyudi, 2024) [16] untuk mendukung proses pengelompokan tema dan validasi hasil analisis. Melalui pendekatan tersebut, penelitian ini tidak hanya memetakan tren publikasi dan perkembangan teknologi, tetapi juga mengidentifikasi research gap, tantangan implementasi, serta arah pengembangan penelitian di masa depan. Dengan demikian, seluruh tahapan penelitian, mulai dari *identification* hingga *inclusion*, dirancang untuk memastikan bahwa artikel yang dianalisis memiliki tingkat relevansi, validitas, dan kualitas ilmiah yang memadai sehingga mampu menghasilkan sintesis literatur yang lebih kredibel dan sistematis.

Untuk menjaga konsistensi proses seleksi artikel, penelitian ini juga menerapkan kriteria *inclusion* dan *exclusion* yang digunakan sebagai acuan dalam menentukan kelayakan setiap studi yang dianalisis. Kriteria tersebut disusun berdasarkan relevansi topik, kualitas publikasi, cakupan metodologi, serta kesesuaian artikel dengan fokus penelitian mengenai OCR dan *information retrieval* pada sistem *Image-based Question Answering*.

3. HASIL DAN PEMBAHASAN

Penelitian ini menyajikan hasil sintesis literatur dari 62 artikel yang telah lolos tahap *inclusion* berdasarkan protokol PRISMA. Analisis difokuskan pada perkembangan penelitian OCR-based dan *information retrieval techniques* pada sistem *Image-based Question Answering* (Image-QA), dengan menekankan tren publikasi, distribusi geografis penelitian, dominasi metode dan algoritma, karakteristik modalitas input, hingga tantangan utama yang dihadapi pada pengembangan sistem multimodal modern. Untuk menjaga koherensi dengan rumusan masalah dan novelty penelitian, pembahasan tidak hanya berfokus pada pemetaan statistik literatur, tetapi juga mengkaji pergeseran paradigma penelitian dari pendekatan berbasis *feature engineering* menuju integrasi *multimodal reasoning*, *knowledge graph*, dan *large language models* (LLMs). Seluruh visualisasi hasil analisis disajikan secara berurutan melalui Gambar 2 hingga Gambar 12, yang mencakup *wordcloud metadata*, klasifikasi negara penelitian, tren publikasi dan sitasi, klasifikasi teori, pendekatan metodologis, teknik analisis, modalitas input, metrik evaluasi, serta tantangan penelitian yang dominan.

Gambar 2 menampilkan *wordcloud metadata* yang menggambarkan distribusi kata kunci dominan dalam literatur yang dianalisis. Visualisasi ini menunjukkan bahwa istilah *Visual Question Answering (VQA)* menjadi tema paling sentral, diikuti oleh *Knowledge-based VQA*, *Medical VQA*, *Deep Learning*, *Attention Mechanism*, dan *Multimodal Fusion*. Selain itu, istilah seperti *Knowledge Graph*, *Large Language Model (LLM)*, dan *Retrieval-Augmented Generation (RAG)* mulai muncul secara signifikan pada publikasi setelah tahun 2023. Visualisasi tersebut mengindikasikan bahwa penelitian terkini tidak lagi hanya berfokus pada pengenalan objek visual, tetapi telah berkembang menuju integrasi pengetahuan eksternal dan penalaran multimodal berbasis konteks.



Gambar 2. Wordcloud Metadata

Korelasi antarkata kunci menunjukkan adanya transisi paradigma dari pendekatan *feature engineering* tradisional menuju model multimodal berbasis representasi end-to-end. Kemunculan kata kunci seperti *semantic reasoning*, *cross-modal attention*, dan *knowledge integration* mengindikasikan bahwa komunitas riset semakin menekankan pentingnya pemahaman semantik dibandingkan sekadar klasifikasi visual. Di sisi lain, munculnya domain spesifik seperti *Traditional Chinese Medicine (TCM)*, *smart agriculture*, dan *medical imaging* menunjukkan bahwa

 https://doi.org/10.25077/TEKNOSI.v12i2.2026.284-299

penelitian VQA kini semakin berorientasi pada aplikasi kontekstual dan domain driven.

Analisis ini juga menunjukkan bahwa integrasi OCR dengan retrieval mechanism menjadi salah satu tema yang berkembang pesat, terutama pada penelitian yang menangani dokumen visual, tangkapan layar (*screenshots*), dan error-code retrieval systems. Kemunculan istilah seperti *cosine similarity*, *OCR algorithm*, dan *FAQ retrieval* mengindikasikan bahwa sistem QA berbasis gambar tidak lagi hanya mengandalkan pemahaman visual, tetapi juga kemampuan retrieval terhadap informasi tekstual yang diekstraksi dari gambar. Hal ini memperkuat novelty penelitian ini yang menempatkan OCR dan *information retrieval* sebagai komponen integral dalam evolusi sistem QA multimodal.

Analisis pada Gambar 3 menunjukkan bahwa China merupakan negara dengan kontribusi penelitian terbesar dalam bidang OCR-based dan *Image-based Question Answering*. Dominasi tersebut terlihat dari tingginya jumlah publikasi yang membahas VQA medis, *knowledge graph*, multimodal reasoning, dan integrasi LLM. Penelitian oleh Liu et al. (2021) [3], Xu et al. (2024) [5], Chunfang et al. (2025) [17], dan Wang et al. (2026) [18] memperlihatkan bahwa China tidak hanya unggul secara kuantitas publikasi, tetapi juga memiliki pengaruh sitasi yang cukup tinggi.

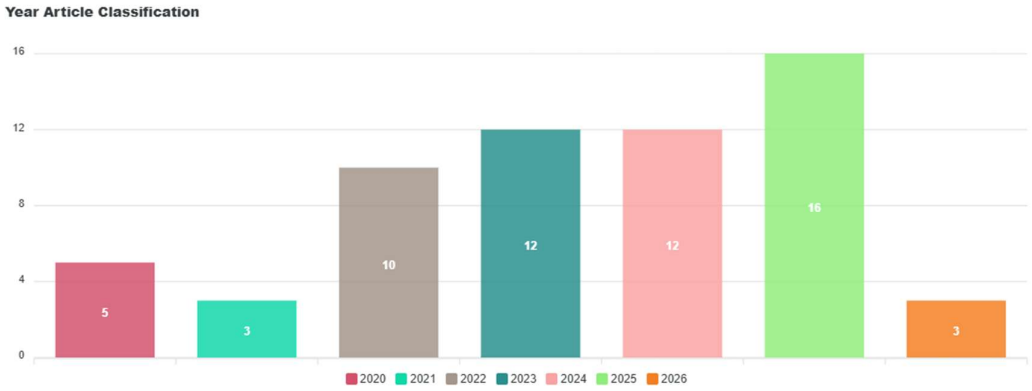
Selain dominasi kuantitatif, publikasi dari China juga menunjukkan dampak ilmiah yang tinggi. Studi Xu et al. (2024) memperoleh jumlah sitasi tertinggi dengan integrasi *Knowledge Graph* dan RAG untuk pelestarian budaya, sementara Liu et al. (2021) menjadi salah satu publikasi paling berpengaruh dalam pengembangan KG-based QA. Kondisi ini mengindikasikan bahwa penelitian dari Asia tidak hanya produktif secara jumlah, tetapi juga menjadi referensi utama dalam perkembangan metodologi QA berbasis multimodal.

Gambar 3. Klasifikasi Negara Artikel

Di sisi lain, negara-negara seperti Amerika Serikat, Korea Selatan, Jepang, Inggris, dan Spanyol menunjukkan kontribusi yang lebih sedikit secara kuantitatif, tetapi cenderung menghasilkan penelitian dengan pendekatan metodologis yang lebih eksploratif. Misalnya, Kim et al. (2025) [7] berfokus pada survei evolusi VQA modern berbasis Transformer dan LLM, sedangkan Salaberria et al. (2023) mengeksplorasi integrasi *knowledge-based reasoning* pada multimodal QA [19].

