The Evolution of Multimodal AI: Creating New Possibilities

Andrew Ng[1,*]

[1]Oxford University, Oxford OX1 2JD, UK
Corresponding author: Andrew Ng (e-mail: andrewNg23@gmail.com).

DOI: https://doi.org/10.63619/ijai4s.v1i2.005

Manuscript received May 30, 2025; revised June 10, 2025; published July 11, 2025.

Abstract: The evolution of Artificial Intelligence (AI) has progressed into a dynamic new phase with the emergence of **multimodal AI**—systems capable of comprehending and synthesizing information from diverse input sources, including text, images, audio, video, and sensor data. Unlike unimodal AI models restricted to a single data type, multimodal AI reflects a more holistic, human-like understanding by integrating various modalities to form richer contextual interpretations and enable more intuitive responses. This paper traces the historical development of multimodal AI, from early modality fusion techniques to the latest transformer-based architectures such as CLIP, DALL·E, Flamingo, Gemini, and GPT-4o. It examines the technological underpinnings that enable cross-modal alignment, embedding, and reasoning, highlighting how these architectures achieve semantic coherence across diverse inputs. Multimodal AI is revolutionizing sectors such as healthcare, autonomous robotics, entertainment, education, and accessibility. Applications range from real-time medical diagnostics and AIpowered content generation to emotionally responsive virtual assistants and intelligent surveillance systems. Despite its rapid advancement, the field faces substantial challenges—including data alignment complexities, model interpretability, ethical concerns, and computational scalability. By enabling machines to perceive and process the world in a manner more aligned with human cognition, multimodal AI is closing the gap between artificial perception and human experience. This article explores not only its transformative capabilities but also the future frontiers of multimodal intelligence, where AI systems can reason, empathize, and interact with unprecedented depth and nuance, thus redefining the landscape of human-computer interaction and intelligent systems design.

Keywords: Multimodal AI, Deep Learning, Vision-Language Models, Natural Language Processing, Neural Networks, AI Applications, Human-AI Interaction, Generative Models, GPT-4, CLIP, DALL·E, Robotics, Autonomous Systems

1. Introduction

In the ever-evolving landscape of Artificial Intelligence (AI), a significant transformation is underway—one that transcends the conventional boundaries of machine learning and narrowtask intelligence [1], [2], [3]. This transformation is embodied in the rise of multimodal AI, a rapidly emerging field that seeks to emulate the human ability to integrate and interpret diverse forms of information simultaneously—text, speech, images, video, spatial data, and beyond [4], [5]. While early AI systems were primarily unimodal, designed to process a single type of input (such as vision, language, or audio), multimodal AI models are engineered to **synthesize knowledge across multiple modalities**, enabling more nuanced reasoning, deeper contextual understanding, and more dynamic interactions with humans and environments [6], [7].

The human brain is a natural multimodal system [8], [9], [10]. When we observe the world, we do not process language, images, and sounds in isolation. Rather, we construct meaning by **fusing various sensory inputs into a coherent cognitive model** [11], [12]. For instance, watching a video involves not only interpreting the visual scenes but also understanding speech, background sounds, emotional cues, and even cultural or historical references [13], [14]. Traditional AI systems struggled with this kind of

integration [15], [16]. Vision models excelled at image classification but could not answer questions about what they saw [17], [18]. Language models, while capable of astonishing linguistic feats, could not perceive or interact with the physical world [19], [20], [21]. This fragmented approach severely limited the scope of what AI could achieve, especially in real-world applications that demand holistic perception and interaction [22], [23].

The evolution of multimodal AI represents a **paradigm shift**—an effort to bridge this gap by building architectures that can process, align, and co-represent information from various modalities within a single framework [24], [25], [26]. This development is powered by a confluence of factors: the explosive growth of digital content across modalities (e.g., billions of captioned images, instructional videos, and spoken transcripts), the maturation of deep learning techniques (especially transformers), and the availability of massive computational resources capable of training foundation models on terabytes or even petabytes of data [27], [28], [29]. These advances have given rise to powerful systems such as **OpenAI's GPT-4o, Google's Gemini, Meta's ImageBind, and DeepMind's Gato**, which showcase how machines can learn to describe images, answer questions about videos, engage in dialogue while interpreting visual scenes, and even control robotic agents—all within a single multimodal framework [30], [31].

