A Survey of Multimodal Large Language Models and Retrieval-Augmented Generation

www.surveyx.cn

Abstract

This survey paper explores the transformative potential of Multimodal Large Language Models (MLLMs) and Retrieval-Augmented Generation (RAG) techniques, which integrate diverse data modalities such as text, images, and audio to enhance AI capabilities. The survey is structured to elucidate the roles of these advanced AI systems in various domains, particularly highlighting their applications in healthcare, where integration of multimodal data has shown significant promise in improving diagnostic and treatment processes. Key findings emphasize the efficacy of MLLMs in enhancing cancer diagnosis and treatment strategies, advocating for further exploration of multimodal learning in oncology. Despite notable advancements, challenges persist, particularly in complex reasoning scenarios and ethical considerations surrounding AI deployment. The survey underscores the necessity for continued research to address these challenges and improve the robustness and generalization of multimodal models. Future research directions include refining data integration techniques, enhancing evaluation frameworks, and exploring ethical guidelines to ensure responsible AI use. Overall, the survey highlights the vast potential of MLLMs and RAG techniques to facilitate more intelligent and contextually aware AI systems, paving the way for future innovations in AI technology.

1 Introduction

1.1 Structure of the Survey

This survey is structured into key sections that address fundamental aspects of multimodal large language models and retrieval-augmented generation. The introduction elucidates the significance of integrating diverse modalities and the role of retrieval-augmented generation in enhancing AI capabilities. The subsequent section provides a comprehensive overview of essential concepts, including multimodal models, large language models, and cross-modal retrieval, along with their evolution and advancements.

The third section discusses the development and capabilities of multimodal large language models, focusing on the integration of multimodal data, training strategies, and real-world applications. The fourth section examines retrieval-augmented generation, detailing techniques that incorporate external data sources to enhance the generative capacities of AI models.

The fifth section highlights the importance and methods of cross-modal retrieval, emphasizing its applications and impact on AI systems. Following this, the sixth section analyzes the role of information retrieval and knowledge augmentation, exploring methods and tools that improve the accuracy and comprehensiveness of generated content.

Practical applications and case studies are presented in the seventh section, demonstrating the integration of these technologies and their real-world benefits. The eighth section addresses challenges and future directions in the field, including limitations in data, evaluation frameworks, and ethical considerations.

Generated by SurveyX.cn.

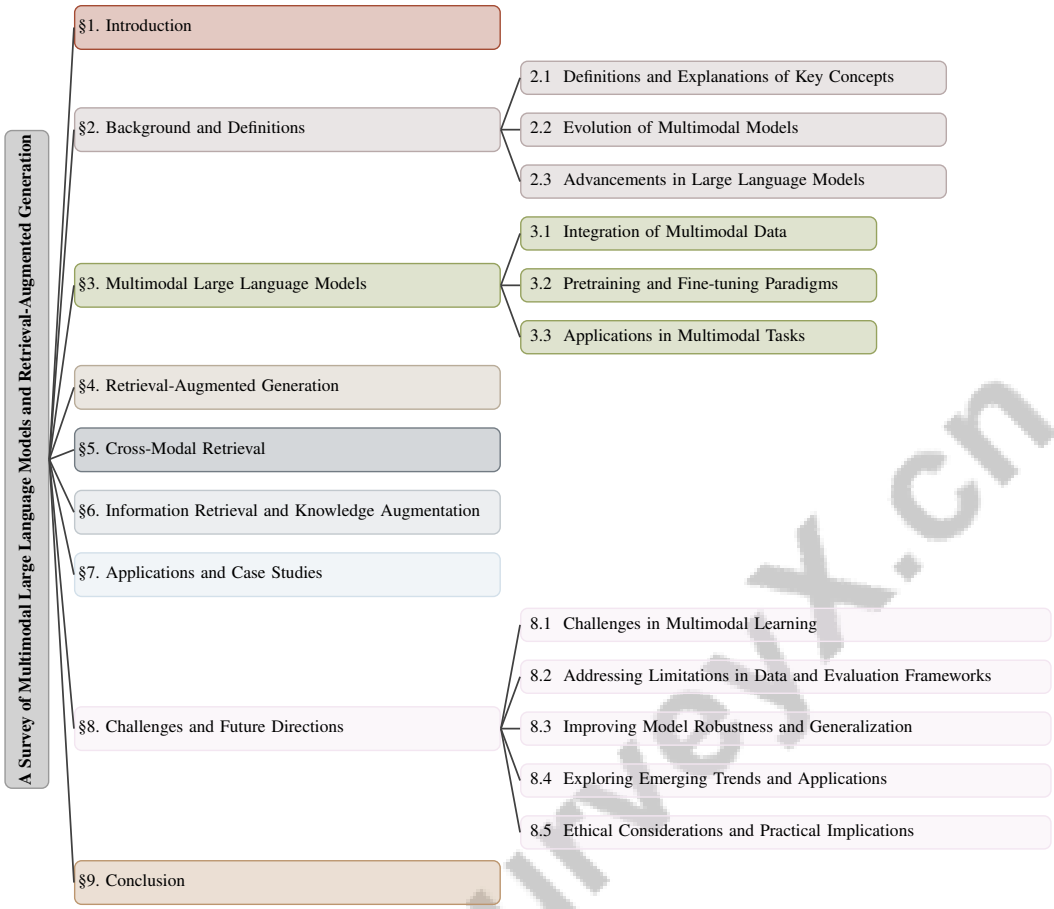


Figure 1: chapter structure

The discussion concludes by encapsulating the essential findings of the research and contemplating the transformative influence of advanced AI systems, particularly multimodal automated academic paper interpretation tools, on future research. These systems enhance literature review efficiency and summarization while addressing critical challenges in managing complex, long-form texts and diverse data formats, potentially reshaping scholarly communication and knowledge dissemination [1, 2].The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Definitions and Explanations of Key Concepts

Multimodal models are advanced AI frameworks designed to process and integrate various data types, such as text, images, and audio, which enhance performance across diverse tasks. These models are critical for applications like multimodal summarization, which synthesizes textual and visual elements [3], and multimodal misinformation detection, where the authenticity of image-caption pairs is assessed [4]. Large Language Models (LLMs) have evolved to incorporate multimodal capabilities, enabling tasks like visual question answering by merging textual and visual data [5]. This integration is particularly beneficial in healthcare, where diverse data types offer a comprehensive understanding of patient information [6].

A significant challenge in multimodal data is ensuring sentence representations are semantically grounded compared to unimodal models [7]. This survey addresses the knowledge gap in preference alignment within Multimodal Large Language Models (MLLMs), focusing on reducing hallucinations and improving alignment with visual data [8]. Retrieval-Augmented Generation (RAG) enhances AI models by retrieving relevant information from extensive databases, enriching responses contextually.

AI-generated, for reference only.

However, existing text-centric multimodal alignment methods lack robustness when dealing with missing or noisy data, affecting task accuracy [9].

Cross-modal retrieval, which aligns information across different data types, is essential for effective multimodal information retrieval, especially when query and reference instances have incomplete modalities [10]. Learning a common representation to bridge heterogeneous data is fundamental for successful retrieval [11]. Knowledge augmentation enriches AI-generated content by incorporating external information, crucial for tasks like knowledge-intensive visual question answering (KI-VQA) [12]. Benchmarks for generating accurate graphical process models from multimodal documents emphasize the importance of comparing representation techniques [13].

Multimodal Knowledge Graphs (MKGs) organize visual-textual factual knowledge, proving useful in structuring complex information [14]. However, reliance on paired datasets for vision-language pre-training poses scalability and performance challenges due to high data annotation costs [15].

2.2 Evolution of Multimodal Models

The evolution of multimodal models has focused on integrating diverse data types to address the limitations of unimodal systems. Initially, efforts aimed to enhance machine learning performance by leveraging complementary modalities for cross-domain tasks, such as product retrieval [16]. The transition from LLMs to MLLMs, particularly in healthcare, has enabled a more comprehensive understanding of patient information through integrated data types, improving diagnostic and treatment processes [6]. However, current benchmarks often emphasize perceptual capabilities over cognitive abilities necessary for understanding text-rich visual scenes [17].

The progression of multimodal models has been driven by the need to overcome the limitations of methods reliant on fully paired datasets and specific task tuning, which restrict generalization. This has led to robust frameworks capable of broader task applicability [9]. The complexity of multimodal content, especially in misinformation detection, highlights the inadequacy of traditional unimodal approaches, necessitating advanced data processing methods [7].

Recent innovations address constraints in text-to-image diffusion models, limited by single language inputs and complex prompts requiring multimodal inputs. Comprehensive models now incorporate stages such as preprocessing, feature extraction, and data fusion for effective integration [14]. Despite advancements, challenges remain in processing long videos exceeding context limits of LLMs and GPU memory, necessitating memory-augmented approaches [18].

The reliance on paired datasets for vision-language pre-training, requiring extensive annotation, poses scalability challenges [15]. MLLMs' tendency to generate hallucinated responses not grounded in visual input highlights a critical area for ongoing research [8].

2.3 Advancements in Large Language Models

Advancements in Large Language Models (LLMs) have significantly expanded their capabilities, enabling a variety of multimodal tasks. Innovations in architectures and training methodologies have facilitated the integration of visual, auditory, and textual data [19]. Multimodal data incorporation has enhanced LLMs' understanding and generation of complex content, exemplified by pretraining paradigms like Croc, which improve visual understanding through cross-modal comprehension [20].

A major challenge remains the large model sizes, resulting in high training and inference costs, which are barriers for researchers outside major enterprises [21]. Traditional multimodal deep learning methods exacerbate these challenges due to computational demands, limiting broader application in low-resource environments [22].

