Open Access Research Journal of
Science and Technology
Journals home page: https://oarjst.com/
ISSN: 2782-9960 (Online)

OARJ OPEN ACCESS RESEARCH JOURNALS

(REVIEW ARTICLE)

Check for updates

Multimodal Foundation Models for Unified Image, Video and Text Understanding

Ayodele R. Akinyele [1], Oseghale Ihayere [2], Osayi Eromhonsele [1], Ehisuoria Aigbogun [3], Adebayo Nurudeen Kalejaiye [4] and Oluwole Olakunle Ajayi [5, *]

[1] Kenan-Flagler Business School, University of North Carolina at Chapel Hill, North Carolina, USA.
[2] Fuqua School of Business, Duke University, Durham, North Carolina, USA.
[3] Booth School of Business, University of Chicago, Illinois, USA.
[4] Scheller College of Business, Georgia Institute of Technology, Georgia, USA.
[5] Community and Program Specialist, UHAI For Health Inc, Worcester, Massachusetts, USA.

Open Access Research Journal of Science and Technology, 2024, 12(02), 081–093

Publication history: Received on 08 October 2024; revised on 16 November 2024; accepted on 19 November 2024

Article DOI: https://doi.org/10.53022/oarjst.2024.12.2.0139

Abstract

Advanced models can now interpret and understand photos, videos, and text. Conventional AI models focused on image classification, text analysis, and video processing. Multimodal foundation models combine data analysis in one framework to meet the requirement for more integrated AI systems. The models may learn joint representations from several modalities to generate text from images, analyze movies with textual context, and answer visual queries. Cross-modal learning's theoretical foundations and architectural advances have helped multimodal foundation models grow rapidly. This study explores their success. Transformer-based architectures have profoundly changed AI model data modalities. Self-attention and contrastive learning help the models align and integrate data across modalities, improving data understanding. The study analyses well-known multimodal models CLIP, ALIGN, Flamingo, and Video BERT, emphasizing on their design, training, and performance across tasks. Their performance in caption generation, video-text retrieval, and visual reasoning has led to more adaptable AI systems that can handle complex real-world scenarios. Despite promising results, multimodal learning faces various obstacles. To create effective models, you need substantial, high-quality datasets. Computers struggle to handle many data formats simultaneously, and bias and interpretability difficulties arise. The limitations and ethical implications of multimodal models in healthcare and autonomous systems are examined in this research. This study investigates the future of multimodal foundation models, focused on reducing processing, enhancing model fairness, and applying them to audio, sensor data, and robotics. Understanding and integrating multimodal information is essential for creating more intuitive and intelligent systems, therefore unified multimodal models could change human-computer interaction. Overall, multimodal foundation models drive the search for generalized and adaptable AI systems. These models' capacity to combine picture, video, and text data could alter many applications, driving creativity across sectors and stimulating AI research.

Keywords: Multimodal; Artificial Intelligence; Transformer-based architectures; Interactions; Models

1. Introduction

With the rapid growth of digitalization, there has been a considerable increase in the quantity and type of data that has been generated across a wide range of platforms, devices, and domains. According to Javaid (2024), there is a demand for artificial intelligence (AI) models that can execute rapid analysis and comprehension of many types of data, including text, photos, videos, audio, and other types of data. These single-modality models have demonstrated effectiveness in their respective domains; however, real-world scenarios frequently involve intricate associations between different data types, which renders them unsuitable for tasks that require a comprehensive understanding of multimodal inputs (Liang et al., 2024). The adoption of artificial intelligence in multiple domains has rapidly grown in recent years (Akinyele et al., 2024, Mudele et al., 2019, Mudele et al., 2021a, Mudele et al., 2021b).

[*] Corresponding author: Oluwole Olakunle Ajayi

Copyright © 2024 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution Liscense 4.0.

Most artificial intelligence systems focus on a single mode, like computer vision for classifying images or natural language processing for analyzing writing. These models can find new skills and make tasks easier because they combine different types of strengths found in data. Textual and visual data can be used together to help people better understand what they are seeing in movies or write more accurate picture captions. Gupta et al. (2024) say that a unified method to multimodal learning is a big part of how artificial intelligence studies are moving forward. This is because it can link a lot of different types of data and make AI systems that are better at adapting to different situations. So, this study looks into the future of multimodal foundation models with the goal of making them fairer, lowering processing, and using them with audio, sensor data, and robotics.

2. Methodology

A systematic search strategy and five scientific databases (PubMed, Google Scholar, Scopus, IEEE, and Science Direct) were used to find research papers related to Multimodal Foundation Models for Unified Image, Video, and Text Understanding (Zhao et al., 2020). There were also meeting proceedings, books, dissertations, and master's theses. What were typed into the search engine include "Multimodal Foundations", "Unified Image, Unified Video, and Text Understanding." For this study, an extensive list of abstracts was compiled and carefully read. Any publications that met the selection criteria were carefully investigated. The study included all papers that came out up until 2024.

3. Results

3.1. Overview of Multimodal Learning

Building models that can handle, combine, and evaluate input from different modes is what multimodal learning is all about. As Duan et al. (2024) say, this method goes beyond single-modality AI models because it combines text, images, videos, audio, and sensor data into a single representation. Because they have a solid understanding of the subject, models can look for images based on textual queries, explain visuals, and answer questions about videos. There are many reasons why multimodal learning is very important in artificial intelligence. Using a variety of methods together helps models understand things better. When you combine MRI scans with written patient records, you get more contextual information that one modality might not have (Xu et al., 2024) that makes medical imaging findings more accurate. Additionally, multimodal models show better performance across a range of jobs. The model can successfully handle new inputs when representations are moved between different modalities (Zhang et al., 2020). Multimodal systems are also more reliable in real life because they can use information from another modality to make up for noisy or lost data in one modality.

If an artificial intelligence model can't tell what something is by looking at it, it might use text or voice prompts to figure out what it is. For the same reason, not having to train models for every mode speeds up development and makes it easier to use AI in many different areas and tasks (Baduge et al., 2022). In the end, the progress made in bidirectional learning could help solve more complicated problems in the real world than old models. The field has progressed significantly with the implementation of transformer architectures like Flamingo and CLIP, which adeptly handle both textual and visual data (Chen et al., 2024). Integration of learning across different modalities is a necessary first step toward creating general-purpose AI systems that can think like humans.