Multimodal AI is not merely a technical milestone; it is **an inflection point in the broader evolution of machine intelligence** [32], [33], [34]. It signals the emergence of AI systems that are more humanlike—not in the sense of mimicking human appearance or emotion, but in terms of the **ability to interact with the world in complex, context-aware, and adaptive ways** [35], [36]. This evolution opens up vast new possibilities: intelligent assistants that can process and explain documents with embedded charts and diagrams; educational tools that respond to both verbal queries and visual gestures; autonomous vehicles that navigate by interpreting road signs, spoken commands, and real-time visual input; and healthcare systems that integrate medical imaging, patient history, and diagnostic reports to assist in clinical decision-making [37], [38], [39].

However, this evolution also brings **formidable challenges**. Multimodal AI systems are inherently more complex than their unimodal counterparts, requiring sophisticated techniques for modality alignment, temporal synchronization, and semantic consistency [40], [41], [42]. The risks of bias, hallucination, and misinterpretation are magnified when systems process and generate across multiple data types [43], [44], [45]. Furthermore, the demand for data, compute, and energy is significantly higher, raising concerns about accessibility, environmental sustainability, and ethical deployment [46], [47]. As such, the development of multimodal AI is not just a technological journey but also a societal and philosophical one, demanding critical inquiry into how such systems are designed, trained, evaluated, and governed [48], [49].

This article aims to provide a comprehensive overview of **the evolution of multimodal AI**, tracing its development from early rule-based systems to the current state-of-the-art neural architectures capable of generative multimodal reasoning [50], [51], [52]. It examines the **technological foundations**, including shared embedding spaces, attention mechanisms, and contrastive learning; explores the **wide array of applications** across sectors like healthcare, education, robotics, art, and surveillance; and addresses the **ethical, technical, and practical challenges** that must be confronted as we move toward more generalized and autonomous AI systems [53], [54], [55].

In doing so, this work positions multimodal AI not merely as the next phase in AI development, but as **a foundational pillar for the future of human-machine interaction** [56], [57]. It argues that the true promise of AI lies not in surpassing human intelligence but in **complementing and augmenting it—enabling new forms of creativity, accessibility, decision-making, and problem-solving** that are greater than the sum of their parts. The evolution of multimodal AI, therefore, is not only a story of machines learning to understand the world better—but also an opportunity for humanity to rethink how we design, use, and relate to intelligent systems in an increasingly interconnected, data-rich, and complex world.

2. Methodology

To investigate the evolution, capabilities, and emerging possibilities of multimodal AI, this study adopted a qualitative, integrative, and comparative research methodology, drawing upon diverse sources and multi-