Recent architectural advancements categorize LLMs into encoder-only, decoder-only, and encoder-decoder structures, incorporating visual encoders to enhance capabilities [23]. The MoVA framework exemplifies these advancements by employing a mixture of vision encoders, improving performance across benchmarks while addressing biases in existing encoders [24]. This contrasts with traditional methods producing single-length outputs, as seen in hierarchical representations like Matryoshka, which allow adaptive granularity of visual tokens.

Frameworks such as Optimus have accelerated MLLM training by over 20

AI-generated, for reference only.

Integrating diverse data types from electronic health records (EHRs) presents a critical challenge, requiring effective processing strategies [25]. Existing benchmarks do not adequately measure qualitative differences in information retention across recursive modality changes, underscoring the need for comprehensive evaluation frameworks [18].

The continuous advancement of LLMs into robust systems addresses growing challenges in processing diverse data types, particularly within scientific literature. This evolution paves the way for sophisticated multimodal systems capable of effectively integrating and analyzing various forms of information across multiple domains, enhancing automated interpretation tools' functionality and adaptability. These developments significantly improve information retrieval and decision-making processes in complex applications [26, 9, 6, 1, 27].

In recent years, the development of multimodal large language models has garnered significant attention due to their ability to process and integrate diverse data types. As illustrated in Figure 2, this figure elucidates the hierarchical structure of these models, focusing on the integration of multimodal data, pretraining and fine-tuning paradigms, and their applications in multimodal tasks. Specifically, it highlights key techniques, challenges, and real-world applications, thereby demonstrating the transformative potential of these models in enhancing AI capabilities across various domains. This comprehensive overview not only underscores the complexity of multimodal learning but also sets the stage for understanding its implications for future research and application.

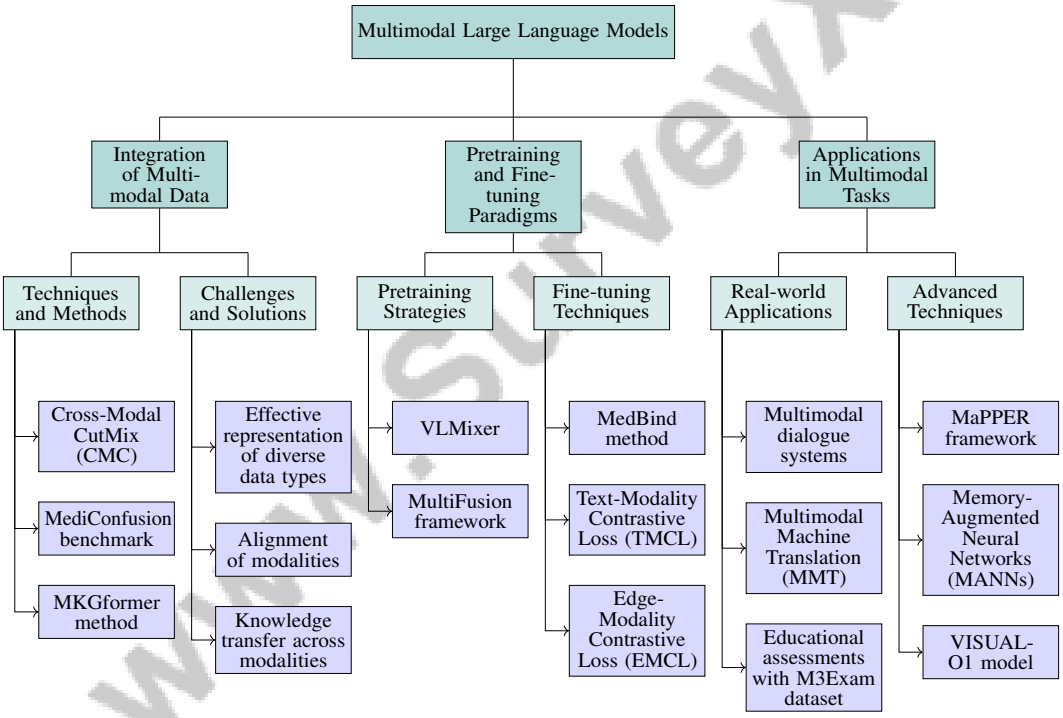


Figure 2: This figure illustrates the hierarchical structure of multimodal large language models, focusing on the integration of multimodal data, pretraining and fine-tuning paradigms, and applications in multimodal tasks. It highlights key techniques, challenges, and real-world applications, demonstrating the transformative potential of these models in integrating and processing diverse data types to enhance AI capabilities across various domains.

3 Multimodal Large Language Models

3.1 Integration of Multimodal Data

The integration of multimodal data within Large Language Models (LLMs) is crucial for enhancing their capability to process and generate content across various modalities, such as text, images, and audio. This process requires sophisticated methodologies to manage heterogeneous data effectively,

AI-generated, for reference only.

enabling models to produce nuanced outputs. Systems like MMAPIS utilize a structured approach for interpreting multimodal academic papers, while the PIN dataset enhances multimodal model performance in complex, knowledge-driven tasks [2, 1, 28]. Integration involves alignment methods and representation learning to combine diverse data types synergistically.

Innovative techniques, such as Cross-Modal CutMix (CMC), transform natural sentences by substituting visually-grounded words with image patches, fostering cross-modal alignments and improving content generation. The MediConfusion benchmark highlights the vulnerabilities of current medical Multimodal Large Language Models (MLLMs), which often perform poorly in distinguishing visually confusing image pairs, raising concerns about their reliability in healthcare and emphasizing the need for better model design and training strategies [6, 5, 29].

Challenges in multimodal integration include representing diverse data types effectively, aligning modalities, and transferring knowledge across them, which are essential for enhancing multimodal systems' performance in real-world applications [30, 28, 9, 1, 31]. The MKGformer method uses a hybrid transformer architecture to unify input-output across tasks, improving entity representation by integrating visual and textual data.

Reducing computational costs while maintaining performance is vital. Frameworks like ETCMA convert various inputs—text, images, and tables—into a unified textual representation, aligning them within a cohesive semantic space. This approach facilitates diverse input interpretation and incorporates advanced techniques like hierarchical discourse-aware summarization, enhancing information clarity and relevance. By addressing modality mismatch and long-form text complexities, these frameworks significantly bolster LLM capabilities in academic and practical applications [30, 9, 1, 2, 32]. The continuous evolution of integration methods reflects an ongoing effort to harness multimodal data's full potential, paving the way for sophisticated AI systems capable of addressing complex tasks with enhanced accuracy and efficiency.

Research is structured into alignment method stages: offline methods, like Direct Preference Optimization, use pre-collected data, while online methods, such as Reinforcement Learning from Human Feedback, sample feedback during training [8]. These methods refine the integration process, ensuring models adaptively learn from diverse inputs and enhance generative capabilities.

3.2 Pretraining and Fine-tuning Paradigms

The development of Multimodal Large Language Models (MLLMs) hinges on effective pretraining and fine-tuning paradigms that facilitate the integration and processing of diverse data modalities. Pretraining leverages extensive datasets to create generalized representations across text, images, and audio, while fine-tuning adapts these models to specific tasks, optimizing performance for applications like sentiment analysis and visual question answering. This process often involves integrating various pretrained unimodal models into a cohesive multimodal framework, enhancing efficiency and outcomes, especially in low-resource scenarios [33, 34, 35, 29].

Pretraining strategies often incorporate methods like VLMixer, which employs cross-modal CutMix to create robust multimodal representations from text and image data [15]. This establishes a strong foundation for adaptability across various tasks. The integration of pretrained language models with multimodal adapters, as seen in the MultiFusion framework, exemplifies the potential of combining text and image inputs to enhance model performance [36].

Fine-tuning paradigms are crucial for refining model outputs through targeted adjustments. The MedBind method illustrates this by employing steps such as data preprocessing for modalities like CXR and ECG, embedding extraction using modality-specific encoders, and applying Text-Modality Contrastive Loss (TMCL) and Edge-Modality Contrastive Loss (EMCL) to fine-tune models for healthcare applications [37]. This underscores the importance of modality-specific adjustments in enhancing output precision and relevance.

Integrating pretraining and fine-tuning strategies is vital for augmenting MLLM capabilities, enabling effective processing and generation of multimodal content—text, images, and audio—with improved accuracy and contextual relevance. Recent advancements indicate that dynamic input scaling and the incorporation of multimodal data, including visual and acoustic information, significantly enhance performance across tasks, particularly in in-context learning and sentiment analysis. By optimizing in-context example selection and refining data filtering methods, researchers can further enhance

AI-generated, for reference only.

MLLMs' efficiency and robustness, paving the way for innovative applications in fields requiring seamless integration of diverse communication modalities [34, 38, 29, 39, 40]. Continuous refinement of these paradigms is essential for developing sophisticated multimodal language models capable of addressing complex tasks across various domains.

3.3 Applications in Multimodal Tasks

Multimodal models demonstrate remarkable capabilities across various real-world applications by effectively integrating and processing diverse data types. In multimodal dialogue systems, models enhance response generation by grounding responses in conversational context and external knowledge, thus improving interaction relevance and accuracy [41]. These systems are essential for creating more natural and contextually aware conversational agents.

In translation, research by Vijayan et al. (2024) highlights the limitations of current Multimodal Machine Translation (MMT) models trained solely on datasets like Multi30k, which often struggle with real-world translation complexities, underscoring the need for advanced approaches to handle diverse linguistic and contextual nuances [42].

In educational contexts, the M3Exam dataset integrates text-rich images and multiple-choice questions requiring visual input for accurate responses, evaluating model capabilities in handling multimodal data and enhancing educational assessments [43]. Similarly, the TRINS dataset, with detailed annotations and a vast collection of text-rich images, serves as a robust benchmark for evaluating multimodal language model performance in complex comprehension tasks [44].