3.1.1. Applications: Image Captioning, Video Analysis, and Text-Based Image Retrieval

The applications of multimodal foundation models span across multiple domains, with a broad range of use cases that leverage the integrated understanding of different data types. Some of the most prominent applications include:

Image captioning generates informative text from an image. Multimodal models learn to analyze image content and write suitable text. CLIP and Flamingo have generated outstanding captions by learning image-text representations (Ghandi et al., 2023). These models recognize visual items and infer relationships, resulting in more contextually correct descriptions. Image captioning helps visually challenged people interpret photos and organizes and retrieves large image collections in stock photography marketplaces and content management systems (Al-Malla et al., 2022).

Video analysis is complicated and requires visual and temporal knowledge. Video BERT and other multimodal models use frames and captions, subtitles, and metadata to recognize activities, segment scenes, and summarize videos (Rafiq et al., 2021). Video question-answering tasks need profound cross-modal thinking as the AI analyses a video clip and answers questions depending on its content. Video analysis is useful in entertainment, surveillance, and education. Media production models can classify sequences or generate highlights based on audience preferences. Video-based learning platforms can personalize video content and create quizzes (Bulathwela, 2023).

Text-based image retrieval use queries to find images. Learning shared text and image representations allows multimodal foundation models to do this. This application is valuable in e-commerce, where customers search for products using descriptive phrases (Zhang et al., 2022). It searches and categorizes massive volumes of visual data using textual queries in content moderation and media curation.

3.2. Single-Modality vs. Multimodal Models: Historical Focus and the Shift Towards Multimodal Models

AI models have mostly been developed to solve problems with a single mode. Xu et al. (2024) say that models must be made that are experts at handling and understanding text, images, and audio data. Early text-only models, such as Bag-of-Words (BoW), and later NLP models, such as Word2Vec and GloVe, noted word embeddings and changed the meanings of words based on the text around them (Incitti et al., 2023). By learning spatial hierarchies in visual data, convolutional neural networks (CNNs) changed the way images are identified in computer vision. For example, Liu et al. (2024) found that single-modality models were great at their own tasks but not so good at multi-modality tasks.

Real-world scenarios rarely involve isolated modalities. People and robots use numerous data types in real life. Images typically have text, films are visual and auditory, etc. Complex tasks revealed single-modality model limitations. An AI reading a video scene may need to consider both the visual aspects and the audio or textual metadata to completely understand the context (Shoumy et al., 2020). This has led to the use of multimodal models, which unify understanding across data kinds. Multimodal models learn joint representations that combine features from various modalities for deeper, context-aware processing. Multimodal learning is based on the idea that integrating diverse types of data helps us comprehend more since it mimics how humans interpret sensory inputs. Multimodal models are used in image captioning, visual question answering, and video analysis (Liang et al., 2024). Research has shown that mixing input from many modalities increases task performance and opens new possibilities that single-modality models cannot (Proca et al., 2024).

3.3. Multimodal Learning Paradigms: Early Fusion vs. Late Fusion

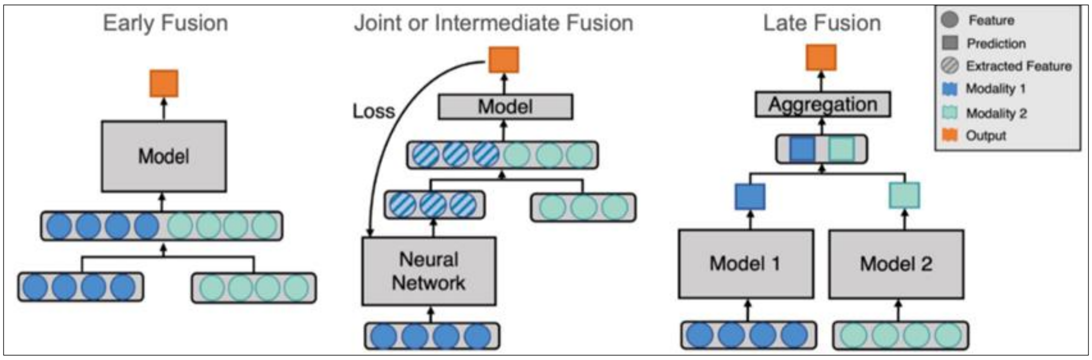

Figure 1 Data Modality Fusion Strategies (Adopted from Huang et al. 2020)

Multimodal learning can be categorized into different paradigms based on when and how the integration of modalities occurs within the model. The two primary approaches are early fusion and late fusion, each with its own advantages and challenges.

Early Fusion (Feature-Level Integration) integrates various modalities at the feature level. Before passing through the model, data from each modality is mixed at the start of the learning process (Boulahia et al., 2021). Pixel values from images and word embeddings from text are concatenated into a single representation in early fusion models. The model learns from this joint feature space to capture early correlations and interactions between modalities.

Late Fusion (Decision-Level Integration) delays modality integration till decision time. This method processes each modality separately through its own pipeline and combines its outputs or predictions at the end to decide (Pawłowski et al., 2023). In an image-text task, an image-processing network and a text-processing network would separately make predictions, which would be combined via weighted averaging or attention-based aggregation.

3.4. The Role of Transformers: Impact on Multimodal Tasks and Cross-Modal Attention Mechanisms

Multimodal learning is all about making models that can take in, combine, and judge information from different learning styles. Duan et al. (2024) say that this method is better than single-modality AI models because it makes a single representation out of text, photos, videos, audio, and sensor data. Models can look for pictures based on text queries, explain images, and answer questions about videos because they know a lot about the subject. Learning in more than one way is very important in artificial intelligence for many reasons.

Models can better understand things when they use a number of different ways together. When you put together MRI scans and written patient records, you get more contextual information that one method might not have (Xu et al., 2024) that helps you get more accurate results from medical imaging. Additionally, mixed models work better for many different types of tasks. It is possible for the model to handle new inputs when representations are changed between senses (Zhang et al., 2020). It's also safer to use multimodal systems in real life because they can use data from another modality to make up for data that is lost or busy in one modality.

If a computer program can't figure out what something is just by looking at it, it may use voice or text hints to do so. In the same way, not having to train models for each mode speeds up progress and makes it easier to use AI in lots of different situations (Baduge et al., 2022). Finally, the progress made in two-way learning could help real-world problems that are more difficult to answer with old models. Transformer designs like Flamingo and CLIP, which can handle both textual and visual data well, have made a lot of progress in the field (Chen et al., 2024). Putting together different ways of learning is the first thing that needs to be done to make general-purpose AI systems that can think like people.