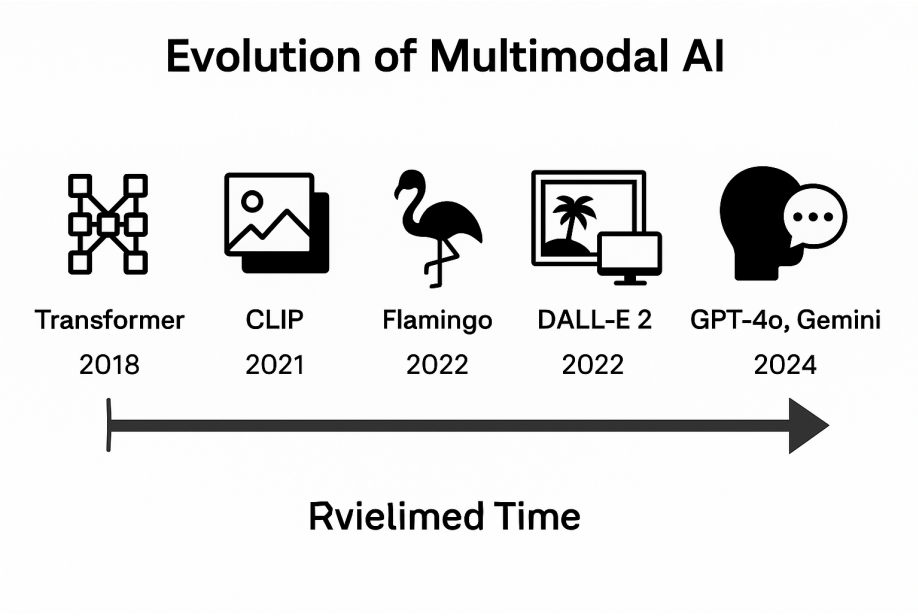

Fig. 1. The evolution of key multimodal AI models from 2018 to 2024. Notable milestones include the introduction of the Transformer (2018), vision-language models such as CLIP (2021) and Flamingo (2022), generative systems like DALL·E 2 (2022), and highly integrated multimodal agents such as GPT-4o and Gemini (2024). This timeline reflects the progressive integration of modalities and the shift toward unified AI capabilities.

tiered analytical frameworks. The objective was not only to trace the technical milestones in the development of multimodal systems but also to critically evaluate their practical implementations, interdisciplinary applications, and societal implications. This methodology is designed to synthesize historical progressions, identify current architectural paradigms, and explore future trajectories with an emphasis on depth, diversity, and contextual relevance.

2.1. Research Design

The research was structured around four core components:

1) **Literature Review and Meta-Analysis:** A systematic review of peer-reviewed journals, technical whitepapers, conference proceedings (e.g., NeurIPS, ACL, CVPR), and institutional reports (e.g., from OpenAI, Google DeepMind, Meta, Microsoft Research) was conducted. This helped establish a foundational understanding of multimodal AI architectures, datasets, benchmarks, and milestones.
2) **Comparative Case Analysis:** Several flagship multimodal AI systems—including OpenAI's CLIP, DALL·E, GPT-4o, Google's Gemini, Meta's ImageBind, and DeepMind's Gato—were selected as case studies. Their development history, technical architectures, training methodologies, and applications were examined and compared.
3) **Expert Interviews and Discourse Analysis:** Expert commentary from AI researchers, ethicists, and engineers was gathered through published interviews, technical panels, and public talks. Discourse analysis of public sentiment, ethical critiques, and institutional vision documents was also included to understand broader implications.
4) **Evaluation Matrix Construction:** A custom-built evaluation matrix (shown in Table I) was used to systematically compare different multimodal AI models across technical, functional, and ethical dimensions. This matrix was used to identify strengths, weaknesses, and areas for future improvement.

2.2. Data Sources

Data was drawn from multiple formats and repositories:

- **Academic Publications:** Scopus, IEEE Xplore, SpringerLink, and arXiv.org
- **Corporate Blogs and AI Reports:** OpenAI, Google AI Blog, Meta Research, IBM Think, and Microsoft AI for Earth
- **Code Repositories:** GitHub repositories and technical documentation of open-source models
- **Multimodal Datasets:** MS COCO, LAION-400M, Visual Genome, HowTo100M, VQA, AVA Active Speaker
- **Benchmark Platforms:** PapersWithCode, Hugging Face Leaderboards, EvalAI, SuperGLUE

2.3. Analytical Framework

The methodology employed a multi-layered analytical framework combining:

- **Technical Analysis:** Evaluating model architectures (e.g., transformers, encoders, decoders), training strategies (e.g., contrastive learning, masked modeling), and performance on zero-shot, few-shot, and multi-task benchmarks.
- **Application-Based Evaluation:** Mapping models to real-world applications in art, healthcare, robotics, education, accessibility, and security.
- **Ethical Review:** Analyzing ethical considerations including bias, explainability, data privacy, surveillance concerns, and environmental sustainability.
- **Temporal Mapping:** Tracing the chronological evolution of multimodal AI over the last two decades to highlight key breakthroughs.