In financial domains, research indicates that Large Language Models (LLMs) can generate long-form summaries of financial reports, often preferred over human-written summaries, highlighting LLMs' potential to efficiently process and summarize complex multimodal documents, enhancing decision-making in financial analysis [2].

The field of computer vision has also benefited from advancements in multimodal models, as exemplified by the MaPPER framework, which employs Dynamic Prior Adapters and Local Convolution Adapters to enhance visual perception and multimodal alignment, addressing limitations of existing methods and improving performance in visual tasks [45].

Memory-Augmented Neural Networks (MANNs) have applications across domains, including natural language processing, computer vision, and multimodal learning, leveraging memory mechanisms to improve information retention and retrieval [46]. The MA-LMM model exemplifies memory augmentation in long-term video understanding tasks, achieving state-of-the-art performance while effectively managing computational resources [47].

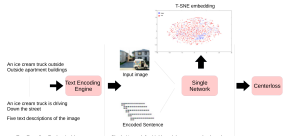
Lastly, the VISUAL-O1 model enhances understanding and execution of ambiguous instructions, enabling general-intelligent models to perform comparably to high-intelligent models, crucial in scenarios requiring precise interpretation and execution of complex instructions [48].

The extensive range of applications for multimodal models across various sectors demonstrates their transformative potential to revolutionize industries by significantly enhancing the accuracy, efficiency, and contextual understanding of artificial intelligence systems. These models integrate diverse data types, such as images, text, and audio, facilitating comprehensive data processing and understanding. Advancements in multimodal representation learning and fusion techniques improve performance in tasks such as image-to-text caption generation, visual question answering, and complex data interpretation in scientific literature. Consequently, multimodal AI is positioned to address the limitations of traditional models, ultimately leading to more effective interactions and insights across numerous fields [26, 1, 27, 31, 49].

As illustrated in Figure 3, multimodal large language models (LLMs) are increasingly utilized to address complex tasks requiring the integration of multiple data modalities, such as text, images, and videos. The examples presented highlight the diverse applications of these models in multimodal tasks. The first example, "Multimodal Answer Refinement with Source Attribution and LLM Integration," demonstrates a sophisticated process where a language model generates a text answer to a user query, refined using multimodal data sources like images and videos for a more comprehensive response. The second example, "Text Encoding and T-SNE Embedding for Image Representation," showcases a technique for encoding and visualizing images in a two-dimensional space using T-SNE embeddings, facilitating clustering of similar data points through center loss computation. Lastly, the

AI-generated, for reference only.

(a) Multimodal Answer Refinement with Source Attribution and LLM Integration[50]



(b) Text Encoding and T-SNE Embedding for Image Representation[51]

(c) Human Annotation and Auto Generation of Text-Related, Image-Related, and Mixed Queries for Bing Concert Hall[52]

Figure 3: Examples of Applications in Multimodal Tasks

third example, "Human Annotation and Auto Generation of Text-Related, Image-Related, and Mixed Queries for Bing Concert Hall," illustrates a method combining human annotation with automated query generation, enhancing the retrieval and understanding of both text and image data related to Bing Concert Hall. Together, these examples underscore the potential of multimodal LLMs to revolutionize data processing and interpretation, paving the way for more intuitive and effective AI applications [50, 51, 52].

4 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) significantly advances AI by enhancing generative capabilities through the integration of external knowledge sources, thereby enriching contextual understanding and improving the relevance and accuracy of generated content. Table 1 offers a detailed comparison of various Retrieval-Augmented Generation techniques, focusing on their integration methods, data types, and unique features. This section delves into specific RAG techniques, illustrating their importance in multimodal integration and processing.

4.1 Retrieval-Augmented Generation Techniques

RAG techniques enhance AI's generative abilities by incorporating external knowledge, enabling the production of contextually rich and precise content. These methods utilize advanced retrieval mechanisms to access and integrate information from extensive databases, thereby improving contextual understanding and output accuracy. The AMPere method exemplifies this by combining visual data with LLM-summarized Automatic Speech Recognition (ASR) text, enhancing product retrieval capabilities through multimodal integration [11].

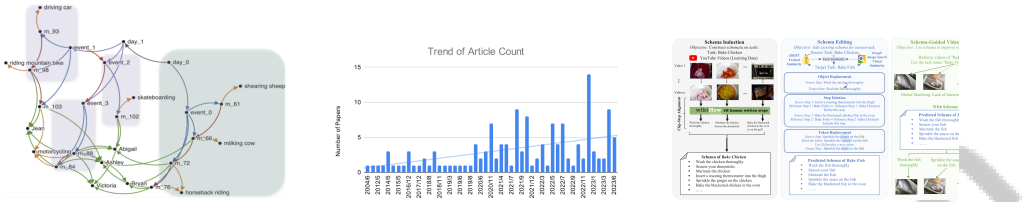
The ChatBridge model integrates modality-specific encoders with a large language model, thus improving the generation of coherent and contextually relevant outputs [53]. This integration is crucial for synthesizing diverse data types and enhancing nuanced response generation. Similarly, the X-REFLECT technique prompts Large Multimodal Models (LMMs) to analyze both textual and visual information, identifying supportive or conflicting elements to improve content accuracy and contextual alignment [54].

Furthermore, the MULTIFUSION framework leverages both textual and visual inputs to enable comprehensive image generation, demonstrating the effectiveness of combining modalities in RAG [36]. In knowledge-intensive visual question answering, the method by Wen et al. uses cross-item interactions between questions and knowledge candidates to derive accurate relevance scores, enhancing response precision [12]. This highlights the importance of effective retrieval mechanisms in improving multimodal output accuracy.

The MSMO method employs a multimodal attention mechanism to fuse text and visual information, facilitating the generation of summaries that incorporate both modalities [3]. This underscores the potential of multimodal attention mechanisms in enhancing the coherence and relevance of generated content. Recent benchmarks emphasize the recursive nature of modality changes, contrasting with previous evaluations focused on static transformations, thus highlighting the need for adaptive retrieval strategies for dynamic content integration [18].

AI-generated, for reference only.

Continuous refinement of RAG techniques, such as enhancing robustness under imperfect conditions through improved alignment, reflects ongoing efforts to advance AI capabilities, enabling the generation of contextually enriched content that aligns with user needs [9]. These advancements pave the way for sophisticated applications across diverse domains, enhancing the accuracy and relevance of AI-generated content.



(a) Social Network Analysis of a Group of Friends[55]

(b) Trend of Article Count[56]

(c) Schema-guided video generation[57]

Figure 4: Examples of Retrieval-Augmented Generation Techniques

As illustrated in Figure 4, RAG enhances content generation by integrating retrieval mechanisms with generative models, leveraging external information sources. The "Social Network Analysis of a Group of Friends" demonstrates RAG's versatility in analyzing social interactions, while the "Trend of Article Count" example highlights RAG's utility in bibliometric analysis, tracking publication trends over nearly two decades. Lastly, the "Schema-guided video generation" showcases a flowchart guiding video creation through schema induction and retrieval from YouTube, emphasizing RAG's potential in multimedia content creation [55, 56, 57].

4.2 Enhancing Multimodal Integration and Processing

RAG significantly enhances the integration and processing capabilities of multimodal systems by leveraging external knowledge to refine contextual understanding and improve content accuracy. Techniques such as AMPere exemplify this enhancement by reducing noise in ASR text, thus improving retrieval accuracy and overall system performance [11]. This approach emphasizes minimizing noise to enhance the coherence and relevance of multimodal outputs.

The CAT framework aggregates relevant clues from audio-visual data, allowing for improved grounding and specificity in responses to questions [58]. By integrating audio-visual cues, CAT enhances the ability of multimodal systems to generate contextually enriched responses.

VISUAL-O1 advances the interpretation of ambiguous instructions through a structured reasoning process that incorporates visual context, enhancing the system's capacity to handle complex tasks [48]. This structured approach is essential for improving multimodal systems' interpretative capabilities.

The ChatBridge model highlights the flexibility of multimodal systems by performing zero-shot tasks across multiple modalities without requiring fully paired training data [53]. This capability enhances adaptability, allowing effective operation across diverse scenarios.

Additionally, the VLMixer approach enhances data diversity and improves instance-level alignment capabilities, reducing reliance on paired datasets and making integration more scalable and efficient [15]. These advancements underscore the critical role of effective integration and processing strategies in multimodal systems, paving the way for sophisticated and contextually aware AI applications across diverse domains.

Figure 5 illustrates the key enhancements in multimodal integration and processing, focusing on noise reduction, contextual enrichment, and adaptability. Techniques like AMPere and VLMixer reduce noise, while CAT and VISUAL-O1 enhance contextual understanding, and ChatBridge improves adaptability for diverse tasks.

www.SurveyX.cn

AI-generated, for reference only.

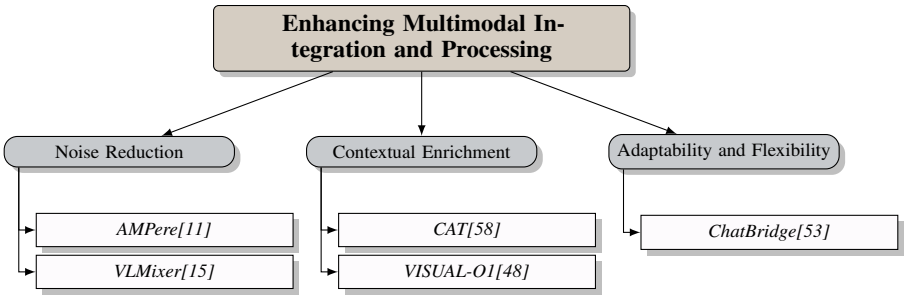


Figure 5: This figure illustrates the key enhancements in multimodal integration and processing, focusing on noise reduction, contextual enrichment, and adaptability. Techniques like AMPere and VLMixer reduce noise, CAT and VISUAL-O1 enhance contextual understanding, and ChatBridge improves adaptability for diverse tasks.