Another example is Flamingo, which enhances the architecture to handle both images and sequences of text, thereby leveraging the transformer's ability to process multi-modal inputs (Binte Rashid et al., 2024). Flamingo employs cross-modal attention to effectively capture the interactions between images and text in tasks such as visual question answering, integrating visual tokens with textual tokens (Lin et al., 2023). The adaptability of transformers, particularly in modelling intricate relationships among various data types, has sparked advancements in cross-modal learning. Transformers allow for dynamic attention across different modalities, facilitating the integration and alignment of multimodal inputs. For roles that require both surface-level modality matching and profound semantic comprehension along with data type reasoning, they are ideal (Bhattacharyya, 2023). Moreover, the emergence of transformer-based multimodal models has contributed to enhancing scalability and generalization. The flexibility of these models for real-world applications is notable, as extensive pretraining on large multimodal datasets, combined with fine-tuning for specific downstream tasks, has enabled them to generalize effectively across various tasks and domains (Han et al., 2021).

4. Key Architectural Advancements

Multimodal foundation models have advanced to create a unique architectural advance that seamlessly integrate data from images, videos, and text. These architectural enhancements have enabled models to handle different tasks, cross-modal thinking, and complicated real-world situations (Chen et al., 2024). This section covers multimodal model design advances such unified model design, pretraining tactics, and fine-tuning methods that have led to world-class performance.

4.1. Unified Model Design

The key innovation behind multimodal foundation models is their unified design, which allows the model to simultaneously process and integrate visual, textual, and sometimes temporal information (Liang et al., 2024). Unified models aim to learn joint representations from multiple modalities, meaning that the model can understand how different forms of data (e.g., images and text) relate to one another (Xiong et al., 2022). This capability is crucial for tasks like image captioning, video analysis, and text-based image retrieval, which require reasoning across different types of inputs.

4.1.1. Vision-Language Models (VLMs): CLIP and Beyond

Multimodal Vision-Language Models (VLMs) align visual and textual data. These models have revolutionized how humans perceive words and visual material by mapping images and text into a shared embedding space (Ghosh et al., 2024). One famous example is OpenAI's CLIP (Contrastive Language-Image Pretraining). CLIP pretrains on a vast dataset of images and text descriptions to match text and images. CLIP's contrastive learning aim, where the model learns to associate matching pairs of images and text (images and captions) while discriminating non-matching pairings, is its main novelty (Shi et al., 2024). ALIGN (Learning from Noisy Text) aligns visual and textual representations using contrastive learning like CLIP. ALIGN outperforms CLIP in cross-modal tasks by pretraining with larger datasets. The approach is adaptable because large-scale pretraining allows it to generalize across more ideas (Jin et al., 2023).

4.1.2. Video-Text Models: Handling Temporal Information

VLMs like CLIP excel at image-text tasks, whereas video-text models are built to handle video data's temporal dimension. Video content includes both frames and continuous occurrences. This adds complexity since the model must learn to capture frame temporal correlations and align them with textual descriptions or enquiries (Madan et al., 2024). Video BERT is a popular video-text model that applies transformer-based text modelling to video understanding. Video BERT interprets video frames as "tokens" in a sequence, way NLP models do text (Khan, 2023). Visual tokens and textual tokens (e.g., speech transcriptions or subtitles) are learnt together to generate joint representations. For action recognition, video summarization, and video question answering, Video BERT's capacity to learn temporal dependencies is its main architectural achievement (Shakil et al., 2024).

4.2. Pretraining and Fine-tuning Techniques

Pretraining on large datasets has become a central part of developing multimodal models. This approach involves training a model on massive amounts of data to learn general-purpose representations that can later be fine-tuned for specific downstream tasks (Han et al., 2021). Pretraining methods in multimodal models often rely on techniques like masked modelling, which allow the model to capture relationships between modalities in a self-supervised or semi-supervised manner (Zong et al., 2024).

Multimodal models pretrain with masked modelling. This method is influenced by NLP models like BERT, which mask particular words in a phrase and train the machine to anticipate them based on context (Bugliarello et al., 2021). Masked modelling is extended to include visual and textual data in multimodal learning to build representations that integrate both modalities. Multimodal model Flamingo pretrains on images and text via masked modelling (Sun et al., 2023). Pretraining in Flamingo masks text and picture input. The algorithm is trained to predict masked sections using data from both modalities. This method teaches the model to fill in gaps and capture intricate text-image interactions. The model can anticipate what a masked image should contain using the associated text. If text is masked, the model can deduce the missing information from the image (Alhabeeb and Al-Shargabi, 2024).

5. Applications of Multimodal Models

Multimodal models have shown remarkable versatility across a wide range of applications that require the integration of visual, textual, and sometimes temporal information (Bayoudh et al., 2022). From improving human-computer interaction in the form of visual question answering (VQA) to enabling highly accurate caption generation and enhancing cross-modal search and retrieval, these models are reshaping how AI interacts with and interprets multimodal data. Below, some of the key applications in discussed in detail (Lu et al., 2023).

5.1. Visual Question Answering (VQA)

Images and natural language processing are used in visual question answering (VQA). In VQA, the model is shown an image and a text-based inquiry on its content (Wu et al., 2017). The model must assess the image, comprehend the inquiry, and respond appropriately. This task is difficult since the model must reason across vision and language and often complicated object relations, scene understanding, or background knowledge (Ali et al., 2023). Visual BERT and ViLBERT use transformers to merge visual and textual information, setting new standards for VQA activities. Visual BERT adds image features to the text-only BERT architecture (Manzoor et al., 2023). However, ViLBERT (Vision-and-Language BERT) merges visual and textual data from two streams using attention techniques. This architecture lets the model acquire relationships between image items and their textual descriptions, boosting its capacity to answer difficult visual enquiries (Khan et al., 2022). VQA models are used in assistive technologies for the visually impaired, automated customer care, and interactive educational systems where users may ask questions about images or videos and get thorough answers (de Freitas et al., 2022).

5.2. Caption Generation

Caption generation is another critical application of multimodal models, where the goal is to generate descriptive text from visual data, be it an image or video. The challenge here lies in capturing the essential elements of the visual content and articulating them in grammatically correct and semantically rich sentences. Models must not only recognize objects and actions but also understand their relationships and contextual significance (Wang et al., 2020).