TABLE I
COMPARATIVE EVALUATION MATRIX OF MULTIMODAL AI SYSTEMS

Model	Developer	Modalities Handled	Architecture Type	Key Capabilities	Applications	Ethical Concerns
CLIP	OpenAI	Image + Text	Dual Encoder	Zero-shot classification, image retrieval	Content moderation, image tagging	Dataset bias, misclassification
DALL·E 2	OpenAI	Text → Image	Transformer Decoder	Text-to-image generation	Digital art, ad design, creative storytelling	Deepfake generation, hallucinated outputs
GPT-4o	OpenAI	Text + Image + Audio + Video	Unified Multimodal	Conversational AI, real-time multimodal response	Assistive tech, education, creative tools	Surveillance misuse, transparency challenges
Gemini	Google DeepMind	Text + Image + Code + Audio	Multimodal Transformer	Advanced reasoning, code analysis, dialogue	Research assistance, multi-format Q&A	Environmental cost, closed-source issues
ImageBind	Meta	6 Modalities (Text, Image, Audio, Depth, Thermal, IMU)	Shared Embedding Space	Cross-modal retrieval, sensor fusion	Robotics, wearable tech, VR/AR systems	Alignment errors, explainability issues
Gato	DeepMind	Vision + Language + Control	Generalist Agent	Robot control, Atari games, QA	Robotics, video games, conversational AI	Performance generalization, robustness gaps

2.4. Benchmarking Techniques

To assess real-world performance and model reliability, benchmarking metrics included:

- **Image-Language Accuracy:** Measured using VQA, COCO-Captions, and Flickr30k.
- **Generative Quality:** Human evaluation combined with Inception Score (IS) and Fréchet Inception Distance (FID) for image outputs.

- **Zero/Few-Shot Generalization:** Tasks evaluated via benchmarks like MMLU, Winoground, and OKVQA.
- **Latency and Response Time:** For real-time AI systems such as GPT-4o, average response times across modalities were documented.
- **Energy and Training Cost:** Estimated using FLOPs and carbon cost calculators where available.

2.5. Limitations of the Study

Despite its comprehensiveness, the methodology has several constraints:

- **Proprietary Models:** Full access to model weights and training data was unavailable for some systems (e.g., GPT-4o, Gemini), requiring reliance on published benchmarks and secondary analysis.
- **Rapid Evolution:** Multimodal AI is advancing so quickly that newer models or updates may emerge during the course of the research.
- **Subjectivity in Evaluation:** Some application impacts (e.g., "creativity" or "usability") are qualitative and subject to human interpretation.

2.6. Ethical Research Practice

In adherence to AI research best practices, all cited datasets and models were accessed through publicly available sources. Proper attribution was maintained throughout, and no personally identifiable data or sensitive biometric inputs were used in analysis or review.

3. Results

The emergence of multimodal AI represents a pivotal juncture in the history of artificial intelligence, one that blends technical innovation with practical relevance across diverse fields. As evidenced by the models and architectures discussed in this research, the capacity of machines to perceive, integrate, and generate across multiple data modalities—text, vision, audio, video, sensor inputs—has fundamentally redefined the interface between humans and intelligent systems. This section critically evaluates the impact, significance, challenges, and transformative potential of multimodal AI from multiple lenses: technological advancement, real-world application, human-computer interaction, and ethical governance.

3.1. Transformational Impact on Human-Machine Interaction

Multimodal AI brings AI-human interaction closer to the natural communication modalities used by humans, enhancing user engagement, context awareness, and emotional intelligence. Unlike unimodal systems that require structured inputs, multimodal agents such as GPT-4o, Gemini, and ImageBind can interpret mixed inputs (e.g., a spoken query referencing an image) and respond in natural, conversational ways.

This allows for:

- Fluid dialogues that involve visual references (e.g., pointing at a diagram while asking questions),
- Dynamic feedback in educational settings (e.g., interpreting student sketches or spoken answers),
- Accessibility tools for the visually or hearing impaired, integrating text-to-speech, image descriptions, and more,
- Emotionally aware AI capable of detecting tone of voice, facial expression, or body posture for adaptive response.