Feature	AMPere	ChatBridge	X-REFLECT
Integration Method	Multimodal Integration	Modality-specific Encoders	Prompting Lmms
Data Types	Visual, Asr Text	Multiple Modalities	Textual, Visual
Unique Feature	Noise Reduction	Zero-shot Tasks	Content Accuracy

Table 1: This table presents a comparative analysis of three Retrieval-Augmented Generation (RAG) techniques: AMPere, ChatBridge, and X-REFLECT. It highlights their integration methods, the types of data they process, and their unique features, providing insights into their respective strengths in multimodal integration and content generation.

5 Cross-Modal Retrieval

5.1 Cross-Modal Retrieval Methods

Cross-modal retrieval methods are crucial for aligning and retrieving information across diverse modalities, such as text and images, by addressing representation discrepancies inherent in these modalities [59]. These techniques aim to bridge the 'heterogeneity gap' caused by distribution differences between vision and language modalities. Among these, the late fusion technique is particularly effective, consistently outperforming alternatives by adeptly combining features from various modalities [60].

A primary challenge in cross-modal retrieval is developing unified representations that accurately capture the semantics of both visual and textual data. Traditional methods often use separate hash codes for correlated data points, leading to inefficiencies. Advancements emphasize feedback in hashing function learning to optimize unified hash codes, enhancing retrieval performance [61].

Multimodal convolutional neural networks (m-CNNs) have been extensively used for bidirectional image and sentence retrieval, showing effectiveness on datasets like Flickr30K and Microsoft COCO. These models leverage convolutional layers to extract and integrate features from multiple modalities, improving retrieval robustness. Techniques such as Any2Any enable direct comparisons across modalities by grounding similarity scores to probabilities, particularly beneficial in scenarios with incomplete data [62].

The implementation of cross-modal consistent cost functions, as demonstrated by the Unified Optimal Transport framework (UOT-RCL), further refines retrieval by addressing label inconsistencies across modalities, ensuring accuracy even with heterogeneous data sources [63]. Additionally, integrating pre-trained image captioning and classification models to generate textual representations of visual content effectively bridges the gap between visual and textual data, facilitating seamless integration and retrieval [64].

In applications like sarcasm detection, cross-modal retrieval is essential, requiring simultaneous analysis of textual and visual information for accurate interpretation [65]. The evolution of cross-modal retrieval methods reflects a commitment to enhancing multimodal data alignment and retrieval, paving the way for more sophisticated and contextually aware AI systems.

AI-generated, for reference only.

5.2 Advancing Cross-Modal Retrieval and Alignment

Recent advancements in cross-modal retrieval focus on improving alignment and retrieval accuracy across diverse modalities by addressing inherent heterogeneity. A notable contribution is the Cross-modal Correlation Learning (CCL) approach, which employs a hierarchical network to model both intra-modality and inter-modality correlations simultaneously [66]. This method effectively integrates multi-grained fusion, leveraging both coarse-grained instances and fine-grained patches to enhance retrieval accuracy by capturing complex modality relationships [66].

The development of unified representations has been pivotal, with methods like Deep Cross-modal Hashing with Unified Code (DCHUC) leading the way. DCHUC is distinguished for jointly learning unified hash codes and hashing functions, achieving improved retrieval accuracy through iterative optimization [61]. This approach addresses the inefficiencies of traditional methods that rely on separate hash codes for correlated data points, thereby enhancing retrieval processes.

Future opportunities in cross-modal retrieval lie in refining methodologies to manage increasingly complex and large-scale datasets. Integrating advanced machine learning techniques, including deep learning and reinforcement learning, could yield sophisticated models capable of adapting to dynamic data environments. Exploring multimodal pre-trained models to generate unified representations across text, images, and audio presents a promising research direction, enhancing our understanding of complex data interactions and improving performance in various applications, such as image captioning, visual question answering, and multimodal classification tasks. These models aim to transcend the limitations of traditional single-modality approaches by integrating diverse data types, facilitating effective information fusion and representation learning [60, 26, 1, 31]. Such advancements are expected to enhance AI capabilities, enabling more accurate and efficient performance in applications requiring multimodal data synthesis.

6 Information Retrieval and Knowledge Augmentation

6.1 Role of Information Retrieval in AI Systems

Information retrieval (IR) is pivotal in advancing artificial intelligence (AI) systems, particularly in the integration and processing of multimodal data. Efficient retrieval and processing of diverse data forms—such as text, images, and audio—enhance AI models' performance across various domains. For example, methods like AMPere leverage LLM-summarized ASR text to improve cross-domain product retrieval, demonstrating IR's capacity to enhance retrieval accuracy and system performance [11]. This integration enhances AI models' contextual understanding, enabling them to generate more precise and relevant outputs.

In healthcare, IR's ability to retrieve and process multimodal data is crucial for improving diagnostic and treatment processes, allowing AI systems to synthesize information from multiple sources to support clinical decision-making [5]. Additionally, IR techniques are vital in multimodal misinformation detection, as illustrated by the COSMOS benchmark, which evaluates MMD models' effectiveness against complex claims involving text and images [4, 1]. This underscores IR's essential role in ensuring AI systems' reliability and trustworthiness.

The ongoing development of IR techniques informs ethical guidelines and safety measures for AI in sensitive contexts, ensuring responsible deployment [67]. Enhancements in robustness to noise, recovery of lost information from other modalities, and improved clarity through summarization and reasoning further highlight IR's significance in advancing AI capabilities [9]. Continuous refinement of IR techniques and evaluation frameworks is crucial for enhancing AI models, enabling them to handle complex multimodal tasks and adapt across various domains, including text-heavy visual content interpretation and academic literature summarization [1, 49].

6.2 Methods and Tools for Effective Information Retrieval

Developing robust information retrieval (IR) methods and tools is essential for optimizing retrieval processes, ensuring accuracy, and improving the contextual relevance of AI outputs. Metrics such as Element Accuracy and Step-Level Success Rate evaluate models' performance in interacting with web user interfaces (UIs), providing a comprehensive assessment of their task execution capabilities [19].

AI-generated, for reference only.

Advanced IR tools utilize machine learning algorithms to enhance retrieval processes, employing techniques like semantic search and relevance feedback to improve result precision. They often leverage pre-trained language models to create embeddings that capture semantic relationships between queries and documents, facilitating accurate retrieval across diverse data types. Multimodal attention mechanisms integrate and analyze data from various sources—including text, images, and audio—enhancing the coherence, relevance, and quality of retrieved content. This integration is crucial for developing automated systems capable of interpreting complex scientific literature and understanding text-heavy visual content, as demonstrated by the superior performance of multimodal approaches in recent studies [1, 27, 49].

Incorporating advanced methods and tools, such as multi-modal automated academic paper interpretation systems, enhanced vision models for text-heavy content, and innovative datasets like PIN for multimodal training, significantly improves AI systems' ability to process large-scale datasets. This not only enhances AI model performance but also broadens their applicability across various fields, enabling more effective analysis and understanding of complex information from diverse sources, including scientific literature and visual data [28, 68, 1, 69, 49]. Continuous refinement of IR techniques and the development of innovative tools are crucial for advancing AI capabilities, ensuring effective utilization of diverse data types to generate accurate and contextually enriched outputs.

6.3 Knowledge Augmentation in AI Systems

Knowledge augmentation is crucial for enhancing AI outputs' accuracy and comprehensiveness by integrating diverse external information sources, enriching reasoning capabilities. Incorporating domain-specific knowledge and varied datasets is vital for advancing multimodal large language models (MLLMs), addressing ethical implications, and ensuring broader applicability across domains [6]. Techniques like Chain-of-Thought (CoT) and Visual Question Answering (VQA) have significantly improved AI performance in multimodal contexts, demonstrating structured reasoning frameworks' potential.

The Valley framework exemplifies effective knowledge augmentation by aligning visual and textual information through a unified approach, contributing to more accurate AI outputs [16]. This alignment is critical for developing AI systems that can interpret and generate content across multiple modalities. The Hybrid Transformer Multi-Level Fusion method further illustrates the ability to model complementary visual and textual information, enhancing knowledge graph completion task accuracy [14].

Additionally, the MCTBench framework provides a comprehensive evaluation of MLLMs' cognitive abilities in real-world applications, offering insights into performance and areas for improvement [17]. This evaluation framework is instrumental in understanding MLLMs' cognitive capabilities and guiding the development of more sophisticated AI systems.

The recursive nature of modality changes can lead to significant information loss, as recent findings indicate [18]. This highlights the necessity for careful evaluation of generative AI tools to ensure that knowledge augmentation processes do not compromise the integrity of processed and generated information.

Knowledge augmentation plays a critical role in enhancing AI capabilities by enabling effective utilization and integration of diverse data types, including textual, visual, and multimodal information. This integration is vital for tasks such as scientific literature interpretation, complex visual content understanding, and advanced question answering, where leveraging various data forms can significantly improve model performance and reasoning abilities [68, 1, 70, 49, 31]. Ongoing refinement of augmentation techniques and evaluation frameworks is essential for enhancing AI models, ensuring their applicability across a wide range of tasks and domains.

7 Applications and Case Studies

7.1 Case Studies and Applications

The integration of multimodal large language models (MLLMs) with retrieval-augmented generation techniques has demonstrated transformative potential across various domains. In augmented reality (AR) and artificial intelligence-generated content (AIGC), these technologies facilitate personalized

AI-generated, for reference only.