5.3. Cross-Modal Search and Retrieval

One of the most practical and impactful applications of multimodal models is their ability to perform cross-modal search and retrieval tasks. These models allow for efficient search across different data types, such as using text to search for images or videos, or generating textual summaries based on visual input (Beltrán et al., 2021). Cross-modal search is critical for applications like search engines, content recommendation systems, and media libraries, where users need to retrieve relevant information across modalities quickly and accurately (Wang et al., 2016).

6. Challenges and Limitations

While multimodal foundation models have made significant advances in unifying the understanding of image, video, and text data, several challenges remain. These challenges stem from the complex nature of working across multiple modalities, as well as the high demands placed on data, computational resources, and model interpretability (Chen et al., 2024).

6.1. Data Requirements

One of the most significant challenges in building effective multimodal models is the sheer scale and quality of data required for training (Liang et al., 2021). Multimodal models, especially those aimed at image, video, and text understanding, need massive, diverse datasets to capture the richness and variety across different modalities (Zhang et al., 2024).

6.1.1. Large-scale Datasets

For multimodal models to perform well across diverse tasks, they must be pretrained on enormous datasets. Leading models like CLIP and ALIGN rely on hundreds of millions, if not billions, of image-text pairs sourced from the internet (Zhang et al., 2024). Datasets like JFT-300M and LAION have become foundational to the training of these models (Takashima et al., 2023). For instance, JFT-300M contains hundreds of millions of labelled images, while LAION offers an open-source, large-scale dataset of image-text pairs designed specifically for training vision-language models (Desai, 2024). However, curating such large datasets presents several challenges. First, acquiring this data at scale requires significant resources in terms of both storage and infrastructure. In addition, these large datasets must capture the full spectrum of visual and linguistic diversity, covering various domains, languages, objects, and activities. Training multimodal models on smaller or less diverse datasets would limit their ability to generalize across different tasks and contexts (Fung et al., 2023).

6.1.2. Data Biases

Another critical issue related to data is bias. Many of the large-scale datasets used to train multimodal models are scraped from the internet, which introduces inherent biases in terms of geography, culture, and social norms (Watson et al., 2023). These biases can significantly impact the fairness and accuracy of the model, especially when deployed in real-world applications. For example, models trained on internet-based datasets might over-represent certain demographics, leading to skewed outcomes in tasks like facial recognition or caption generation (Rice et al., 2019). Gender, racial, and cultural biases are common in multimodal datasets. If a model trained on biased data is tasked with generating captions for images or answering questions about visual content, it may perpetuate harmful stereotypes or produce inaccurate results (Hirota et al., 2022). These biases not only affect the fairness of AI systems but also raise ethical concerns about the widespread deployment of such technologies. Addressing these biases requires careful dataset curation and post-processing techniques, such as data augmentation and bias mitigation strategies, to ensure that the models perform equitably across all user groups (Siddique et al., 2023).

6.2. Model Interpretability

One of the biggest hurdles in deploying multimodal foundation models is their inherent black-box nature. These models often operate as highly complex systems, with millions or billions of parameters that make their decision-making processes difficult to interpret (Hassija et al., 2024). As AI systems are increasingly applied to high-stakes domains such as healthcare, autonomous driving, and legal systems there is growing demand for more transparent and interpretable models (Chau, 2024).

7. Evaluation Metrics for Multimodal Models

Multimodal foundation model evaluation is complicated. Multimodal models must examine how well they integrate and reason across modalities, unlike single-modality models that only need task-specific metrics (Wang et al., 2024). Evaluation often considers task-specific performance, cross-modal consistency, and generalization to new tasks or datasets.

7.1. Task-specific Metrics

The activities multimodal models accomplish are evaluated using specified metrics to measure their efficacy. Multimodal models can do caption creation, visual question answering (VQA), and cross-modal retrieval, hence their assessment metrics vary by task (Xi et al., 2020). Accuracy, precision, recall, and F1-score are commonly used for VQA, picture classification, and object detection. These metrics show how well the model predicts or detects meaningful information across modalities. Accuracy is the percentage of model predictions that are correct. VQA measures how often the model answers a question correctly based on visual input (Sharma and Jalal, 2021). Precision measures how well the model makes favourable predictions, whereas recall measures its capacity to catch all relevant instances. Precision and recall are combined into the F1-score to quantify the model's performance, especially in class imbalance tasks like spotting rare items in photos (Juba and Le, 2019).

7.2. BLEU and METEOR for Caption Generation

Specified criteria are used to evaluate the quality and fluency of natural language descriptions generated by the model for image and video captioning. Bilingual Evaluation Understudy (BLEU) and METEOR are common metrics in this domain (de Souza Inácio and Lopes, 2023). BLEU measures the overlap between model-generated captions and human-written reference captions. BLEU counts the number of n-grams (word sequences) in the generated caption that match the reference captions (Rafiq et al., 2021). BLEU is a popular metric for machine translation and caption generation; however, it doesn't account for synonyms or text fluency. BLEU scores precise matches, but METEOR scores synonyms and paraphrases to provide a more sophisticated assessment of created text (Callison-Burch et al., 2006). METEOR scores fluency and semantic accuracy by stemming, synonym matching, and paraphrasing generated captions against reference captions. Captioning tasks are evaluated using BLEU, METEOR, CIDEr, and ROUGE (Khurana and Deshpande, 2021). CIDEr focusses on the consensus of numerous reference captions to match human caption quality judgements, while ROUGE measures recall by counting reference n-grams in the generated caption (de Souza Inácio and Lopes, 2023). These metrics are useful, but they cannot capture sophisticated multimodal learning characteristics like semantic understanding and contextual reasoning, which are critical for video captioning and VQA.

8. Future Directions

As multimodal foundation models continue to gain prominence, significant research and development opportunities lie ahead, particularly in broadening their capabilities, improving efficiency, and addressing ethical concerns. These future directions highlight potential advancements that could reshape how multimodal AI is applied across industries and scientific disciplines.

8.1. Unified Models for More Modalities

8.1.1. Beyond Image, Video, and Text

Presently, multimodal foundation models prioritize images, videos, and text. In the domains of picture captioning, video understanding, and text-based retrieval, current methods have proven adequate; however, there is an increasing interest in integrating additional modalities to create more comprehensive artificial intelligence systems (Chen et al., 2024). The incorporation of audio would enhance the interpretation of multimodal videos and facilitate the generation of content driven by speech. A model capable of recognizing both audio and video have the potential to enhance the richness of video summaries and improve voice recognition by integrating visual signals (Sulubacak et al., 2020).