Temuan ini memiliki implikasi strategis terhadap arah perkembangan penelitian masa depan. Dominasi China sebagian besar didorong oleh ketersediaan data domain-spesifik, investasi besar dalam AI, dan fokus pada solusi kontekstual berbasis kebutuhan domestik. Namun, kondisi ini juga memunculkan tantangan terkait generalisasi model lintas budaya dan lintas bahasa. Mayoritas dataset dan model yang dikembangkan masih berorientasi pada konteks lokal, sehingga berpotensi menghasilkan bias semantik ketika diterapkan pada domain global. Oleh karena itu, pengembangan dataset multibahasa dan framework multimodal lintas budaya menjadi agenda penelitian yang semakin mendesak.

Gambar 4 menunjukkan tren publikasi dan sitasi penelitian dari tahun 2020 hingga 2026. Hasil analisis memperlihatkan peningkatan signifikan jumlah publikasi sejak tahun 2022 dan mencapai puncaknya pada tahun 2025. Lonjakan ini menunjukkan bahwa penelitian mengenai *image-based question answering* sedang berada pada fase ekspansi yang sangat cepat, terutama setelah kemunculan multimodal LLM dan RAG-based systems.



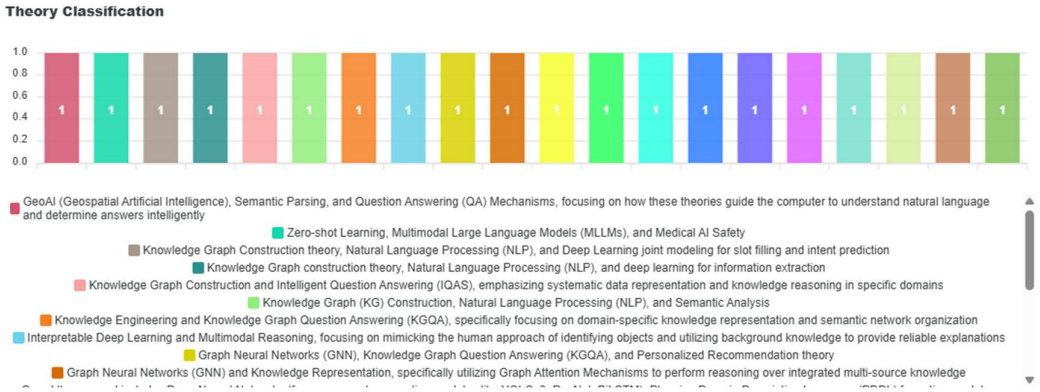
Gambar 4. Tahun Publikasi Artikel

Meskipun jumlah publikasi meningkat tajam, distribusi sitasi menunjukkan pola yang berbeda. Artikel-artikel pada tahun 2021 dan 2022 memiliki jumlah sitasi yang lebih tinggi dibandingkan publikasi terbaru. Hal ini menunjukkan bahwa penelitian awal yang memperkenalkan integrasi *knowledge graph*, Transformer, dan multimodal reasoning memiliki dampak fundamental terhadap perkembangan bidang ini. Sebaliknya, artikel tahun 2024–2026 masih relatif baru sehingga belum memiliki cukup waktu untuk membangun pengaruh sitasi. Fenomena ini memperlihatkan adanya transisi dari penelitian foundational menuju penelitian aplikatif.

Pada fase awal, fokus penelitian berada pada pengembangan model dasar VQA dan multimodal learning. Namun, sejak 2023 terjadi pergeseran menuju integrasi pengetahuan eksternal, *zero-shot reasoning*, dan domain-specific deployment seperti medis, pendidikan, dan infrastruktur cerdas. Pergeseran tersebut memperlihatkan bahwa tantangan penelitian tidak lagi hanya terkait akurasi klasifikasi, tetapi juga kemampuan reasoning, interpretability, dan trustworthiness.

Tren sitasi juga menunjukkan bahwa penelitian berbasis *knowledge integration* cenderung memiliki pengaruh ilmiah lebih tinggi dibandingkan pendekatan murni berbasis CNN atau RNN. Hal ini menandakan bahwa komunitas ilmiah mulai menganggap kemampuan reasoning sebagai aspek utama dalam evaluasi sistem QA modern. Dengan demikian, perkembangan masa depan kemungkinan akan bergerak menuju sistem yang lebih adaptif, explainable, dan mampu melakukan penalaran lintas domain secara kontekstual.

Gambar 5 menunjukkan klasifikasi teori yang mendominasi penelitian *image-based question answering*. Hasil analisis memperlihatkan bahwa *Multimodal Deep Learning* dan *Natural Language Processing* menjadi fondasi teoretis utama dalam mayoritas penelitian. Dominasi ini mencerminkan kebutuhan untuk mengintegrasikan representasi visual dan linguistik secara simultan dalam sistem QA berbasis gambar.



Gambar 5. Klasifikasi Teori

Secara khusus, teori berbasis Transformer dan *Attention Mechanism* menjadi pendekatan paling berpengaruh dalam literatur terkini. Evolusi dari CNN-RNN menuju Transformer menunjukkan perubahan mendasar dalam cara model memproses relasi multimodal. Pendekatan attention memungkinkan model memusatkan fokus pada fitur visual dan tekstual yang relevan secara dinamis, sehingga meningkatkan kemampuan reasoning dan contextual understanding. Studi Kim et al. (2025)

 https://doi.org/10.25077/TEKNOSI.v12i2.2026.284-299

memperlihatkan bahwa multimodal LLM berbasis Transformer telah melampaui performa model tradisional pada berbagai benchmark VQA [7].

Selain Transformer, *Knowledge Graph* (KG) muncul sebagai teori penting dalam domain-domain yang membutuhkan reasoning berbasis pengetahuan eksternal. Integrasi KG banyak digunakan pada bidang medis, warisan budaya, dan maritim karena mampu menyediakan representasi relasional yang lebih explainable dibandingkan model black-box berbasis deep learning. Integrasi KG dengan RAG, sebagaimana dilakukan Xu et al. (2024), menunjukkan arah baru pengembangan sistem QA yang lebih kontekstual dan dapat dipercaya. Penelitian oleh Xu et al. (2024), Chunfang et al. (2025), dan Jiang dan Meng (2023) menunjukkan bahwa integrasi antara KG dan multimodal reasoning mampu meningkatkan kualitas jawaban sekaligus mengurangi kesalahan inferensi [5], [17], [20].

Namun demikian, analisis ini juga mengungkap adanya kesenjangan metodologis. Sebagian besar penelitian masih berorientasi pada peningkatan performa teknis tanpa memberikan perhatian memadai terhadap aspek interpretability dan fairness. Model berbasis Transformer memang unggul dalam akurasi, tetapi sering mengalami hallucination ketika menghadapi pertanyaan yang membutuhkan pengetahuan domain mendalam. Dengan demikian, integrasi teori reasoning berbasis pengetahuan eksternal menjadi kebutuhan yang semakin penting.

Tabel 1. Evaluasi Metode Penelitian

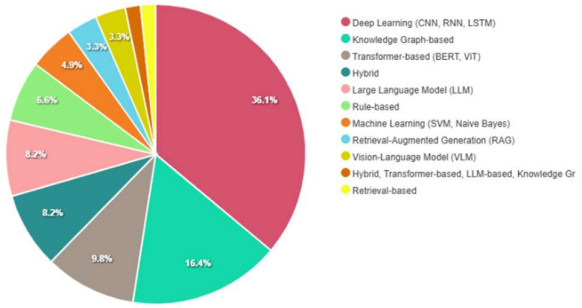
No	Fungsi	Jumlah	Evaluasi	Authors
1	Visual Question Answering (VQA)	24	Akurasi 68,62% Presisi 79,57% F1 -	[21], [22], [23], [24], [25], [4], [26], [27], [28], [20], [29], [19], [30], [31], [32], [33], [34], [7], [35], [36], [37], [38], [39], [40]
2	FAQ Retrieval	18	Akurasi 88,41% Presisi 92,38% F1 92,26%	[41], [42], [43], [44], [45], [46], [47], [9], [29], [48], [49], [6], [50], [5], [17], [51], [52], [53]
3	Medical Image QA	8	Akurasi 71,21% Presisi 56,80% F1 38,43%	[54], [55], [56], [57], [58], [59], [60], [18]
4	Document Image QA (OCR-based)	3	Akurasi 86,12% Presisi 78,00% F1 93,48%	[3], [2], [1]
5	Multi-modal QA	3	Akurasi 84,31% Presisi 93,65% F1 84,98%	[61], [62], [63]
6	Community QA (CQA)	2	Akurasiisi 74,98% Presisi 73,38% F1 75,21%	[64], [8]
7	Image-based QA (error Code / Screen-shot)	1	Akurasi 93,38 Presisi 93,29 F1 93,37	[65]