AR fitting rooms, real-time media generation, and enhanced multi-user collaboration, showcasing MLLMs' ability to create interactive environments [71]. In multimodal cooperation, the sample-level modality valuation method enhances performance in scenarios with imbalanced modalities, underscoring effective modality integration's role in boosting AI capabilities [72].

The MMHQA-ICL framework exemplifies MLLMs' proficiency in question answering, achieving state-of-the-art results in few-shot settings on the MultimodalQA dataset, highlighting their ability to integrate diverse data types for accurate responses [73]. Experiments with the Multi-modal Dialogue Dataset (MMD) demonstrate a proposed framework's effectiveness in multimodal dialogue tasks, emphasizing MLLMs' potential to enhance conversational agents [74].

In video understanding, the Valley framework achieves state-of-the-art performance across benchmarks, illustrating MLLMs' prowess in processing complex video data [16]. For emotion recognition, models like MulT outperform baselines on datasets such as CMU-MOSI and CMU-MOSEI, enhancing sentiment analysis and human-computer interaction [75]. User satisfaction studies indicate a 12.4% increase for pictorial over text-only summaries, underscoring the benefits of integrating visual elements [3].

These case studies highlight MLLMs' broad applications and transformative potential, emphasizing the strategic selection of multimodal examples to enhance task performance without updating pre-trained parameters. Integrating diverse modalities—textual, visual, and auditory data—improves AI systems' effectiveness and contextual awareness. By exploring novel supervised retrieval methods, these studies pave the way for sophisticated AI systems capable of understanding and interacting with complex information [56, 29].

7.2 Comparative Analysis of MMAPIS and GPT-4

Comparative analysis of MMAPIS and GPT-4 reveals distinct capabilities and performance differences in academic tasks. MMAPIS excels in scientific summarization through hybrid modality preprocessing, hierarchical discourse-aware summarization, and diverse user interfaces. In contrast, GPT-4 struggles with multimodal data and lengthy text summarization, highlighting each model's strengths and limitations in handling complex multimodal academic literature [43, 1, 28]. MMAPIS's integration of diverse data types is crucial for tasks requiring nuanced understanding, particularly in multimodal dialogue systems and emotion recognition, where it leverages visual, textual, and auditory data for enhanced contextual understanding and response accuracy.

Despite GPT-4's adaptation for multimodal inputs, its performance in tasks requiring numeric data integration or key information recognition is surpassed by Claude 2, noted for its summarization capabilities [2]. While GPT-4 faces challenges in synthesizing visual and textual information, MMAPIS's robust multimodal integration framework provides a comprehensive approach to these tasks. This distinction underscores the importance of specialized architectures in improving AI performance, especially in applications requiring multimodal data integration, such as scientific literature interpretation and text-heavy visual content analysis [1, 49].

The evaluation of MMAPIS and GPT-4 highlights significant advancements in artificial intelligence, particularly in multimodal large language models (MLLMs). It emphasizes the need for ongoing innovation in multimodal integration techniques, essential for enhancing AI systems' accuracy and applicability across diverse fields. Notably, MMAPIS demonstrates superior performance in scientific summarization by processing and interpreting multimodal data, underscoring the importance of developing specialized tools to meet researchers' evolving demands in the expanding scientific literature landscape [76, 1].

7.3 PathChat in Pathology Education and Clinical Decision-Making

PathChat exemplifies MLLMs' transformative potential in pathology education and clinical decision-making by integrating diverse data types to enhance diagnostic accuracy and educational outcomes. In education, MLLMs synthesize textual, visual, and auditory elements, enriching comprehension of complex medical information and aligning with the evolving landscape of medical education. This integration supports clinical decision support, patient engagement, and research, addressing challenges in understanding patient health and improving educational outcomes [23, 6, 76, 29, 40].

AI-generated, for reference only.

In clinical practice, PathChat enhances diagnostic accuracy by analyzing diverse patient data types, such as medical images, clinical notes, and laboratory results, supporting improved clinical decision-making [6, 77]. Its ability to process diverse data types enhances utility in complex cases where traditional unimodal approaches may fall short.

PathChat also serves as an interactive learning platform, supporting critical thinking and diagnostic skills development. By simulating real-world scenarios, it allows students to engage with multimodal content, fostering a deeper understanding of pathology concepts and practical applications. This approach improves educational outcomes by leveraging advanced techniques like retrieval-augmented generation and multimodal integration, equipping students with the critical reasoning and contextual understanding necessary for effective navigation of clinical complexities [70, 6, 1].

PathChat plays a crucial role in continuing medical education, offering access to cutting-edge research and clinical guidelines, supporting lifelong learning for healthcare practitioners. Its multimodal capabilities allow interaction with various data types—text, images, and audio—enhancing understanding of complex medical concepts and improving patient care outcomes. This platform facilitates continuous professional development and addresses challenges posed by the rapid expansion of scientific literature, helping practitioners navigate and apply new knowledge effectively in clinical environments [77, 6, 1, 78, 49]. This model is essential for maintaining high care standards and adapting to the evolving medical practice landscape.

PathChat's transformative potential is highlighted by its ability to enhance learning and patient care through the seamless integration of diverse data types, including text, images, and audio, facilitating improved clinical decision-making, patient engagement, and personalized healthcare delivery [23, 77, 79, 6, 29].

7.4 VISUAL-O1's Impact on Ambiguous Instruction Processing

VISUAL-O1 represents a significant advancement in processing ambiguous instructions, particularly in contexts requiring clarity and precision. By employing a structured reasoning process that integrates visual context, VISUAL-O1 enhances AI systems' interpretative capabilities, enabling them to handle complex and nuanced tasks effectively. This capability is crucial in interactive systems and autonomous agents, where multimodal data integration—text, images, and structured visual content—enhances understanding and interaction, facilitating effective communication and task performance in complex environments [19, 1, 69, 49].

The model's ability to process ambiguous instructions is supported by its sophisticated multimodal integration framework, which combines visual, textual, and contextual data to derive meaning. This approach significantly enhances instruction processing accuracy and empowers the system to adaptively learn from diverse inputs, improving robustness and flexibility in rapidly changing environments. By integrating advanced multimodal capabilities, such as dynamic input scaling and hierarchical discourse-aware summarization, the system effectively categorizes and interprets various data types, including text, tables, and figures, while employing tailored user interfaces for diverse applications. This comprehensive methodology addresses traditional models' challenges and ensures superior performance in scientific summarization and vision-language tasks [38, 1].

In practical applications, VISUAL-O1 has demonstrated utility across domains, including human-computer interaction and autonomous navigation, where accurate instruction interpretation is critical for successful task execution. The model employs a structured reasoning process to disambiguate instructions by effectively integrating contextual cues and visual information. This approach minimizes error risks and significantly boosts overall system performance, as evidenced by advancements in multimodal reasoning frameworks like VISUAL-O1 and self-consistency training methods that enhance rationale generation. These frameworks enable models to interpret ambiguous instructions and complex visual-textual data, showcasing their potential for human-like reasoning in uncertain and ambiguous real-world scenarios [80, 48, 81, 76, 49].

VISUAL-O1's impact extends to educational and training applications, serving as a tool for developing critical thinking and problem-solving skills. It enhances learning outcomes by providing interactive scenarios that require interpreting ambiguous instructions, fostering a deeper understanding of complex concepts. This multimodal multi-turn chain-of-thought reasoning framework addresses the challenges posed by ambiguous language in real-world tasks, enabling models to integrate visual context and common sense for accurate interpretation. Unlike traditional methods demanding high

AI-generated, for reference only.

intelligence for lengthy text comprehension, VISUAL-O1 effectively supports models of varying intelligence levels without significantly increasing computational overhead. Experimental results demonstrate that this approach improves model performance on ambiguous instructions and enhances capabilities in processing general datasets, ultimately simulating human-like reasoning amidst uncertainty and ambiguity [48, 49].

VISUAL-O1 signifies a substantial advancement in multimodal AI systems, effectively processing and interpreting ambiguous instructions through a sophisticated multimodal multi-turn chain-of-thought reasoning framework. This approach enhances instruction comprehension accuracy by integrating visual context and common sense while improving AI models' adaptability across various real-world scenarios, addressing challenges posed by ambiguous language and weak reasoning abilities. By simulating human-like reasoning processes, VISUAL-O1 promises to elevate the performance of both highly intelligent and generally intelligent models, contributing to the evolution of AI systems capable of operating effectively in uncertain environments [82, 48, 49].

8 Challenges and Future Directions

8.1 Challenges in Multimodal Learning

Multimodal learning faces significant challenges in scalability and effectiveness due to the complexities of integrating diverse data types and existing methodological limitations. The development of robust multimodal models is hampered by computational expense and the need for large aligned datasets, which restrict scalability compared to self-supervised approaches that do not heavily rely on extensive annotations [9]. The limited accessibility of high-quality datasets further complicates this issue [7]. Complex user prompts pose another challenge, necessitating advanced methods for parsing detailed instructions across modalities [16]. Traditional large language models (LLMs) are insufficient for tasks that require complex reasoning and understanding of physical scenes, exposing a critical gap in current methodologies [5].

Benchmarks often suffer from unimodal bias, as seen with datasets like COSMOS, which inadequately represent multimodal task complexities [4, 7]. Additionally, limitations in datasets such as MCTBench, which focus predominantly on English, restrict model generalization to multilingual contexts [17]. Continual learning for multimodal models introduces further challenges, particularly in integrating knowledge from previous tasks without significant computational costs, requiring careful tuning of hyperparameters and model components [14]. The inherent randomness in LLM text generation also complicates consistency and reliability [9].