Future multimodal models may require the integration of environmental sensor data, lidar, and radar technologies. This has the potential to revolutionize intelligent environments and self-driving vehicles, where immediate decision-making necessitates the analysis of intricate data flows from various origins (Alabyad et al., 2024). The integration of these data formats, alongside images, videos, and text, would facilitate the development of more complex and contextualized models capable of managing real-world intricacies (Yan et al., 2024). The integration of wearable sensor data, patient histories, and medical imaging has the potential to improve patient monitoring systems, tailor treatment plans, and refine diagnosis (Singh et al., 2024). In medical fields, the potential of multimodal learning is highly encouraging. Integrating multiple data sources into a single model can enhance the precision and effectiveness of healthcare delivery (Shaik et al., 2023; Ajayi et al., 2024; Davies et al., 2024).

8.1.2. Multimodal Foundation Models in Robotics

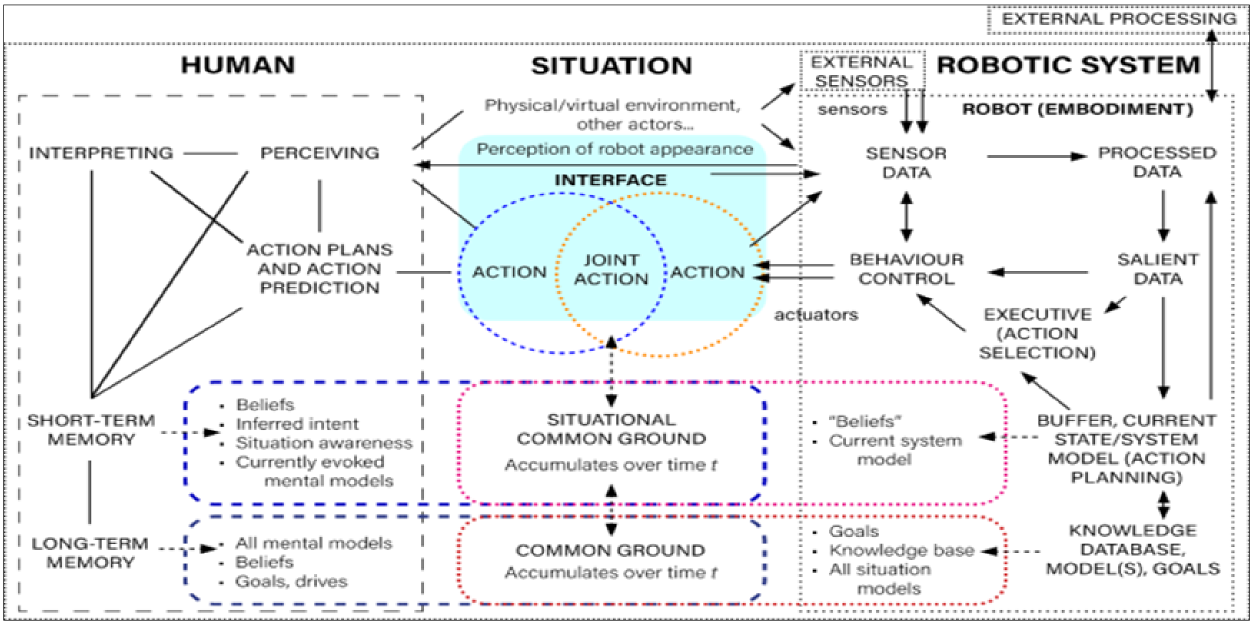

Figure 2 Robotics related interactions (Frijns et al., 2023)

The exploration of multimodal models in the realm of future robotics presents a thrilling opportunity. Robots are required to analyze and integrate multiple data streams to navigate intricate environments (Wang et al., 2024). The enhancement of robotic comprehension and reaction to auditory, gestural, and visual directives may be achieved through the integration of cohesive multimodal models. To determine the suitable response, smart home robots are capable of interpreting vocal commands, recognizing gestures, and visually assessing their environment (Saunders et al., 2015).

By integrating camera, radar, and lidar data alongside map annotations, multimodal models have the potential to enhance the navigation capabilities of autonomous vehicles and drones, enabling them to avoid obstacles and improve their understanding of the environment (Yeong et al., 2021). Unified models may facilitate the development of robots capable of executing intricate tasks such as search-and-rescue, by allowing them to make decisions in unpredictable and dynamic environments through the integration of multimodal input streams (Queralta et al., 2020).

9. Conclusion

Multimodal foundation models let machines to process and understand images, videos, and text in a single framework, advancing artificial intelligence. This review has examined the architectural improvements and methods that have advanced AI research with these models. Transformers, contrastive learning, and pretraining procedures like masked modelling helped transform single-modality models into multimodal systems. CLIP, ALIGN, Flamingo, and Video BERT demonstrate multimodal learning's potential by performing challenging image classification, video analysis, and cross-modal retrieval tasks. Despite these advances, the field faces substantial obstacles. Large datasets, model interpretability, and multimodal model training computational demands hinder acceptance and deployment. Additionally, bias, justice, and multimodal AI ethics must be addressed.

Future research in multimodal foundation models is promising. Multimodal AI applications will grow as AI systems include audio, sensor, and robotic inputs. Effective model designs, knowledge distillation, and transfer learning will make these models more scalable and useful by more organizations. To develop trust in these technologies, bias mitigation and ethical, transparent AI systems must be prioritized. In conclusion, multimodal foundation models have transformed AI by providing a more comprehensive and integrated view of data across modalities. As research improves these models, they will unlock the full potential of unified picture, video, and text interpretation, enabling a new generation of intelligent devices that seamlessly interact with the world.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Ajayi, O. O., Wright-Ajayi, B., Mosaku, L. A., Davies, G. K., Moneke, K. C., & Adeleke, O. R. Enhancing Infectious Disease Management in Nigeria: The Role of Artificial Intelligence in Diagnosis and Treatment. *Clin Case Rep Int. 2024; 8, 1670*.

[2] Ajayi, O. O., Wright-Ajayi, B., Mosaku, L. A., Davies, G. K., Moneke, K. C., Adeleke, O. R., ... & Mudele, O. (2024). Application of satellite imagery for vector-borne disease monitoring in sub-Saharan Africa: An overview. *GSC Advanced Research and Reviews*, *18*(3), 400-411.