Berdasarkan analisis terhadap tabel 1 yang disajikan, menunjukkan klasifikasi jenis *Question Answering* (QA) berbasis gambar yang menjadi fokus penelitian. Dari 60 studi yang tercantum, terdapat lima kategori utama, yaitu: *Visual Question Answering* (VQA), *Medical Image QA*, *FAQ Retrieval*, *Document Image QA (OCR-based)*, *Multi-modal QA*, *Community QA (CQA)*, dan *Image-based QA (error code / screenshot)*. Kategori yang paling dominan adalah Visual Question Answering (VQA) yang muncul sebanyak 24 kali (41,7% dari total studi). Dominasi ini menunjukkan bahwa VQA merupakan bidang yang paling matang dan banyak diminati, karena menantang kemampuan sistem untuk mengintegrasikan pemrosesan gambar dan bahasa alami secara mendalam, seperti yang terlihat pada penelitian oleh Chen et al. (2020), Wang et al. (2023), dan Kim et al. (2025) [22], [28], [7]. Kategori terbanyak kedua adalah FAQ Retrieval dengan 18 studi (28,3%), yang banyak diterapkan pada domain spesifik seperti pendidikan, teknik, dan kedokteran tradisional. Sementara itu, Medical Image QA menempati posisi ketiga dengan 8 studi (15%), yang relevan dengan tekanan diagnostik di rumah sakit.

Implikasi dari tren ini adalah adanya pergeseran fokus dari sistem QA berbasis teks semata menuju sistem yang mampu menangani multimodalitas secara cerdas. Hampir semua studi pada kategori VQA dan Medical Image QA menggunakan metode *Deep Learning* berbasis Transformer dan model bahasa besar (LLM) seperti yang dilakukan oleh Salaberria et al. (2023) dan Lee et al. (2024). Penggunaan *Attention Mecha-*

nism dan *Word Embedding* (seperti GloVe dan Word2Vec) menjadi fondasi teknis yang sangat dominan, hadir di lebih dari 40 studi. Ini menunjukkan bahwa komunitas riset sepakat bahwa kemampuan untuk memusatkan perhatian pada fitur gambar yang relevan dan menghubungkannya dengan representasi linguistik adalah kunci keberhasilan (Chen et al., 2024). Selain itu, negara asal peneliti didominasi oleh China, yang menunjukkan konsentrasi riset yang kuat di Asia Timur, terutama dalam aplikasi medis dan teknik.

Analisis terhadap Gambar 6 menunjukkan bahwa pendekatan *Deep Learning* berbasis CNN, RNN, dan LSTM masih menjadi metode yang paling dominan digunakan dalam penelitian *Image-based Question Answering*.



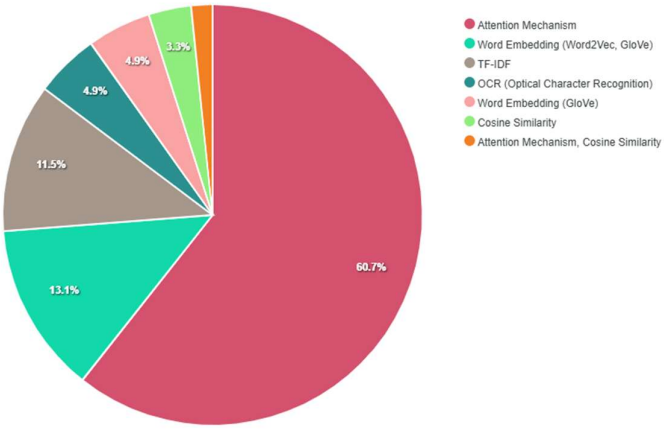
Gambar 6. Klasifikasi Berdasarkan Metode / Pendekatan

Dominasi ini terutama terlihat pada penelitian yang berfokus pada klasifikasi visual dan ekstraksi fitur multimodal dasar. Namun demikian, hasil analisis juga menunjukkan adanya pergeseran metodologis yang cukup signifikan menuju pendekatan berbasis Transformer, Vision-Language Model (VLM), dan Large Language Model (LLM).

Perubahan tersebut terjadi karena arsitektur Transformer memiliki kemampuan yang lebih baik dalam memahami hubungan lintas-modal dan menangani dependensi jangka panjang dibandingkan CNN-RNN tradisional. Penelitian oleh Wang et al. (2023), Islam et al. (2025), dan Hu et al. (2025) menunjukkan bahwa Transformer-based architecture mampu menghasilkan performa yang lebih stabil pada dataset multimodal kompleks seperti OK-VQA dan VQA v2.0 [27], [36], [39].

Gambar 9 memperlihatkan bahwa *Attention Mechanism* menjadi teknik algoritmik paling dominan dalam penelitian *image-based question answering*. Dominasi ini mencerminkan pentingnya mekanisme attention dalam membangun hubungan antara representasi visual dan linguistik secara dinamis. Self-attention, cross-attention, dan multi-head attention menjadi fondasi utama dalam arsitektur Transformer dan multimodal LLM.

Berdasarkan analisis terhadap tabel penelitian terdahulu, gambar 7 mengklasifikasikan teknik atau pendekatan spesifik yang digunakan dalam *image-based question answering* menunjukkan dominasi yang sangat jelas pada Attention Mechanism. Dominasi ini mencakup berbagai sub-tipe seperti *self-attention*, *cross-attention*, *co-attention*, dan *multi-head attention,* yang terlihat pada karya-karya seperti Chen (2024), Bi et al. (2025), serta Wang et al. (2026). Menariknya, distribusi sitasi dari kelompok ini sangat bervariasi.



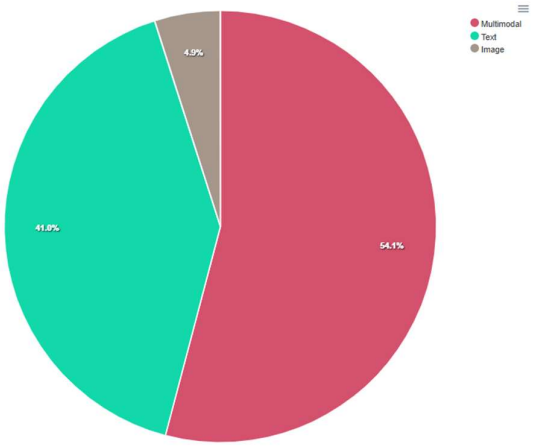
Gambar 7. Klasifikasi Berdasarkan Teknik/Algoritma

Meskipun attention mendominasi secara kuantitatif, penelitian dengan sitasi tinggi justru banyak berasal dari pendekatan yang mengintegrasikan semantic embedding, Vision-Language Model, dan knowledge reasoning. Hal ini menunjukkan bahwa attention mechanism saja tidak cukup untuk menghasilkan reasoning yang robust. Model modern membutuhkan integrasi antara representasi visual, retrieval pengetahuan, dan semantic grounding.

Temuan ini memperlihatkan bahwa arah perkembangan algoritma bergerak menuju hybrid architecture yang menggabungkan Transformer, KG reasoning, dan retrieval mechanism. Integrasi tersebut memungkinkan model tidak hanya memahami gambar, tetapi juga mengakses pengetahuan eksternal untuk menjawab pertanyaan yang kompleks.

 https://doi.org/10.25077/TEKNOSI.v12i2.2026.284-299

Gambar 8 menunjukkan bahwa modalitas multimodal mendominasi penelitian dengan proporsi lebih dari separuh total studi. Dominasi ini menunjukkan bahwa integrasi teks dan gambar telah menjadi paradigma utama dalam pengembangan QA modern. Sistem multimodal memungkinkan model memahami konteks data visual dan tekstual secara simultan untuk menjawab pertanyaan berbasis gambar.