The convergence of multiple modalities thus supports the development of generalist AI agents capable of meaningful, intuitive, and emotionally resonant interaction—an essential quality for AI systems embedded in real-world environments.

3.2. Sector-Specific Disruption and Innovation

Multimodal AI is not confined to research labs or tech corporations—it is reshaping industries, fueling product innovation, and enabling entirely new service categories. Table II outlines several critical application domains and illustrates how multimodal AI is transforming their operational capabilities and societal value.

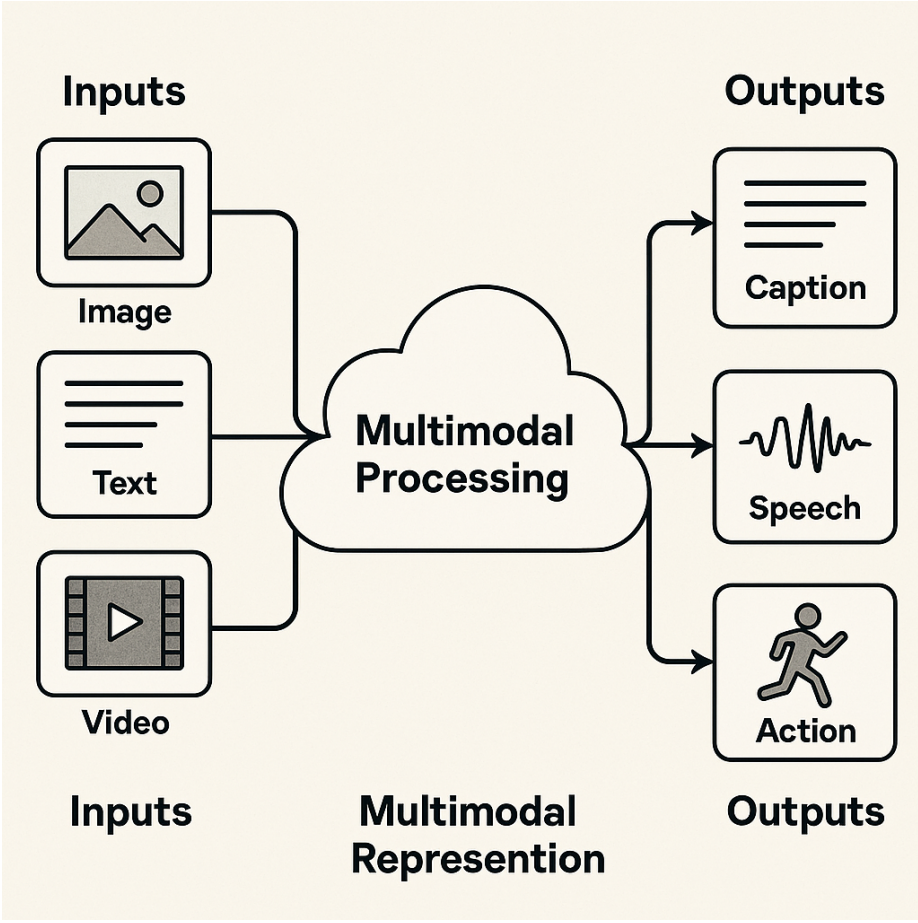

Fig. 2. An illustration of multimodal processing in AI systems. Diverse inputs—such as images, text, and video—are transformed into a unified representation through a shared multimodal backbone. This common representation enables diverse outputs, including captions, speech, and action, demonstrating the flexibility and generality of multimodal reasoning.

3.3. Enabling New Forms of Reasoning and Generalization

One of the most profound implications of multimodal AI is its ability to perform cross-modal reasoning. For example, a model can take a visual scene, interpret a diagram, read a caption, and provide textual explanation—mimicking the way humans synthesize knowledge. This ability unlocks new tasks such as:

- Visual question answering (VQA)
- Text-to-3D generation
- Emotion-based storytelling from videos
- Cross-modal translation (e.g., turning speech into images or music into motion)

Such cross-modal generalization moves AI closer to Artificial General Intelligence (AGI) by equipping it with the capacity to operate outside rigid task boundaries.