[3] Akinyele, A.R., Ajayi, O.O., Munyaneza, G., Ibecheozor, U.H. and Gopakumar, N., (2024). Leveraging Generative Artificial Intelligence (AI) for cybersecurity: Analyzing diffusion models in detecting and mitigating cyber threats. GSC Advanced Research and Reviews, 21(2), pp.001-014.

[4] Alabyad, N., Hany, Z., Mostafa, A., Eldaby, R., Tagen, I. A., & Mehanna, A. (2024, March). From Vision to Precision: The Dynamic Transformation of Object Detection in Autonomous Systems. In *2024 6th International Conference on Computing and Informatics (ICCI)* (pp. 332-344). IEEE.

[5] Alhabeeb, S. K., & Al-Shargabi, A. A. (2024). Text-to-Image Synthesis With Generative Models: Methods, Datasets, Performance Metrics, Challenges, and Future Direction. *IEEE Access*.

[6] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., ... & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information fusion*, *99*, 101805.

[7] Al-Malla, M. A., Jafar, A., & Ghneim, N. (2022). Image captioning model using attention and object features to mimic human image understanding. *Journal of Big Data*, *9*(1), 20.

[8] Baduge, S. K., Thilakarathna, S., Perera, J. S., Arashpour, M., Sharafi, P., Teodosio, B., ... & Mendis, P. (2022). Artificial intelligence and smart vision for building and construction 4.0: Machine and deep learning methods and applications. *Automation in Construction*, *141*, 104440.

[9] Bayoudh, K., Knani, R., Hamdaoui, F., & Mtibaa, A. (2022). A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, *38*(8), 2939-2970.

[10] Beltrán, L. V. B., Caicedo, J. C., Journet, N., Coustaty, M., Lecellier, F., & Doucet, A. (2021). Deep multimodal learning for cross-modal retrieval: One model for all tasks. *Pattern Recognition Letters*, *146*, 38-45.

[11] Bhattacharyya, A. (2023). *Multi-Modal Semantic Role Labeling and Its Application* (Doctoral dissertation, University of Colorado at Boulder).

[12] Binte Rashid, M., Rahaman, M. S., & Rivas, P. (2024). Navigating the Multimodal Landscape: A Review on Integration of Text and Image Data in Machine Learning Architectures. *Machine Learning and Knowledge Extraction*, *6*(3), 1545-1563.

[13] Boulahia, S. Y., Amamra, A., Madi, M. R., & Daikh, S. (2021). Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Machine Vision and Applications*, *32*(6), 121.

[14] Bugliarello, E., Cotterell, R., Okazaki, N., & Elliott, D. (2021). Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics*, *9*, 978-994.

[15] Bulathwela, M. S. S. (2023). *Novel datasets, user interfaces and learner models to improve learner engagement prediction on educational videos* (Doctoral dissertation, UCL (University College London)).

[16] Callison-Burch, C., Osborne, M., & Koehn, P. (2006, April). Re-evaluating the role of BLEU in machine translation research. In *11th conference of the european chapter of the association for computational linguistics* (pp. 249-256).

[17] Chau, H. M. (2024). Developing Interpretable and Explainable AI Models for High-stakes Decision Making in Societal Contexts. *Journal of Sustainable Urban Futures*, *14*(1), 13-26.

[18] Chen, Z., Xu, L., Zheng, H., Chen, L., Tolba, A., Zhao, L., ... & Feng, H. (2024). Evolution and Prospects of Foundation Models: From Large Language Models to Large Multimodal Models. *Computers, Materials & Continua*, *80*(2).

[19] Chen, Z., Xu, L., Zheng, H., Chen, L., Tolba, A., Zhao, L., ... & Feng, H. (2024). Evolution and Prospects of Foundation Models: From Large Language Models to Large Multimodal Models. *Computers, Materials & Continua*, *80*(2).

[20] Chen, Z., Xu, L., Zheng, H., Chen, L., Tolba, A., Zhao, L., ... & Feng, H. (2024). Evolution and Prospects of Foundation Models: From Large Language Models to Large Multimodal Models. *Computers, Materials & Continua*, *80*(2).

[21] Davies, G. K., Ajayi, O. O., Wright-Ajayi, B., Mosaku, L. A., Moneke, K. C., Adeleke, O. R., ... & Mudele, O. (2024). Unravelling the complexity of environmental exposures and health: A novel exposome-centered framework for occupational and environmental epidemiology. *GSC Advanced Research and Reviews*, *19*(1), 026-032.

[22] de Freitas, M. P., Piai, V. A., Farias, R. H., Fernandes, A. M., de Moraes Rossetto, A. G., & Leithardt, V. R. Q. (2022). Artificial intelligence of things applied to assistive technology: a systematic literature review. *Sensors*, *22*(21), 8531.

[23] de Souza Inácio, A., & Lopes, H. S. (2023). Evaluation metrics for video captioning: A survey. *Machine Learning with Applications*, *13*, 100488.

[24] de Souza Inácio, A., & Lopes, H. S. (2023). Evaluation metrics for video captioning: A survey. *Machine Learning with Applications*, *13*, 100488.

[25] Desai, K. P. (2024). *Language Supervision for Computer Vision* (Doctoral dissertation).

[26] Dobrzycki, A. D., Bernardos, A. M., Bergesio, L., Pomirski, A., & Sáez-Trigueros, D. (2023). Exploring the Use of Contrastive Language-Image Pre-Training for Human Posture Classification: Insights from Yoga Pose Analysis. *Mathematics*, *12*(1), 76.

[27] Duan, J., Xiong, J., Li, Y., & Ding, W. (2024). Deep learning based multimodal biomedical data fusion: An overview and comparative review. *Information Fusion*, 102536.

[28] Frijns, H. A., Schürer, O., & Koeszegi, S. T. (2023). Communication models in human–robot interaction: an asymmetric MODel of ALterity in human–robot interaction (AMODAL-HRI). *International Journal of Social Robotics*, *15*(3), 473-500.

[29] Fung, A., Benhabib, B., & Nejat, G. (2023). Robots autonomously detecting people: A multimodal deep contrastive learning method robust to intraclass variations. *IEEE Robotics and Automation Letters*, *8*(6), 3550-3557.

[30] Ghandi, T., Pourreza, H., & Mahyar, H. (2023). Deep learning approaches on image captioning: A review. *ACM Computing Surveys*, *56*(3), 1-39.

[31] Ghosh, A., Acharya, A., Saha, S., Jain, V., & Chadha, A. (2024). Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214*.