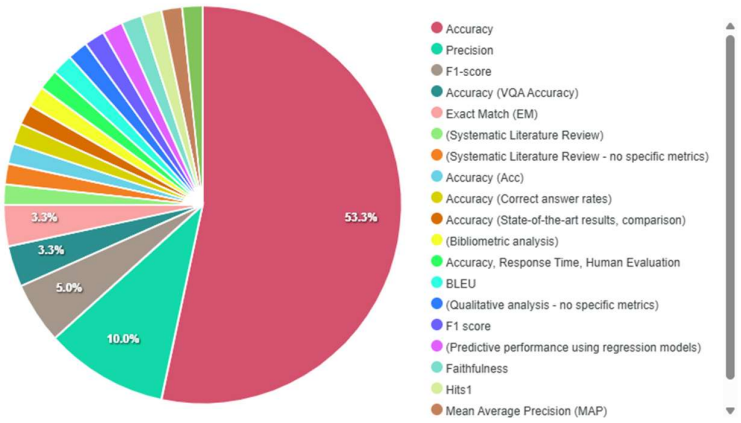
Gambar 8. Klasifikasi Berdasarkan Input

Pendekatan multimodal dianggap lebih efektif karena mampu menghubungkan informasi visual dengan konteks linguistik secara bersamaan. Sistem tidak hanya melakukan identifikasi objek visual, tetapi juga memahami hubungan semantik antara pertanyaan, gambar, dan konteks pengetahuan yang relevan. Penelitian oleh Wang et al. (2026), Silva et al. (2023), dan Kim et al. (2025) menunjukkan bahwa integrasi multimodal mampu meningkatkan performa model secara signifikan dibandingkan pendekatan unimodal [18], [55], [7].

Namun demikian, penelitian berbasis multimodal masih menghadapi tantangan besar terkait kebutuhan data yang tinggi, kompleksitas komputasi, dan bias modalitas. Banyak model cenderung lebih bergantung pada teks dibandingkan informasi visual, sehingga menghasilkan jawaban yang tampak benar tetapi tidak benar-benar memahami gambar.

Temuan ini menunjukkan bahwa masa depan penelitian tidak hanya bergantung pada peningkatan ukuran model, tetapi juga pada kemampuan membangun representasi multimodal yang lebih seimbang dan robust. Integrasi OCR dengan retrieval mechanism menjadi salah satu pendekatan potensial untuk meningkatkan grounding visual pada sistem multimodal.

Gambar 9 dan Gambar 10 menunjukkan bahwa Accuracy (Akurasi) masih menjadi metrik evaluasi paling dominan dalam penelitian VQA dan multimodal QA. Dominasi metrik ini menunjukkan bahwa sebagian besar penelitian di bidang VQA masih berorientasi pada pengukuran performa kuantitatif dasar. Namun, pendekatan tersebut memiliki keterbatasan karena tidak mampu menangkap kualitas reasoning dan explainability model secara komprehensif.



Gambar 9. Matriks Evaluasi

Gambar 10. Tantangan Utama

Analisis tantangan penelitian menunjukkan bahwa *multimodal fusion difficulty* dan *semantic gap* menjadi masalah paling dominan dalam literatur. Kesulitan mengintegrasikan representasi visual dan tekstual secara efektif masih menjadi hambatan utama dalam pengembangan sistem QA berbasis gambar. Selain itu, masalah hallucination, bias dataset, dan domain adaptation juga semakin banyak dibahas pada penelitian terbaru.

Implikasi dari temuan ini sangat signifikan terhadap agenda penelitian masa depan. Pengembangan model yang hanya berorientasi pada akurasi tidak lagi memadai untuk aplikasi kritis seperti medis dan pendidikan. Penelitian masa depan perlu mengembangkan framework evaluasi yang lebih holistik, mencakup interpretability, fairness, robustness, dan contextual reasoning. Dengan demikian, evolusi penelitian *OCR-based and Information Retrieval Techniques for Image-Based Question Answering* menunjukkan bahwa bidang ini sedang bergerak menuju paradigma baru yang menempatkan multimodal reasoning, knowledge integration, dan explainable AI sebagai fondasi utama pengembangan sistem QA generasi berikutnya.

Hasil *Systematic Literature Review* ini menunjukkan bahwa penelitian mengenai *OCR-Based and Information Retrieval Techniques for Image-Based Question Answering* mengalami perkembangan yang sangat cepat dalam lima tahun terakhir, khususnya setelah munculnya arsitektur Transformer, *Vision-Language Models* (VLM), dan *Large Language Models* (LLM). Temuan ini sejalan dengan studi Kim et al. (2025) dan Salaberria et al. (2023) yang menegaskan bahwa paradigma *Visual Question Answering* (VQA) telah bergeser dari pendekatan berbasis ekstraksi fitur menuju sistem multimodal yang mengintegrasikan penalaran visual, linguistik, dan pengetahuan eksternal secara simultan [66], [19]. Namun demikian, ulasan ini menunjukkan bahwa perkembangan tersebut tidak hanya merupakan evolusi teknis, melainkan juga transformasi epistemologis dalam cara sistem AI memahami hubungan antara gambar, teks, dan konteks pengetahuan.

Berbeda dengan penelitian sebelumnya yang umumnya memfokuskan kajian pada performa model VQA secara umum, ulasan ini menawarkan perspektif baru dengan menempatkan integrasi OCR dan *information retrieval* sebagai komponen sentral dalam evolusi *image-based question answering*. Temuan ini memperluas hasil Roy et al. (2022) dan Mohamed et al. (2025) yang masih memandang OCR dan retrieval sebagai komponen pendukung, bukan sebagai inti mekanisme reasoning [8] [51]. Dalam studi-studi terkini, OCR tidak lagi sekadar berfungsi untuk mengekstraksi teks dari gambar, tetapi menjadi jembatan penting dalam membangun *semantic grounding* antara representasi visual dan textual retrieval. Hal ini terlihat jelas pada penelitian Xu et al. (2024) dan Li et al. (2024) yang mengintegrasikan *knowledge retrieval* dan OCR untuk meningkatkan kemampuan sistem menjawab pertanyaan berbasis dokumen visual dan warisan budaya [5], [49].

Temuan lain yang signifikan adalah dominasi pendekatan *Multimodal Deep Learning* dan *Attention Mechanism* dalam sebagian besar penelitian. Hasil ini sejalan dengan Wang et al. (2023), Bi et al. (2025), dan Islam et al. (2025) yang menunjukkan bahwa *attention-based architecture* menjadi fondasi utama dalam menjembatani hubungan visual dan tekstual [28], [62], [36]. Namun, ulasan ini juga menambahkan nuansa penting bahwa dominasi attention mechanism tidak selalu berkorelasi dengan dampak ilmiah tertinggi. Beberapa studi dengan sitasi paling tinggi justru menggunakan pendekatan berbasis *semantic embedding* dan *knowledge graph reasoning*, seperti Liu et al. (2021) dan Jiang dan Meng (2023). Temuan ini mengindikasikan bahwa komunitas riset mulai menyadari keterbatasan attention mechanism sebagai solusi tunggal terhadap persoalan multimodal reasoning [3][20].

Secara akademik, kondisi tersebut menunjukkan bahwa tantangan utama bidang ini telah bergeser dari sekadar *feature extraction* menuju kemampuan reasoning dan contextual understanding. Pergeseran ini memperluas temuan Xiao dan Zhang (2025) yang menyatakan bahwa tantangan utama VQA modern bukan lagi akurasi klasifikasi, tetapi kemampuan model dalam memahami konteks dan melakukan inferensi lintas-modal [38]. Ulasan ini mengungkap aspek yang kurang dipelajari dari penelitian sebelumnya, yaitu bagaimana *semantic gap* dan *multimodal*

fusion difficulty masih menjadi hambatan fundamental meskipun performa model terus meningkat. Dengan kata lain, peningkatanakurasi belum tentu menunjukkan peningkatan kemampuan pemahaman konseptual model.