The reliance on extensive preprocessing and the variable quality of source data further constrain multimodal learning systems' effectiveness [7]. Robust preprocessing techniques are essential for handling diverse data inputs, and sophisticated temporal modeling techniques are needed to ground long-range temporal information in videos and audios [53]. Integrating more than two modalities remains challenging, limiting the effectiveness of current methods in managing complex clinical data [37]. The increased processing time for extremely long videos poses a significant challenge as sequence lengths grow [47]. Moreover, noisy ASR text and the limitations of LLM-based summarizers complicate the retrieval process [11].

Addressing these challenges is crucial for enhancing multimodal AI systems' capabilities across various domains, including politics, health, environment, sports, and finance, where systems leverage both visual and textual information to improve performance in tasks like news retrieval, image-to-text caption generation, and visual question answering [27, 31].

8.2 Addressing Limitations in Data and Evaluation Frameworks

Overcoming significant limitations in data and evaluation frameworks is vital for the development of robust multimodal AI systems. Challenges like position bias and the integration of numeric data in summaries are critical for enhancing AI-generated content's accuracy and relevance [2]. Addressing these challenges requires refining data processing techniques and developing comprehensive evaluation metrics that accurately reflect multimodal tasks' complexities. Future research should focus on creating domain-specific embeddings and exploring advanced pre-training techniques to improve multimodal deep learning's applicability and performance, especially in low-resource environments

AI-generated, for reference only.

Benchmark	Size	Domain	Task Format	Metric
HierarCaps[83]	73,000	Multimodal Learning	Hierarchical Image-Text Matching	Precision, Recall
CoMMuTE[42]	1,000	Multimodal Machine Translation	Translation	BLEU4, CoMMuTE
VLM-Benchmark[64]	39,928	Visual Reasoning	Classification	Accuracy, F1-score
MIR[27]	611	News Retrieval	Multimodal Retrieval	Average Precision
COMET[55]	103,400	Multimodal Dialog Systems	Task-oriented Dialogs	Accuracy, F1
MMBench[84]	3,217	Multimodal Understanding	Multiple-choice Question Answering	Accuracy, F1-score
ACES[85]	29,645	Audio Captioning	Evaluation Metric	ACES
MIMIC-IT[86]	2,800,000	Visual Language Processing	Instruction Following	Accuracy, Elo Rating

Table 2: This table presents a comprehensive overview of representative benchmarks used in multimodal AI research. It details the size, domain, task format, and evaluation metrics of each benchmark, providing essential insights into their application and significance in advancing multimodal AI systems.

[22]. Tailoring embeddings to specific domains can enhance precision and contextual understanding across diverse applications.

Exploring methods to integrate natural differences between modalities into the valuation process could significantly improve multimodal large language models' performance, allowing them to leverage each modality's unique characteristics effectively [72]. Recognizing these differences may lead to a more nuanced understanding of multimodal data, improving integration and processing capabilities. Existing benchmarks often emphasize linear hierarchies of texts, which may not fully capture the complexity of branching structures in natural descriptions [83]. Future evaluation frameworks should incorporate sophisticated hierarchical models that accommodate diverse and intricate relationships between multimodal data elements, crucial for accurately assessing AI systems' performance in real-world scenarios.

Addressing limitations in data and evaluation frameworks is essential for enhancing multimodal AI systems' capabilities, particularly in improving vision models for processing text-heavy visual content and developing automated interpretation systems for academic literature [1, 49]. Developing comprehensive and contextually relevant evaluation metrics will ensure AI models can handle multimodal tasks' complexities, paving the way for accurate and reliable applications across various domains. Table 2 provides a detailed summary of key benchmarks utilized in the evaluation of multimodal AI systems, highlighting their domain, task format, and associated metrics.

8.3 Improving Model Robustness and Generalization

Enhancing multimodal models' robustness and generalization is crucial for their effective deployment across diverse applications. Future research should prioritize hybrid approaches combining few-shot learning and gradient-based methods to improve in-context learning robustness and mitigate biases in model outputs [87]. Such methods could leverage both paradigms' strengths, ensuring models can adaptively learn from limited data while maintaining high performance across varied tasks. Addressing biases, particularly those related to agreeableness in multimodal systems, requires exploring a broader range of models and practical interventions to enhance fairness and reliability [88].

Optimizing feature selection and refining attention mechanisms are critical for improving model robustness and generalization. Enhancing these components can contribute to more resilient models capable of maintaining performance across various conditions [89]. Improving the model's ability to handle phrase-level semantics will enhance performance, enabling a more nuanced understanding and processing of complex inputs [90]. In multimodal federated learning, improving communication efficiency and addressing challenges posed by heterogeneous client environments remain pivotal. Future work should focus on developing strategies that enhance federated learning frameworks' adaptability and scalability, ensuring their effectiveness in real-world scenarios [91].

Enhancing retrieval strategies' robustness is essential to accommodate dynamic knowledge retrieval needs effectively. By focusing on adaptive retrieval mechanisms, future research can ensure models remain responsive and accurate in rapidly changing information landscapes [92]. A comprehensive strategy is essential for improving multimodal models' robustness and generalization, encompassing learning paradigms optimization, bias mitigation, model components refinement, and adaptive

AI-generated, for reference only.

retrieval strategies development. This involves leveraging text-centric approaches to unify diverse modalities into coherent textual representations, improving model interpretation and robustness in the face of missing data or noise. Addressing biases in in-context example selection, particularly the overemphasis on visual data, through supervised retrievers can significantly enhance multimodal in-context learning efficiency. Integrating these multifaceted approaches can foster multimodal representations' adaptability and effectiveness, making them more suitable for dynamic, real-world applications [9, 29]. These efforts are crucial for multimodal AI systems' continued evolution and applicability across diverse domains.

8.4 Exploring Emerging Trends and Applications

Emerging trends in multimodal AI systems focus on optimizing model architectures and enhancing multimodal data integration to improve robustness and adaptability. Models like CAT exemplify advancements in handling complex audio-visual tasks and improving robustness in ambiguous scenarios, crucial for enhancing AI systems' interpretative capabilities [58]. Innovations in multimodal product representation learning, particularly using ASR text, demonstrate the potential for improving retrieval performance by leveraging diverse data types [11].

Future research directions include expanding existing datasets' capabilities, such as the WanJuan dataset, to integrate more diverse data sources and ensure high data quality, supporting more comprehensive model training [10]. Enhancing models like ChatBridge to better handle long-range temporal data and integrate additional modalities, such as sketches and point clouds, could significantly improve their performance and applicability across various domains [53]. In interactive image generation, future efforts could focus on expanding the range of modalities supported by frameworks like MULTIFUSION, enhancing their versatility and application scope [36]. Developing scalable multimodal frameworks in fields like oncology is essential for improving interpretability, addressing data privacy concerns, and leveraging federated learning to enhance model performance across diverse datasets [13].

Hierarchical processing methods for long videos, integrating video-based encoders, and utilizing large-scale video-text datasets for pre-training are promising directions for advancing memory-augmented multimodal models [47]. Improving alignment between text and images and extending methods to include other modalities, such as video, could enhance multimodal AI systems' effectiveness [3]. Addressing unimodal biases and integrating additional evidence for improved detection accuracy are critical for developing more reliable multimodal misinformation detection methods. Establishing clearer guidelines for creating real-world MMD benchmarks will contribute to these systems' robustness and applicability [4]. Expanding evaluation frameworks like MCTBench to encompass a broader range of models and potentially multilingual datasets will provide more comprehensive insights into multimodal AI systems' cognitive capabilities [17].

Emerging trends and applications underscore the potential for continued innovation in multimodal AI systems, paving the way for more sophisticated and contextually aware technologies that can effectively address complex challenges across various domains. Future research should explore the effects of using different models for each iteration and investigate additional types of images to enhance understanding of information loss [18]. As the Valley framework evolves, incorporating audio inputs and expanding multilingual capabilities will further enhance its performance [16]. Additionally, future research should refine visual and textual modalities integration, explore emerging trends in multimodal representation learning, and address identified limitations [7]. Exploring the application of MKGformer to additional tasks, such as multimodal event extraction and sentiment analysis, along with enhancing visual representations with textual data, presents promising opportunities for further advancements [14]. Future research could investigate enhancements to the cross-modal CutMix technique and its applicability to other modalities beyond vision and language [15].

8.5 Ethical Considerations and Practical Implications

The deployment of multimodal large language models (MLLMs) and retrieval-augmented generation systems raises significant ethical considerations and practical implications that must be addressed to ensure responsible and effective use. A primary ethical concern is the potential for biases embedded in pre-trained models, leading to skewed outputs and reinforcing existing social biases. This issue is particularly pertinent in recommendation systems, where there is a risk of promoting biased or harmful

AI-generated, for reference only.

content, necessitating robust content filtering mechanisms to mitigate such risks [54]. In healthcare, deploying unreliable medical MLLMs poses significant ethical challenges, underscoring the need for improved benchmarks to ensure safety and efficacy in clinical applications [5]. Integrating MLLMs in clinical settings requires establishing ethical and regulatory frameworks to ensure responsible technology use, safeguarding patient safety and privacy [6].

The lack of comprehensive coverage in current multimodal preference datasets presents challenges in generalizing findings, hindering robust and reliable models' development [8]. This limitation underscores the necessity for ongoing assessments and developing a more nuanced understanding of AI capabilities beyond human comparisons as AI technology evolves [67]. Privacy concerns are particularly pertinent in multi-user scenarios, where the hierarchical complexity of data sharing and user interactions necessitates comprehensive design guidance to protect user information. As these systems become more sophisticated and capable of processing sensitive data, careful consideration of privacy implications is essential [48].