[32] Gupta, P., Ding, B., Guan, C., & Ding, D. (2024). Generative AI: A systematic review using topic modelling techniques. *Data and Information Management*, 100066.

[33] Han, X., Wang, Y. T., Feng, J. L., Deng, C., Chen, Z. H., Huang, Y. A., ... & Hu, P. W. (2023). A survey of transformer-based multimodal pre-trained modals. *Neurocomputing*, *515*, 89-106.

[34] Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., ... & Zhu, J. (2021). Pre-trained models: Past, present and future. *AI Open*, *2*, 225-250.

[35] Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., ... & Zhu, J. (2021). Pre-trained models: Past, present and future. *AI Open*, *2*, 225-250.

[36] Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., ... & Hussain, A. (2024). Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, *16*(1), 45-74.

[37] Hazmoune, S., & Bougamouza, F. (2024). Using transformers for multimodal emotion recognition: Taxonomies and state of the art review. *Engineering Applications of Artificial Intelligence*, *133*, 108339.

[38] Hirota, Y., Nakashima, Y., & Garcia, N. (2022, June). Gender and racial bias in visual question answering datasets. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1280-1292).

[39] Huang, S. C., Pareek, A., Seyyedi, S., Banerjee, I., & Lungren, M. P. (2020). Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ digital medicine*, *3*(1), 136.

[40] Incitti, F., Urli, F., & Snidaro, L. (2023). Beyond word embeddings: A survey. *Information Fusion*, *89*, 418-436.

[41] Javaid, M., Haleem, A., Singh, R. P., & Sinha, A. K. (2024). Digital economy to improve the culture of industry 4.0: A study on features, implementation and challenges. *Green Technologies and Sustainability*, 100083.

[42] Jin, Y., Li, Y., Yuan, Z., & Mu, Y. (2023). Learning instance-level representation for large-scale multi-modal pretraining in e-commerce. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11060-11069).

[43] Juba, B., & Le, H. S. (2019, July). Precision-recall versus accuracy and the role of large data sets. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 4039-4048).

[44] Khan, S. (2023). *Evaluating state-of-the-art vision-language models for video recognition on real world dataset* (Master's thesis, S. Khan).

[45] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. *ACM computing surveys (CSUR)*, *54*(10s), 1-41.

[46] Khurana, K., & Deshpande, U. (2021). Video question-answering techniques, benchmark datasets and evaluation metrics leveraging video captioning: a comprehensive survey. *IEEE Access*, *9*, 43799-43823.

[47] Liang, P. P., Lyu, Y., Fan, X., Wu, Z., Cheng, Y., Wu, J., ... & Morency, L. P. (2021). Multibench: Multiscale benchmarks for multimodal representation learning. *Advances in neural information processing systems*, *2021*(DB1), 1.

[48] Liang, P. P., Zadeh, A., & Morency, L. P. (2024). Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, *56*(10), 1-42.

[49] Lin, W., Chen, J., Mei, J., Coca, A., & Byrne, B. (2023). Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems*, *36*, 22820-22840.

[50] Liu, X., Zhang, C., & Zhang, L. (2024). Vision mamba: A comprehensive survey and taxonomy. *arXiv preprint arXiv:2405.04404*.

[51] Lu, S., Liu, M., Yin, L., Yin, Z., Liu, X., & Zheng, W. (2023). The multi-modal fusion in visual question answering: a review of attention mechanisms. *PeerJ Computer Science*, *9*, e1400.

[52] Madan, N., Møgelmose, A., Modi, R., Rawat, Y. S., & Moeslund, T. B. (2024). Foundation Models for Video Understanding: A Survey. *arXiv preprint arXiv:2405.03770*.

[53] Mai, X., Tao, Z., Lin, J., Wang, H., Chang, Y., Kang, Y., ... & Zhang, W. (2024). From Efficient Multimodal Models to World Models: A Survey. *arXiv preprint arXiv:2407.00118*.

[54] Manzoor, M. A., Albarri, S., Xian, Z., Meng, Z., Nakov, P., & Liang, S. (2023). Multimodality representation learning: A survey on evolution, pretraining and its applications. *ACM Transactions on Multimedia Computing, Communications and Applications*, *20*(3), 1-34.

[55] Mudele, O. and Gamba, P., (2019), May. Mapping vegetation in urban areas using Sentinel-2. In 2019 Joint Urban Remote Sensing Event (JURSE) (pp. 1-4). IEEE.

[56] Mudele, O., Frery, A. C., Zanandrez, L. F. R., Eiras, A. E., & Gamba, P. (2021a). Dengue vector population forecasting using multisource earth observation products and recurrent neural networks. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 4390-4404.

[57] Mudele, O., Frery, A. C., Zanandrez, L. F. R., Eiras, A. E., & Gamba, P. (2021b). Modeling dengue vector population with earth observation data and a generalized linear model. Acta Tropica, 215, 105809.

[58] Pawłowski, M., Wróblewska, A., & Sysko-Romańczuk, S. (2023). Effective techniques for multimodal data fusion: A comparative analysis. *Sensors*, *23*(5), 2381.

[59] Proca, A. M., Rosas, F. E., Luppi, A. I., Bor, D., Crosby, M., & Mediano, P. A. (2024). Synergistic information supports modality integration and flexible learning in neural networks solving multiple tasks. *PLOS Computational Biology*, *20*(6), e1012178.

[60] Queralta, J. P., Taipalmaa, J., Pullinen, B. C., Sarker, V. K., Gia, T. N., Tenhunen, H., ... & Westerlund, T. (2020). Collaborative multi-robot search and rescue: Planning, coordination, perception, and active vision. *Ieee Access*, *8*, 191617-191643.

[61] Rafiq, M., Rafiq, G., & Choi, G. S. (2021). Video description: Datasets & evaluation metrics. *IEEE Access*, *9*, 121665-121685.

[62] Rafiq, M., Rafiq, G., & Choi, G. S. (2021). Video description: Datasets & evaluation metrics. *IEEE Access*, *9*, 121665-121685.

[63] Rice, C. A., Beekhuizen, B., Dubrovsky, V., Stevenson, S., & Armstrong, B. C. (2019). A comparison of homonym meaning frequency estimates derived from movie and television subtitles, free association, and explicit ratings. *Behavior research methods*, *51*, 1399-1425.