Temuan ini juga memperlihatkan adanya paradoks metodologis dalam penelitian VQA modern. Di satu sisi, model berbasis Transformer dan LLM berhasil meningkatkan performa pada berbagai benchmark. Namun di sisi lain, sebagian besar penelitian masih bergantung pada metrik evaluasi tradisional seperti Accuracy, BLEU, dan F1-score. Kondisi ini sejalan dengan kritik Cong dan Mo (2025) serta Zhu et al. (2023) yang menyatakan bahwa evaluasi berbasis akurasi cenderung gagal menangkap kualitas reasoning, interpretabilitas, dan robustness model [33], [67]. Ulasan ini mempertegas kritik tersebut dengan menunjukkan bahwa lebih dari separuh studi masih menggunakan akurasi sebagai indikator utama keberhasilan sistem. Dengan demikian, bidang ini menghadapi ketegangan antara optimasi performa numerik dan kebutuhan akan sistem yang dapat dipercaya (*trustworthy AI*).

Berbeda dengan SLR sebelumnya yang lebih berorientasi pada klasifikasi algoritma atau benchmark dataset, ulasan ini menekankan bahwa tantangan sebenarnya terletak pada integrasi pengetahuan eksternal dan kemampuan reasoning berbasis konteks. Temuan ini terlihat dari meningkatnya penggunaan *Knowledge Graph*, *Retrieval-Augmented Generation* (RAG), dan *Large Language Models* dalam penelitian terbaru. Studi Xu et al. (2024), Wang et al. (2022), dan Chunfang et al. (2025) menunjukkan bahwa sistem QA berbasis gambar membutuhkan akses terhadap pengetahuan eksternal agar mampu menjawab pertanyaan kompleks yang tidak dapat diselesaikan hanya melalui representasi visual. Dengan demikian, penelitian ini memperluas temuan Kim et al. (2025) yang sebelumnya menyoroti pentingnya multimodal learning, dengan menambahkan bahwa multimodal reasoning modern membutuhkan integrasi retrieval mechanism secara real-time.

Selain itu, ulasan ini juga menunjukkan bahwa dominasi geografis penelitian mengalami pergeseran signifikan dari negara Barat menuju Asia, khususnya China. Temuan ini sejalan dengan Xu et al. (2024) dan Liu et al. (2021) yang menunjukkan tingginya kontribusi China dalam pengembangan *knowledge-based VQA*. Namun, penelitian ini menambahkan dimensi reflektif bahwa dominasi tersebut tidak hanya disebabkan oleh produktivitas publikasi, tetapi juga oleh fokus pada aplikasi kontekstual seperti *Traditional Chinese Medicine* (TCM), pertanian cerdas, dan warisan budaya. Dengan kata lain, perkembangan VQA modern semakin dipengaruhi oleh kebutuhan domestik dan konteks lokal.

Implikasi akademik dari fenomena ini cukup signifikan. Selama ini, sebagian besar teori VQA dibangun berdasarkan asumsi universalitas dataset dan model. Namun, temuan ulasan ini menunjukkan bahwa konteks budaya, bahasa, dan domain aplikasi memiliki pengaruh yang sangat besar terhadap desain sistem QA multimodal. Hal ini bertentangan dengan paradigma *one-size-fits-all* yang mendominasi penelitian awal VQA. Oleh karena itu, penelitian masa depan perlu bergerak menuju pendekatan *context-aware AI* dan *localized multimodal intelligence* yang lebih adaptif terhadap kebutuhan regional.

Dalam aspek metodologi, hasil review menunjukkan bahwa mayoritas penelitian masih menggunakan pendekatan kuantitatif eksperimental berbasis benchmark. Kondisi ini sejalan dengan Devmane et al. (2026) dan Chen (2024) yang menekankan pentingnya validasi empiris dalam pengembangan model multimodal [63], [57]. Namun, ulasan ini mengungkap bahwa dominasi metode kuantitatif justru menciptakan kekosongan dalam eksplorasi aspek manusiawi, seperti pengalaman pengguna, interpretabilitas, dan penerimaan teknologi. Temuan ini memperluas hasil Holland et al. (2024) yang menyoroti pentingnya aspek kognitif manusia dalam interaksi dengan sistem VQA.

Dengan demikian, salah satu kontribusi baru dari ulasan ini adalah penekanan bahwa perkembangan *image-based question answering* tidak dapat hanya dipahami sebagai persoalan optimasi algoritmik. Bidang ini juga berkaitan erat dengan dimensi kognitif, sosial, dan etis. Temuan ini mengungkap aspek yang kurang dipelajari dari penelitian sebelumnya, khususnya terkait bias dataset, fairness, dan hallucination pada model multimodal besar. Sebagian besar studi masih berfokus pada peningkatan performa teknis tanpa mengevaluasi konsekuensi sosial dari penggunaan model tersebut pada domain kritis seperti medis dan pendidikan.

Ulasan ini juga menunjukkan bahwa integrasi OCR dengan retrieval mechanism membuka peluang baru dalam pengembangan sistem QA berbasis dokumen visual dan screenshot error-code. Berbeda dengan penelitian sebelumnya yang memisahkan OCR dan QA sebagai dua proses independen, penelitian terkini mulai mengintegrasikan keduanya dalam satu pipeline multimodal. Temuan ini memberikan perspektif baru bahwa OCR bukan lagi sekadar tahap preprocessing, tetapi bagian integral dari proses reasoning multimodal. Dengan demikian, penelitian ini menawarkan kontribusi konseptual baru dalam memandang OCR sebagai komponen reasoning, bukan hanya ekstraksi teks.

Dari perspektif teoritis, hasil review ini memperhalus teori *multimodal fusion* yang selama ini berfokus pada penyatuan representasi visual dan linguistik. Temuan penelitian menunjukkan bahwa integrasi pengetahuan eksternal melalui *knowledge graph* dan retrieval system menjadi elemen penting dalam mengatasi *semantic gap*. Oleh karena itu, teori multimodal modern perlu bergerak dari paradigma *feature alignment* menuju *knowledge-grounded multimodal reasoning*. Pendekatan ini sejalan dengan Wang et al. (2022) dan Jiang dan Meng (2023), tetapi ulasan ini menambahkan bahwa grounding semantik juga harus mempertimbangkan konteks budaya dan domain aplikasi.

Implikasi teoritis lainnya adalah perlunya redefinisi terhadap konsep keberhasilan sistem VQA. Selama ini, keberhasilan diukur berdasarkan akurasi jawaban. Namun, hasil review menunjukkan

bahwa sistem dengan akurasi tinggi belum tentu memiliki kemampuan reasoning yang baik atau dapat dipercaya dalam konteks nyata. Oleh karena itu, penelitian masa depan perlu mengembangkan framework evaluasi yang lebih holistik dengan memasukkan aspek interpretability, fairness, robustness, dan explainability. Temuan ini memperluas diskusi Zhu et al. (2023) dan Seki et al. (2025) mengenai pentingnya *explainable AI* dalam sistem multimodal [29], [60].

Secara praktis, hasil review ini memberikan implikasi penting bagi pengembangan sistem QA berbasis gambar di sektor medis, pendidikan, dan layanan teknis. Dalam bidang medis, integrasi OCR, retrieval mechanism, dan knowledge graph dapat meningkatkan kemampuan sistem dalam memahami dokumen klinis dan citra diagnostik secara simultan. Namun, tantangan hallucination dan interpretabilitas tetap menjadi isu kritis karena kesalahan jawaban dapat berdampak langsung terhadap keselamatan pengguna. Oleh karena itu, pengembangan sistem medis berbasis VQA harus menempatkan explainability sebagai prioritas utama.

Dalam konteks pendidikan, hasil review menunjukkan bahwa sistem QA multimodal memiliki potensi besar untuk mendukung pembelajaran adaptif dan aksesibilitas informasi. Penelitian Xiao dan Zhang (2025) serta Bernasconi et al. (2025) menunjukkan bahwa sistem berbasis multimodal dapat meningkatkan interaksi pembelajaran berbasis visual. Namun, ulasan ini menambahkan bahwa keberhasilan implementasi tidak hanya ditentukan oleh performa model, tetapi juga oleh kemampuan sistem memahami konteks budaya dan bahasa pengguna [38], [52].