Future research should focus on developing methods that do not rely on high-intelligent models during optimization to extend systems like VISUAL-O1's applicability, ensuring adaptability and cultural awareness [48]. This approach will help mitigate biases and improve instruction-tuned models' generalizability and reliability in real-world scenarios. Deploying advanced AI systems necessitates a comprehensive approach to address ethical considerations and practical implications, including strategies for bias mitigation, robust privacy protection, and formulating ethical guidelines that inform their development and application across diverse fields. This approach is crucial given the rapid advancements in large language models and generative AI, which raise significant concerns about their capabilities, potential biases, and the need for effective integration into existing frameworks, particularly in areas such as augmented reality and scientific literature interpretation [71, 1, 67, 76, 31].

9 Conclusion

The exploration of multimodal large language models (MLLMs) and retrieval-augmented generation (RAG) techniques reveals their substantial impact on the advancement of AI technologies across multiple sectors. MLLMs, which seamlessly integrate modalities such as text, images, and audio, have shown remarkable potential, especially in healthcare applications. Frameworks like MEDBind, which incorporate CXR, ECG, and textual data, have demonstrated superior performance in both zero-shot and downstream tasks, underscoring the need to address current research gaps. The potential of MLLMs in enhancing cancer diagnosis and treatment strategies highlights the importance of further exploration in oncology. Additionally, advancements in mitigating hallucinations in MLLMs, as evidenced by HDPO's performance, are pivotal for enhancing the reliability of AI outputs. Despite these advancements, significant challenges remain, particularly in scenarios requiring complex reasoning where MLLMs still face limitations. The lack of stable human-like personality traits in models like GPT-3.5 points to the need for further research into cognitive consistency and scalability in real-world applications. The survey underscores the critical role of ongoing research in addressing these challenges and ethical considerations, with the promise of developing more intelligent and contextually aware AI systems. As these technologies evolve, they hold the potential to transform AI capabilities, offering innovative solutions to complex problems across diverse fields.

www.surveyx.cn

AI-generated, for reference only.

References

[1] Feng Jiang, Kuang Wang, and Haizhou Li. Bridging research and readers: A multi-modal automated academic papers interpretation system, 2024.

[2] Tianyu Cao, Natraj Raman, Danial Dervovic, and Chenhao Tan. Characterizing multimodal long-form summarization: A case study on financial reports, 2024.

[3] Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. Msmo: Multimodal summarization with multimodal output. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4154–4164, 2018.

[4] Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C. Petrantonakis. Synthetic misinformers: Generating and combating multimodal misinformation, 2023.

[5] Mohammad Shahab Sepehri, Zalan Fabian, Maryam Soltanolkotabi, and Mahdi Soltanolkotabi. Mediconfusion: Can you trust your ai radiologist? probing the reliability of multimodal medical foundation models, 2024.

[6] Qian Niu, Keyu Chen, Ming Li, Pohsun Feng, Ziqian Bi, Lawrence KQ Yan, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Junyu Liu, Benji Peng, Tianyang Wang, Yunze Wang, Silin Chen, and Ming Liu. From text to multimodality: Exploring the evolution and impact of large language models in medical practice, 2024.

[7] Jindřich Libovický and Pranava Madhyastha. Probing representations learned by multimodal recurrent and transformer models, 2019.

[8] Elmira Amirloo, Jean-Philippe Fauconnier, Christoph Roesmann, Christian Kerl, Rinu Boney, Yusu Qian, Zirui Wang, Afshin Dehghan, Yinfei Yang, Zhe Gan, and Peter Grasch. Understanding alignment in multimodal llms: A comprehensive study, 2024.

[9] Ting-Yu Yen, Yun-Da Tsai, Keng-Te Liao, and Shou-De Lin. Enhance the robustness of text-centric multimodal alignments, 2024.

[10] Conghui He, Zhenjiang Jin, Chao Xu, Jiantao Qiu, Bin Wang, Wei Li, Hang Yan, Jiaqi Wang, and Dahua Lin. Wanjuan: A comprehensive multimodal dataset for advancing english and chinese large models, 2023.

[11] Ruixiang Zhao, Jian Jia, Yan Li, Xuehan Bai, Quan Chen, Han Li, Peng Jiang, and Xirong Li. Asr-enhanced multimodal representation learning for cross-domain product retrieval, 2024.

[12] Haoyang Wen, Honglei Zhuang, Hamed Zamani, Alexander Hauptmann, and Michael Bendersky. Multimodal reranking for knowledge-intensive visual question answering, 2024.

[13] Asim Waqas, Aakash Tripathi, Ravi P. Ramachandran, Paul Stewart, and Ghulam Rasool. Multimodal data integration for oncology in the era of deep neural networks: A review, 2024.

[14] Xiang Chen, Ningyu Zhang, Lei Li, Shumin Deng, Chuanqi Tan, Changliang Xu, Fei Huang, Luo Si, and Huajun Chen. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion, 2023.

[15] Teng Wang, Wenhao Jiang, Zhichao Lu, Feng Zheng, Ran Cheng, Chengguo Yin, and Ping Luo. Vlmixer: Unpaired vision-language pre-training via cross-modal cutmix. In *International Conference on Machine Learning*, pages 22680–22690. PMLR, 2022.

[16] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Da Li, Pengcheng Lu, Tao Wang, Linmei Hu, Minghui Qiu, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability, 2023.

[17] Bin Shan, Xiang Fei, Wei Shi, An-Lan Wang, Guozhi Tang, Lei Liao, Jingqun Tang, Xiang Bai, and Can Huang. Mctbench: Multimodal cognition towards text-rich visual scenes benchmark, 2024.

AI-generated, for reference only.

www.surveyx.cn

[18] Javier Conde, Tobias Cheung, Gonzalo Martínez, Pedro Reviriego, and Rik Sarkar. Analyzing recursiveness in multimodal generative artificial intelligence: Stability or divergence?, 2024.

[19] Junpeng Liu, Tianyue Ou, Yifan Song, Yuxiao Qu, Wai Lam, Chenyan Xiong, Wenhu Chen, Graham Neubig, and Xiang Yue. Harnessing webpage uis for text-rich visual understanding, 2024.

[20] Yin Xie, Kaicheng Yang, Ninghua Yang, Weimo Deng, Xiangzi Dai, Tiancheng Gu, Yumeng Wang, Xiang An, Yongle Zhao, Ziyong Feng, Roy Miles, Ismail Elezi, and Jiankang Deng. Croc: Pretraining large multimodal models with cross-modal comprehension, 2024.

[21] Yizhang Jin, Jian Li, Yexin Liu, Tianjun Gu, Kai Wu, Zhengkai Jiang, Muyang He, Bo Zhao, Xin Tan, Zhenye Gan, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. Efficient multimodal large language models: A survey, 2024.

[22] David Restrepo, Chenwei Wu, Sebastián Andrés Cajas, Luis Filipe Nakayama, Leo Anthony Celi, and Diego M López. Multimodal deep learning for low-resource settings: A vector embedding alignment approach for healthcare applications, 2024.

[23] Hanguang Xiao, Feizhong Zhou, Xingyue Liu, Tianqi Liu, Zhipeng Li, Xin Liu, and Xiaoxuan Huang. A comprehensive survey of large language models and multimodal large language models in medicine, 2024.

[24] Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. Mova: Adapting mixture of vision experts to multimodal context. *arXiv preprint arXiv:2404.13046*, 2024.

[25] Jun-En Ding, Phan Nguyen Minh Thao, Wen-Chih Peng, Jian-Zhe Wang, Chun-Cheng Chug, Min-Chen Hsieh, Yun-Chien Tseng, Ling Chen, Dongsheng Luo, Chi-Te Wang, Pei fu Chen, Feng Liu, and Fang-Ming Hung. Large language multimodal models for 5-year chronic disease cohort prediction using ehr data, 2024.

[26] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and S Yu Philip. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE, 2023.

[27] Golsa Tahmasebzadeh, Sherzod Hakimov, Eric Müller-Budack, and Ralph Ewerth. A feature analysis for multimodal news retrieval, 2020.

[28] Junjie Wang, Yin Zhang, Yatai Ji, Yuxiang Zhang, Chunyang Jiang, Yubo Wang, Kang Zhu, Zekun Wang, Tiezhen Wang, Wenhao Huang, Jie Fu, Bei Chen, Qunshu Lin, Minghao Liu, Ge Zhang, and Wenhu Chen. Pin: A knowledge-intensive dataset for paired and interleaved multimodal documents, 2024.

[29] Yang Luo, Zangwei Zheng, Zirui Zhu, and Yang You. How does the textual information affect the retrieval of multimodal in-context learning?, 2024.

[30] Yun-Da Tsai, Ting-Yu Yen, Pei-Fu Guo, Zhe-Yan Li, and Shou-De Lin. Text-centric alignment for multi-modality learning, 2024.

[31] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493, 2020.

[32] Shaeke Salman, Md Montasir Bin Shams, and Xiuwen Liu. Unaligning everything: Or aligning any text to any image in multimodal models, 2024.

[33] Guilherme Lourenço de Toledo and Ricardo Marcondes Marcacini. Transfer learning with joint fine-tuning for multimodal sentiment analysis, 2022.

[34] Weizhi Wang, Khalil Mrini, Linjie Yang, Sateesh Kumar, Yu Tian, Xifeng Yan, and Heng Wang. Finetuned multimodal language models are high-quality image-text data filters, 2024.

[35] Soyeon Caren Han, Feiqi Cao, Josiah Poon, and Roberto Navigli. Multimodal large language models and tunings: Vision, language, sensors, audio, and beyond, 2024.

AI-generated, for reference only.

www.SurveyX.cn

[36] Marco Bellagente, Manuel Brack, Hannah Teufel, Felix Friedrich, Björn Deiseroth, Constantin Eichenberg, Andrew Dai, Robert Baldock, Souradeep Nanda, Koen Oostermeijer, Andres Felipe Cruz-Salinas, Patrick Schramowski, Kristian Kersting, and Samuel Weinbach. Multifusion: Fusing pre-trained models for multi-lingual, multi-modal image generation, 2023.