[64] Saunders, J., Syrdal, D. S., Koay, K. L., Burke, N., & Dautenhahn, K. (2015). "teach me–show me"—end-user personalization of a smart home and companion robot. *IEEE Transactions on Human-Machine Systems*, *46*(1), 27-40.

[65] Shaik, T., Tao, X., Li, L., Xie, H., & Velásquez, J. D. (2023). A survey of multimodal information fusion for smart healthcare: Mapping the journey from data to wisdom. *Information Fusion*, 102040.

[66] Shakil, H., Farooq, A., & Kalita, J. (2024). Abstractive text summarization: State of the art, challenges, and improvements. *Neurocomputing*, 128255.

[67] Sharma, H., & Jalal, A. S. (2021). A survey of methods, datasets and evaluation metrics for visual question answering. *Image and Vision Computing*, *116*, 104327.

[68] Shi, H., Liu, M., Mu, X., Song, X., Hu, Y., & Nie, L. (2024). Breaking Through the Noisy Correspondence: A Robust Model for Image-Text Matching. *ACM Transactions on Information Systems*, *42*(6), 1-26.

[69] Shoumy, N. J., Ang, L. M., Seng, K. P., Rahaman, D. M., & Zia, T. (2020). Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals. *Journal of Network and Computer Applications*, *149*, 102447.

[70] Siddique, S., Haque, M. A., George, R., Gupta, K. D., Gupta, D., & Faruk, M. J. H. (2023). Survey on machine learning biases and mitigation techniques. *Digital*, *4*(1), 1-68.

[71] Singh, B., Kaunert, C., Vig, K., & Gautam, B. K. (2024). Wearable Sensors Assimilated With Internet of Things (IoT) for Advancing Medical Imaging and Digital Healthcare: Real-Time Scenario. In *Inclusivity and Accessibility in Digital Health* (pp. 275-297). IGI Global.

[72] Sulubacak, U., Caglayan, O., Grönroos, S. A., Rouhe, A., Elliott, D., Specia, L., & Tiedemann, J. (2020). Multimodal machine translation through visuals and speech. *Machine Translation*, *34*, 97-147.

[73] Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., ... & Wang, X. (2023). Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*.

[74] Sun, Q., Yu, Q., Cui, Y., Zhang, F., Zhang, X., Wang, Y., ... & Wang, X. (2023). Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*.

[75] Takashima, S., Hayamizu, R., Inoue, N., Kataoka, H., & Yokota, R. (2023). Visual atoms: Pre-training vision transformers with sinusoidal waves. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 18579-18588).

[76] Wang, H., Zhang, Y., & Yu, X. (2020). An overview of image caption generation methods. *Computational intelligence and neuroscience*, *2020*(1), 3062706.

[77] Wang, J., Jiang, H., Liu, Y., Ma, C., Zhang, X., Pan, Y., ... & Zhang, S. (2024). A Comprehensive Review of Multimodal Large Language Models: Performance and Challenges Across Different Tasks. *arXiv preprint arXiv:2408.01319*.

[78] Wang, J., Jiang, H., Liu, Y., Ma, C., Zhang, X., Pan, Y., ... & Zhang, S. (2024). A Comprehensive Review of Multimodal Large Language Models: Performance and Challenges Across Different Tasks. *arXiv preprint arXiv:2408.01319*.

[79] Wang, K., Yin, Q., Wang, W., Wu, S., & Wang, L. (2016). A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*.

[80] Wang, T., Zheng, P., Li, S., & Wang, L. (2024). Multimodal Human–Robot Interaction for Human-Centric Smart Manufacturing: A Survey. *Advanced Intelligent Systems*, *6*(3), 2300359.

[81] Watson, E., Viana, T., & Zhang, S. (2023). Augmented behavioral annotation tools, with application to multimodal datasets and models: a systematic review. *AI*, *4*(1), 128-171.

[82] Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., & Van Den Hengel, A. (2017). Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, *163*, 21-40.

[83] Xi, Y., Zhang, Y., Ding, S., & Wan, S. (2020). Visual question answering model based on visual relationship detection. *Signal Processing: Image Communication*, *80*, 115648.

[84] Xiong, B., Yang, X., Qi, F., & Xu, C. (2022). A unified framework for multi-modal federated learning. *Neurocomputing*, *480*, 110-118.

[85] Xu, X., Li, J., Zhu, Z., Zhao, L., Wang, H., Song, C., ... & Pei, Y. (2024). A Comprehensive Review on Synergy of Multi-Modal Data and AI Technologies in Medical Diagnosis. *Bioengineering*, *11*(3), 219.

[86] Xu, X., Li, J., Zhu, Z., Zhao, L., Wang, H., Song, C., ... & Pei, Y. (2024). A Comprehensive Review on Synergy of Multi-Modal Data and AI Technologies in Medical Diagnosis. *Bioengineering*, *11*(3), 219.

[87] Yan, L., Martinez-Maldonado, R., & Gasevic, D. (2024, March). Generative artificial intelligence in learning analytics: Contextualising opportunities and challenges through the learning analytics cycle. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 101-111).

[88] Yeong, D. J., Velasco-Hernandez, G., Barry, J., & Walsh, J. (2021). Sensor and sensor fusion technology in autonomous vehicles: A review. *Sensors*, *21*(6), 2140.

[89] Yu, J., Li, J., Yu, Z., & Huang, Q. (2019). Multimodal transformer with multi-view visual representation for image captioning. *IEEE transactions on circuits and systems for video technology*, *30*(12), 4467-4480.

[90] Zhang, C., Yang, Z., He, X., & Deng, L. (2020). Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, *14*(3), 478-493.

[91] Zhang, H., Xu, J., Sun, H., & Zhao, Z. (2022, April). Commodity Image Retrieval Based on Image and Text Data. In *International Conference on Advanced Machine Learning Technologies and Applications* (pp. 99-111). Cham: Springer International Publishing.

[92] Zhang, S., Yang, Y., Chen, C., Zhang, X., Leng, Q., & Zhao, X. (2024). Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and prospects. *Expert Systems with Applications*, *237*, 121692.

[93] Zhang, Y., Zhang, C., Tang, Y., & He, Z. (2024). Cross-modal concept learning and inference for vision-language models. *Neurocomputing*, *583*, 127530.

[94] Zhao, Y., Liu, Z., & Wu, J. (2020). Grassland ecosystem services: a systematic review of research advances and future directions. *Landscape Ecology*, *35*, 793-814.

[95] Zong, Y., Mac Aodha, O., & Hospedales, T. (2024). Self-Supervised Multimodal Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.