Pada akhirnya, ulasan ini menegaskan bahwa masa depan *OCR-Based and Information Retrieval Techniques for Image-Based Question Answering* akan bergerak menuju sistem yang lebih adaptif, explainable, dan knowledge-driven. Berbeda dengan penelitian sebelumnya yang menempatkan OCR, retrieval, dan VQA sebagai domain terpisah, ulasan ini menunjukkan bahwa ketiganya kini berkembang menuju ekosistem multimodal yang saling terintegrasi. Dengan demikian, novelty utama penelitian ini terletak pada pemetaan evolusi integratif antara OCR, information retrieval, multimodal reasoning, dan knowledge-enhanced AI dalam konteks *image-based question answering*. Temuan ini tidak hanya memperluas horizon teoritis bidang VQA, tetapi juga memberikan arah strategis bagi pengembangan sistem AI yang lebih kontekstual, terpercaya, dan berdampak sosial di masa depan.

4. KESIMPULAN

Systematic Literature Review ini berhasil memetakan perkembangan penelitian mengenai *OCR-Based and Information Retrieval Techniques for Image-Based Question Answering* secara komprehensif melalui analisis terhadap 61 studi terpilih. Hasil review menunjukkan bahwa bidang ini mengalami transformasi signifikan dari pendekatan berbasis *feature engineering* menuju sistem multimodal berbasis Transformer, *Large Language Models* (LLMs), dan *knowledge-enhanced reasoning*. Dominasi *Visual Question Answering* (VQA), *Attention Mechanism*, serta pendekatan *Multimodal Deep Learning* mengindikasikan bahwa penelitian modern tidak lagi berfokus pada pengenalan visual semata, tetapi telah bergerak menuju integrasi pengetahuan eksternal dan penalaran kontekstual yang lebih kompleks. Temuan ini memperkuat argumentasi bahwa OCR dan *information retrieval* kini berkembang menjadi komponen inti dalam sistem QA multimodal, bukan sekadar tahap pendukung ekstraksi informasi.

Secara akademik, ulasan ini memberikan kontribusi penting dengan menunjukkan bahwa tantangan utama bidang ini tidak lagi terbatas pada peningkatan akurasi model, melainkan pada kemampuan sistem dalam menjembatani *semantic gap*, melakukan reasoning lintas-modal, dan menghasilkan jawaban yang dapat dipercaya. Berbeda dengan penelitian review sebelumnya yang cenderung mengkaji VQA, OCR, atau retrieval secara terpisah, studi ini menawarkan perspektif integratif mengenai hubungan antara OCR, retrieval mechanism, *knowledge graph*, dan multimodal reasoning dalam ekosistem *image-based question answering*. Dengan demikian, penelitian ini memperluas pemahaman konseptual mengenai evolusi sistem QA modern dari pendekatan berbasis representasi menuju sistem *knowledge-driven AI* yang lebih adaptif dan kontekstual.

Hasil review juga menunjukkan adanya pergeseran geografis penelitian dari negara Barat menuju Asia, khususnya China, yang mendominasi baik dari sisi produktivitas maupun dampak ilmiah. Dominasi ini memperlihatkan bahwa pengembangan sistem QA berbasis gambar semakin dipengaruhi oleh kebutuhan domain-spesifik, seperti medis, pertanian cerdas, dan warisan budaya. Namun demikian, kondisi tersebut juga menimbulkan tantangan terkait generalisasi lintas budaya, bias dataset, dan keterbatasan representasi multibahasa yang masih belum banyak dieksplorasi dalam literatur saat ini.

Berdasarkan gap yang teridentifikasi, arah penelitian masa depan perlu difokuskan pada beberapa aspek utama. Pertama, pengembangan model multimodal yang lebih explainable dan robust untuk mengurangi *hallucination* serta meningkatkan transparansi reasoning pada aplikasi kritis seperti medis dan pendidikan. Kedua, integrasi *Knowledge Graph* dan *Retrieval-Augmented Generation* (RAG) secara real-time perlu diperluas untuk mendukung kemampuan reasoning berbasis konteks dan pengetahuan eksternal. Ketiga, penelitian mendatang perlu mengembangkan dataset multibahasa dan multikultural guna meningkatkan fairness dan generalisasi model lintas domain. Keempat, framework evaluasi perlu diperluas melampaui metrik Accuracy menuju evaluasi yang mencakup interpretability, fairness, robustness, dan contextual understanding. Secara keseluruhan, bidang *OCR-Based and Information Retrieval Techniques for Image-Based Question Answering* sedang bergerak menuju paradigma baru yang menempatkan multimodal reasoning, knowledge integration, dan explainable AI sebagai fondasi utama pengembangan sistem cerdas generasi berikutnya. Oleh karena itu, keberhasilan penelitian di masa depan tidak hanya ditentukan oleh kemampuan model mencapai performa tinggi, tetapi juga oleh kapasitasnya

dalam menghadirkan sistem yang adaptif, transparan, terpercaya, dan relevan terhadap kebutuhan sosial yang semakin kompleks.

DAFTAR PUSTAKA

[1] A. Lee, H. Y. Yu, and G. Min, "An algorithm of line segmentation and reading order sorting based on adjacent character detection: A post-processing of OCR for digitization of Chinese historical texts," *J. Cult. Herit.*, vol. 67, pp. 80–91, May 2024, doi: 10.1016/j.culher.2024.02.001.

[2] D. Banerjee, P. Bhowal, S. Malakar, E. Cuevas, M. Pérez-Cisneros, and R. Sarkar, "Z-Transform-Based Profile Matching to Develop a Learning-Free Keyword Spotting Method for Handwritten Document Images," *International Journal of Computational Intelligence Systems*, vol. 15, no. 1, p. 93, 2022, doi: 10.1007/s44196-022-00148-8.

[3] S. Liu, N. Tan, H. Yang, and N. Lukač, "An Intelligent Question Answering System of the Liao Dynasty Based on Knowledge Graph," *International Journal of Computational Intelligence Systems*, vol. 14, no. 1, p. 170, 2021, doi: 10.1007/s44196-021-00010-3.

[4] Q. Li, F. Xiao, B. Bhanu, B. Sheng, and R. Hong, "Inner Knowledge-based Img2Doc Scheme for Visual Question Answering," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 18, no. 3, pp. 1–21, 2022, doi: 10.1145/3489142.

[5] L. Xu, L. Lu, M. Liu, C. Song, and L. Wu, "Nanjing Yunjin intelligent question-answering system based on knowledge graphs and retrieval augmented generation technology," *Herit. Sci.*, vol. 12, no. 1, p. 118, 2024, doi: 10.1186/s40494-024-01231-3.

[6] C. Xie, Z. Zhong, and L. Zhang, "Intelligent maritime question-answering and recommendation system based on maritime vessel activity knowledge graph," *Ocean Engineering*, vol. 312, p. 119115, 2024, doi: 10.1016/j.oceaneng.2024.119115.

[7] B. S. Kim, J. Kim, D. Lee, and B. Jang, "Visual Question Answering: A Survey of Methods, Datasets, Evaluation, and Challenges," *ACM Comput. Surv.*, vol. 57, no. 10, pp. 1–35, 2025, doi: 10.1145/3728635.

[8] P. K. Roy, S. Saumya, J. P. Singh, S. Banerjee, and A. Gutub, "Analysis of community question-answering issues via machine learning and deep learning: State-of-the-art review," *CAAI Trans. Intell. Technol.*, vol. 8, no. 1, pp. 95–117, 2023, doi: 10.1049/cit2.12081.

[9] H. Sun, S. Wang, Y. Zhu, W. Yuan, and Z. Zou, "Question Classification for Intelligent Question Answering: A Comprehensive Survey," Oct. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/ijgi12100415.

[10] D. Moher, L. Stewart, and P. Shekelle, "Implementing PRISMA-P: Recommendations for prospective authors," Jan. 28, 2016, *BioMed Central Ltd*. doi: 10.1186/s13643-016-0191-y.

[11] N. Panic, E. Leoncini, G. De Belvis, W. Ricciardi, and S. Boccia, "Evaluation of the endorsement of the preferred reporting items for systematic reviews and meta-analysis (PRISMA) statement on the quality of published systematic review and meta-analyses," Dec. 26, 2013, *Public Library of Science*. doi: 10.1371/journal.pone.0083138.