[37] Yuan Gao, Sangwook Kim, David E Austin, and Chris McIntosh. Medbind: Unifying language and multimodal medical data embeddings, 2024.

[38] Yonghui Wang, Wengang Zhou, Hao Feng, and Houqiang Li. Adaptvision: Dynamic input scaling in mllms for versatile scene understanding, 2024.

[39] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2020, page 2359, 2020.

[40] Tianyi Bai, Hao Liang, Binwang Wan, Yanran Xu, Xi Li, Shiyu Li, Ling Yang, Bozhou Li, Yifan Wang, Bin Cui, Ping Huang, Jiulong Shan, Conghui He, Binhang Yuan, and Wentao Zhang. A survey of multimodal large language model from a data-centric perspective, 2024.

[41] Shubham Agarwal, Ondrej Dusek, Ioannis Konstas, and Verena Rieser. A knowledge-grounded multimodal search-based conversational agent, 2018.

[42] Vipin Vijayan, Braeden Bowen, Scott Grigsby, Timothy Anderson, and Jeremy Gwinnup. The case for evaluating multimodal translation models on text datasets, 2024.

[43] Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models, 2023.

[44] Ruiyi Zhang, Yanzhe Zhang, Jian Chen, Yufan Zhou, Jiuxiang Gu, Changyou Chen, and Tong Sun. Trins: Towards multimodal language models that can read, 2024.

[45] Ting Liu, Zunnan Xu, Yue Hu, Liangtao Shi, Zhiqiang Wang, and Quanjun Yin. Mapper: Multimodal prior-guided parameter efficient tuning for referring expression comprehension, 2025.

[46] Savya Khosla, Zhen Zhu, and Yifei He. Survey on memory-augmented neural networks: Cognitive insights to ai applications, 2023.

[47] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding, 2024.

[48] Minheng Ni, Yutao Fan, Lei Zhang, and Wangmeng Zuo. Visual-o1: Understanding ambiguous instructions via multi-modal multi-turn chain-of-thoughts reasoning, 2024.

[49] Adithya TG, Adithya SK, Abhinav R Bharadwaj, Abhiram HA, and Surabhi Narayan. Enhancing vision models for text-heavy content understanding and interaction, 2024.

[50] Zhengyuan Zhu, Daniel Lee, Hong Zhang, Sai Sree Harsha, Loic Feujio, Akash Maharaj, and Yunyao Li. Murar: A simple and effective multimodal retrieval and answer refinement framework for multimodal question answering, 2024.

[51] Shah Nawaz, Muhammad Kamran Janjua, Alessandro Calefati, and Ignazio Gallo. Revisiting cross modal retrieval, 2018.

[52] Cheng-An Hsieh, Cheng-Ping Hsieh, and Pu-Jen Cheng. Mr. right: Multimodal retrieval on representation of image with text, 2022.

[53] Zijia Zhao, Longteng Guo, Tongtian Yue, Sihan Chen, Shuai Shao, Xinxin Zhu, Zehuan Yuan, and Jing Liu. Chatbridge: Bridging modalities with large language model as a language catalyst, 2023.

AI-generated, for reference only.

www.SurveyX.cn

[54] Hanjia Lyu, Ryan Rossi, Xiang Chen, Md Mehrab Tanjim, Stefano Petrangeli, Somdeb Sarkhel, and Jiebo Luo. X-reflect: Cross-reflection prompting for multimodal recommendation, 2024.

[55] Seungwhan Moon, Satwik Kottur, Alborz Geramifard, and Babak Damavandi. Navigating connected memories with a task-oriented dialog system, 2022.

[56] Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, and Shafiq Joty. Retrieving multimodal information for augmented generation: A survey, 2023.

[57] Yue Yang, Joongwon Kim, Artemis Panagopoulou, Mark Yatskar, and Chris Callison-Burch. Induce, edit, retrieve: Language grounded multimodal schema for instructional video retrieval, 2021.

[58] Qilang Ye, Zitong Yu, Rui Shao, Xinyu Xie, Philip Torr, and Xiaochun Cao. Cat: Enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios, 2024.

[59] Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2787–2797, 2023.

[60] Maciej Pawłowski, Anna Wróblewska, and Sylwia Sysko-Romańczuk. Does a technique for building multimodal representation matter? – comparative analysis, 2022.

[61] Rong-Cheng Tu, Xian-Ling Mao, Bing Ma, Yong Hu, Tan Yan, Wei Wei, and Heyan Huang. Deep cross-modal hashing with hashing functions and unified hash codes jointly learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(2):560–572, 2020.

[62] Po han Li, Yunhao Yang, Mohammad Omama, Sandeep Chinchali, and Ufuk Topcu. Any2any: Incomplete multimodal retrieval with conformal prediction, 2024.

[63] Haochen Han, Minnan Luo, Huan Liu, and Fang Nan. A unified optimal transport framework for cross-modal retrieval with noisy labels, 2024.

[64] Sherzod Hakimov and David Schlangen. Images in language space: Exploring the suitability of large language models for vision language tasks, 2023.

[65] Sajal Aggarwal, Ananya Pandey, and Dinesh Kumar Vishwakarma. Modelling visual semantics via image captioning to extract enhanced multi-level cross-modal semantic incongruity representation with attention for multimodal sarcasm detection, 2024.

[66] Yuxin Peng, Jinwei Qi, Xin Huang, and Yuxin Yuan. Ccl: Cross-modal correlation learning with multigrained fusion by hierarchical network. *IEEE Transactions on Multimedia*, 20(2):405–420, 2017.

[67] Ann Speed. Assessing the nature of large language models: A caution against anthropocentrism, 2024.

[68] Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. Data augmentation using large language models: Data perspectives, learning paradigms and challenges, 2024.

[69] Marvin Voelter, Raheleh Hadian, Timotheus Kampik, Marius Breitmayer, and Manfred Reichert. Leveraging generative ai for extracting process models from multimodal documents, 2024.

[70] Hessa Abdulrahman Alawwad, Areej Alhothali, Usman Naseem, Ali Alkhathlan, and Amani Jamal. Enhancing textual textbook question answering with large language models and retrieval augmented generation, 2025.

[71] Yongquan Hu, Dawen Zhang, Mingyue Yuan, Kaiqi Xian, Don Samitha Elvitigala, June Kim, Gelareh Mohammadi, Zhenchang Xing, Xiwei Xu, and Aaron Quigley. Investigating the design considerations for integrating text-to-image generative ai within augmented reality environments, 2024.

AI-generated, for reference only.

[72] Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. Enhancing multimodal cooperation via sample-level modality valuation, 2024.

[73] Weihao Liu, Fangyu Lei, Tongxu Luo, Jiahe Lei, Shizhu He, Jun Zhao, and Kang Liu. Mmhqa-icl: Multimodal in-context learning for hybrid question answering over text, tables and images, 2023.

[74] Mauajama Firdaus, Avinash Madasu, and Asif Ekbal. A unified framework for slot based response generation in a multimodal dialogue system, 2023.

[75] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, page 6558, 2019.

[76] Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning, 2024.

[77] Ling Yang, Zhanyu Wang, Zhenghao Chen, Xinyu Liang, and Luping Zhou. Medxchat: A unified multimodal large language model framework towards cxrs understanding and generation, 2024.

[78] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Melissa Zhao, Aaron K Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, et al. A multimodal generative ai copilot for human pathology. *Nature*, 634(8033):466–473, 2024.

[79] Qu Yang, Mang Ye, and Bo Du. Emollm: Multimodal emotional understanding meets large language models, 2024.

[80] Cheng Tan, Jingxuan Wei, Zhangyang Gao, Linzhuang Sun, Siyuan Li, Ruifeng Guo, Bihui Yu, and Stan Z. Li. Boosting the power of small multimodal reasoning models to match larger models with self-consistency training, 2024.

[81] Suyash Vardhan Mathur, Jainit Sushil Bafna, Kunal Kartik, Harshita Khandelwal, Manish Shrivastava, Vivek Gupta, Mohit Bansal, and Dan Roth. Knowledge-aware reasoning over multimodal semi-structured tables, 2024.

[82] Large multimodal models: Notes o.

[83] Morris Alper and Hadar Averbuch-Elor. Emergent visual-semantic hierarchies in image-text representations, 2024.

[84] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024.

[85] Gijs Wijngaard, Elia Formisano, Bruno L. Giordano, and Michel Dumontier. Aces: Evaluating automated audio captioning models on the semantics of sounds, 2024.

[86] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning, 2023.

[87] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022.

[88] Jaehyuk Lim and Bruce W. Lee. Measuring agreeableness bias in multimodal models, 2024.

[89] Yi-Ting Yeh, Tzu-Chuan Lin, Hsiao-Hua Cheng, Yu-Hsuan Deng, Shang-Yu Su, and Yun-Nung Chen. Reactive multi-stage feature fusion for multimodal dialogue modeling, 2019.

AI-generated, for reference only.

www.surveyX.cn

[90] Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval, 2023.

[91] Qiying Yu, Yang Liu, Yimu Wang, Ke Xu, and Jingjing Liu. Multimodal federated learning via contrastive representation ensemble, 2023.

[92] Yangning Li, Yinghui Li, Xinyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, Hui Wang, Hai-Tao Zheng, Pengjun Xie, Philip S. Yu, Fei Huang, and Jingren Zhou. Benchmarking multimodal retrieval augmented generation with dynamic vqa dataset and self-adaptive planning agent, 2024.

www.SurveyX.cn

AI-generated, for reference only.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.

www.SurveyX.cn

AI-generated, for reference only.