[12] A. P. Siddaway, A. M. Wood, and L. V Hedges, "How to Do a Systematic Review: A Best Practice Guide for Conducting and Reporting Narrative Reviews, Meta-Analyses, and Meta-Syntheses," *Annu. Rev. Psychol*, vol. 70, pp. 747–770, 2019, doi: 10.1146/annurev-psych-010418.

[13] M. ter Huurne, A. Ronteltap, R. Corten, and V. Buskens, "Antecedents of trust in the sharing economy: A systematic review," *Journal of Consumer Behaviour*, vol. 16, no. 6, pp. 485–498, Nov. 2017, doi: 10.1002/cb.1667.

[14] E. M. Lasda Bergman, "Finding Citations to Social Work Literature: The Relative Benefits of Using Web of Science, Scopus, or Google Scholar," 2012.

[15] P. I. Rocha, J. H. Caldeira de Oliveira, and J. de M. E. Giraldi, "Marketing communications via celebrity endorsement: an integrative review," Aug. 19, 2020, *Emerald Group Holdings Ltd.* doi: 10.1108/BIJ-05-2018-0133.

[16] E. Hariningsih, B. Haryanto, L. Wahyudi, and C. Sugiarto, "Ten years of evolving traditional versus non-traditional celebrity endorser study: review and synthesis," *Management Review Quarterly*, vol. 75, no. 3, pp. 1937–1997, Sep. 2025, doi: 10.1007/s11301-024-00425-0.

[17] Z. Chunfang, G. Qingyue, Z. Wendong, Z. Jinyang, and L. Huidan, "TCMLCM: an intelligent question-answering model for traditional Chinese medicine lung cancer based on the KG2TRAG method," *Digital Chinese Medicine*, vol. 8, no. 1, pp. 36–45, 2025, doi: 10.1016/j.dcmed.2025.03.011.

[18] Z. Wang, Q. Deng, T. Y. So, W. H. Chiu, and E. S. Hui, "DoKE: Domain knowledge–enhanced multimodal large language model–based framework for chest X–ray follow–up medical visual question answering," *Biomed. Signal Process. Control*, vol. 120, p. 109908, 2026, doi: 10.1016/j.bspc.2026.109908.

[19] A. Salaberria, G. Azkune, O. Lopez De Lacalle, A. Soroa, and E. Agirre, "Image captioning for effective use of lan-

guage models in knowledge-based visual question answering," *Expert Syst. Appl.*, vol. 212, p. 118669, 2023, doi: 10.1016/j.eswa.2022.118669.

[20] L. Jiang and Z. Meng, "Knowledge-Based Visual Question Answering Using Multi-Modal Semantic Graph," *Electronics (Basel).*, vol. 12, no. 6, p. 1390, 2023, doi: 10.3390/electronics12061390.

[21] I. Kim, "Visual Experience-Based Question Answering with Complex Multimodal Environments," *Math. Probl. Eng.*, vol. 2020, pp. 1–18, 2020, doi: 10.1155/2020/8567271.

[22] C. Chen, D. Han, and J. Wang, "Multimodal Encoder-Decoder Attention Networks for Visual Question Answering," *IEEE Access*, vol. 8, pp. 35662–35671, 2020, doi: 10.1109/ACCESS.2020.2975093.

[23] S. Zhang, Y. Zhang, Z. Chen, and Z. Li, "VSAM-Based Visual Keyword Generation for Image Caption," *IEEE Access*, vol. 9, pp. 27638–27649, 2021, doi: 10.1109/ACCESS.2021.3058425.

[24] P. Gao, H. Sun, G. Chen, R. Wang, and M. Li, "Visual Question Answering for Intelligent Interaction," *Mobile Information Systems*, vol. 2022, pp. 1–6, 2022, doi: 10.1155/2022/4232968.

[25] Q. Li, X. Tang, and Y. Jian, "Learning to Reason on Tree Structures for Knowledge-Based Visual Question Answering," *Sensors*, vol. 22, no. 4, p. 1575, 2022, doi: 10.3390/s22041575.

[26] R. Wang, S. Wu, and X. Wang, "The Core of Smart Cities: Knowledge Representation and Descriptive Framework Construction in Knowledge-Based Visual Question Answering," *Sustainability*, vol. 14, no. 20, p. 13236, 2022, doi: 10.3390/su142013236.

[27] R. Wang *et al.*, "FTN–VQA: Multimodal Reasoning By Leveraging A Fully Transformer-Based Network For Visual Question Answering," *Fractals*, vol. 31, no. 06, p. 2340133, 2023, doi: 10.1142/S0218348X23401333.

[28] H. Wang, J. Li, and J. Wang, "Retrieving Chinese Questions and Answers Based on Deep-Learning Algorithm," *Mathematics*, vol. 11, no. 18, pp. 1–18, 2023, doi: 10.3390/math11183843.

[29] H. Zhu, R. Togo, T. Ogawa, and M. Haseyama, "Multimodal Natural Language Explanation Generation for Visual Question Answering Based on Multiple Reference Data," *Electronics (Basel).*, vol. 12, no. 10, p. 2183, 2023, doi: 10.3390/electronics12102183.

[30] J. He *et al.*, "PERS: Parameter-Efficient Multimodal Transfer Learning for Remote Sensing Visual Question Answering," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 17, pp. 14823–14835, 2024, doi: 10.1109/JSTARS.2024.3447086.

[31] T. Yamane, P. Chun, J. Dang, and T. Okatani, "Deep learning-based bridge damage cause estimation from multiple images using visual question answering," *Structure and Infrastructure Engineering*, vol. 22, no. 4, pp. 734–747, 2026, doi: 10.1080/15732479.2024.2355929.

[32] C. Song, "Enhancing Multimodal Understanding With LIUS: A Novel Framework for Visual Question Answering in Digital Marketing," *Journal of Organizational and End User Computing*, vol. 36, no. 1, pp. 1–17, 2024, doi: 10.4018/JOEUC.336276.

[33] Y. Cong and H. Mo, "A visual question answering method based on task decomposition," *PLoS One*, vol. 20, no. 11, p. e0336623, 2025, doi: 10.1371/journal.pone.0336623.

[34] Z. Liu, K. Yang, Y. Weng, Z. He, X. Liu, and H. Gao, "SCAG: Semantic Co-occurring Attention Guided Alignment for Knowledge-based Visual Question Answering," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 21, no. 7, pp. 1–20, 2025, doi: 10.1145/3734220.

[35] C. Huang and Z. Hu, "A multimodal transformer-based visual question answering method integrating local and global information," *PLoS One*, vol. 20, no. 7, p. e0324757, 2025, doi: 10.1371/journal.pone.0324757.

[36] M. S. Islam *et al.*, "MRAN-VQA: Multimodal Recursive Attention Network for Visual Question Answering," *Engineering Science and Technology, an International Journal*, vol. 72, p. 102232, 2025, doi: 10.1016/j.jestch.2025.102232.

[37] C. Gupta, N. S. Gill, P. Gulia, and G. Pau, "CODNet: Context-based object detection network for multimodal image captioning and virtual question answering," *Image Vis. Comput.*, vol. 163, p. 105768, 2025, doi: 10.1016/j.imavis.2025.105768.

[38] J. Xiao and Z. Zhang, "EduVQA: A multimodal Visual Question Answering framework for smart education," *Alexandria Engineering Journal*, vol. 122, pp. 615–624, 2025, doi: 10.1016/j.aej.2025.03.005.

[39] N. Hu, X. Zhang, Q. Zhang, W. Huo, and S. You, "ZPVQA: Visual Question Answering of Images Based on Zero-Shot Prompt Learning," *IEEE Access*, vol. 13, pp. 50849–50859, 2025, doi: 10.1109/ACCESS.2025.3550942.

[40] J. Ma *et al.*, "DeepSORT-OCR: Design and Application Research of a Maritime Ship Target Tracking Algorithm Incorporating Hull Number Features," *Mathematics*, vol. 14, no. 6, p. 1062, 2026, doi: 10.3390/math14061062.

[41] H. Mohammadi and S. H. Khasteh, "A fast text similarity measure for large document collections using multireferencecosine and genetic algorithm," *Turkish Journal Of Electrical Engineering & Computer Sciences*, vol. 28, no. 2, pp. 999–1013, 2020, doi: 10.3906/elk-1906-30.

[42] W. Budiharto, V. Andreas, and A. A. S. Gunawan, "Deep learning-based question answering system for intelligent humanoid robot," *J. Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00341-6.

[43] L.-Q. Cai, M. Wei, S.-T. Zhou, and X. Yan, "Intelligent Question Answering in Restricted Domains Using Deep Learning and Question Pair Matching," *IEEE Access*, vol. 8, pp. 32922–32934, 2020, doi: 10.1109/ACCESS.2020.2973728.

[44] Y. Wen, X. Zhu, and L. Zhang, "CQACD: A Concept Question-Answering System for Intelligent Tutoring Using a Domain Ontology With Rich Semantics," *IEEE Access*, vol. 10, pp. 67247–67261, 2022, doi: 10.1109/ACCESS.2022.3185400.

[45] X. Yang, J. Yang, R. Li, H. Li, H. Zhang, and Y. Zhang, "Complex Knowledge Base Question Answering for Intelligent Bridge Management Based on Multi-Task Learning and Cross-Task Constraints," *Entropy*, vol. 24, no. 12, p. 1805, 2022, doi: 10.3390/e24121805.

[46] Y. Fang, J. Deng, F. Zhang, and H. Wang, "An Intelligent Question-Answering Model over Educational Knowledge Graph for Sustainable Urban Living," *Sustainability*, vol. 15, no. 2, p. 1139, 2023, doi: 10.3390/su15021139.

[47] M. Gao, M. Li, T. Ji, N. Wang, G. Lin, and Q. Wu, "Key Technologies of Intelligent Question-Answering System for Power System Rules and Regulations Based on Improved BERTserini Algorithm," *Processes*, vol. 12, no. 1, p. 58, 2023, doi: 10.3390/pr12010058.

[48] S. Zhang and Q. Jaamour, "English Intelligent Question Answering System Based on elliptic fitting equation," *Applied Mathematics and Nonlinear Sciences*, vol. 8, no. 1, pp. 1743–1752, 2023, doi: 10.2478/amns.2022.2.0162.

[49] R. Li, G. Ren, J. Yan, B. Zou, and Q. Liu, "Intelligent question answering system for traditional Chinese medicine based on BSG deep learning model: taking prescription and Chinese materia medica as examples," *Digital Chinese Medicine*, vol. 7, no. 1, pp. 47–55, Mar. 2024, doi: 10.1016/j.dcmed.2024.04.006.

[50] P. Lyu, J. Fu, C. Liu, W. Yu, and L. Xia, "GPB and BAC: two novel models towards building an intelligent motor fault maintenance question answering system," *Journal of Engineering Design*, vol. 36, no. 11, pp. 2007–2027, 2025, doi: 10.1080/09544828.2024.2335135.

[51] A. Mohamed, K. Abdelqader, and K. Shaalan, "Machine learning and deep learning techniques in Arabic question answering systems: innovations and challenges," *PeerJ Comput. Sci.*, vol. 11, pp. 1–35, 2025, doi: 10.7717/peerj-cs.3331.

[52] E. Bernasconi, D. Redavid, and S. Ferilli, "Enhancing Personalised Learning with a Context-Aware Intelligent Question-Answering System and Automated Frequently Asked Question Generation," *Electronics (Switzerland)*, vol. 14, no. 7, pp. 1–26, 2025, doi: 10.3390/electronics14071481.

[53] B. Zhang, X. Zhang, Q. Wang, G. Gui, and L. Shan, "Intelligent Question Answering System Design with Domain-Specific Knowledge Graphs," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E108.A, no. 3, pp. 546–554, 2025, doi: 10.1587/transfun.2024EAP1090.

[54] D. Sharma, S. Purushotham, and C. K. Reddy, "MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain," *Sci. Rep.*, vol. 11, no. 1, p. 19826, 2021, doi: 10.1038/s41598-021-98390-1.

[55] J. D. Silva, B. Martins, and J. Magalhães, "Contrastive training of a multimodal encoder for medical visual question answering," *Intelligent Systems with Applications*, vol. 18, p. 200221, 2023, doi: 10.1016/j.iswa.2023.200221.

[56] J. Holland, A. McGarvey, M. Flood, P. Joyce, and T. Pawlikowska, "A Qualitative Exploration of Student Cognition When Answering Text-Only or Image-Based Histology Multiple-Choice Questions," *Med. Sci. Educ.*, vol. 34, no. 6, pp. 1317–1329, 2024, doi: 10.1007/s40670-024-02104-x.

[57] L. Chen, "Medical education and artificial intelligence: Question answering for medical questions based on intelligent interaction," *Concurr. Comput.*, vol. 36, no. 14, p. e8079, 2024, doi: 10.1002/cpe.8079.

[58] B. Lei and P. Yin, "Technological Evolution and Research Trends of Intelligent Question-Answering Systems in Healthcare," *Healthcare*, vol. 13, no. 18, p. 2269, 2025, doi: 10.3390/healthcare13182269.

[59] F. Lu, S. Liu, W. Lu, P. Chen, and B. Ding, "Medical Knowledge-Based Differential Image Visual Question Answering," *IEEE Access*, vol. 13, pp. 93818–93829, 2025, doi: 10.1109/ACCESS.2025.3565695.

[60] T. Seki, Y. Kawazoe, H. Ito, Y. Akagi, T. Takiguchi, and K. Ohe, "Assessing the performance of zero-shot visual question answering in multimodal large language models for 12-lead ECG image interpretation," *Front. Cardiovasc. Med.*, vol. 12, p. 1458289, 2025, doi: 10.3389/fcvm.2025.1458289.

[61] Y. Lu *et al.*, "Application of Multimodal Transformer Model in Intelligent Agricultural Disease Detection and Question-Answering Systems," *Plants*, vol. 13, no. 7, p. 972, 2024, doi: 10.3390/plants13070972.

[62] W. Bi, Q. Xiong, X. Chen, Q. Du, J. Wu, and Z. Zhuang, "Intelligent visual question answering in TCM education: An innovative application of IoT and multimodal fusion," *Alexandria Engineering Journal*, vol. 118, no. December 2024, pp. 325–336, 2025, doi: 10.1016/j.aej.2024.12.052.

[63] S. Devmane, O. Rana, and C. Perera, "OntoSage: Intelligent Human-Building Smartbot for Semantic Smart Building Question Answering," *World Wide Web*, vol. 29, no. 2, p. 17, 2026, doi: 10.1007/s11280-026-01403-0.

[64] Y. Huang, H. Liu, S. Li, W. Wang, and Z. Zhou, "Effective Prediction and Important Counseling Experience for Perceived Helpfulness of Social Question and Answering-Based Online Counseling: An Explainable Machine Learning Model," *Front. Public Health*, vol. 10, no. December, pp. 1–19, 2022, doi: 10.3389/fpubh.2022.817570.

[65] Y. Lan, Y. Guo, Q. Chen, S. Lin, Y. Chen, and X. Deng, "Visual question answering model for fruit tree disease decision-making based on multimodal deep learning," *Front. Plant Sci.*, vol. 13, p. 1064399, 2023, doi: 10.3389/fpls.2022.1064399.

[66] S. H. Kim, J. S. Shin, H. Lee, S. Y. Shin, K. M. Kang, and S. Y. Song, "A Comparative Analysis of GPT-3.5, GPT-4, GPT–4 Omni, Gemini Advanced, and Gemini 1.5 in Answering Frequently Asked Questions Regarding High Tibial Osteotomy," *Orthop. J. Sports Med.*, vol. 13, no. 11, pp. 1–13, 2025, doi: 10.1177/23259671251385127.

[67] R. Zhu, B. Liu, R. Zhang, S. Zhang, and J. Cao, "OEQA: Knowledge- and Intention-Driven Intelligent Ocean Engineering Question-Answering Framework," *Applied Sciences*, vol. 13, no. 23, p. 12915, 2023, doi: 10.3390/app132312915.