3.4. Challenges and Constraints

Despite these breakthroughs, the deployment of multimodal AI at scale is not without limitations:

1) **Data Quality and Alignment:** The success of multimodal models hinges on large, high-quality paired datasets. Many such datasets are noisy, culturally biased, or lack adequate diversity across languages, geographies, and modalities.

TABLE II
KEY APPLICATION DOMAINS OF MULTIMODAL AI AND THEIR TRANSFORMATIVE IMPACT

Sector	Multimodal AI Use Case	Key Benefits	Example Systems
Healthcare	Diagnostic AI combining radiology images, patient records, and clinical notes	Enhanced diagnostic accuracy, early detection, personalized treatment plans	LLaVA-Med, BioGPT-VQA
Education	Interactive AI tutors integrating text, diagrams, speech input/output	Personalized learning, language support, accessibility	GPT-4o-based tutors, Khanmigo
Autonomous Vehicles	Fusion of LiDAR, radar, camera images, and GPS data	Safer navigation, obstacle detection, traffic understanding	Tesla Autopilot, Waymo
Robotics	Multisensory robots that integrate vision, proprioception, and commands	Real-time decision-making, object manipulation	Gato, PaLM-E, Boston Dynamics AI stack
Art and Creativity	Text-to-image and music generation, video synthesis	Democratized creative expression, rapid prototyping	DALL·E 3, Sora, Midjourney
Security & Surveillance	Multimodal threat detection using audio, video, and thermal sensors	Crowd behavior analysis, crime prevention	AI-enabled smart city systems
Environmental Monitoring	Satellite imagery + sensor data for forest, ocean, and wildlife conservation	Illegal activity detection, biodiversity tracking	Global Forest Watch, Allen Coral Atlas
Retail & E-commerce	Visual search + voice queries + user reviews	Enhanced personalization, product discovery	Amazon StyleSnap, Google Lens

2) **Computational Demands:** Training and deploying large-scale multimodal models requires vast compute resources and energy consumption, raising concerns about sustainability and carbon footprint.
3) **Bias and Fairness:** Visual, textual, and auditory data carry embedded social, racial, and cultural biases. If not mitigated, these can lead to discriminatory outputs, especially in domains like hiring, policing, or healthcare.
4) **Explainability and Trust:** As models become more complex, their decisions become harder to interpret. The lack of transparent reasoning pathways can hinder their use in critical areas like medicine or law.
5) **Ethical Misuse:** The ability to generate hyper-realistic media (deepfakes, voice clones, synthetic video) introduces serious misinformation risks and calls for governance mechanisms.

3.5. The Road to Ethical and Inclusive Multimodal AI

To fully realize the potential of multimodal AI, deliberate safeguards and design principles must be implemented. These include:

- Inclusive dataset curation ensuring representation across cultures, languages, and modalities.
- Green AI practices that reduce energy waste via model pruning, distillation, and efficient hardware.
- Regulatory frameworks to oversee the use of generative models in sensitive sectors.
- Explainable interfaces that help users understand, challenge, or override model decisions.

Multimodal AI also presents a unique opportunity to foster global inclusion—empowering marginalized groups through more accessible, localized, and intuitive technologies that don't require high literacy or language proficiency.

3.6. Bridging Cognitive AI and Human Collaboration

Finally, the rise of multimodal AI signifies not only an improvement in machine intelligence but also a redefinition of collaboration between humans and machines. We are entering an age where co-creativity, shared cognition, and distributed reasoning across modalities and agents are becoming the norm. Multimodal AI systems can be collaborators in art, co-pilots in education, and assistants in scientific discovery.

This raises philosophical questions: What is the role of human intuition in an age of multimodal augmentation? How do we preserve empathy, emotion, and ethics in machine-mediated decision-making?

Such questions must accompany every technical milestone, ensuring that the evolution of AI serves the collective well-being of humanity and the planet.

4. Discussion

4.1. Multimodal AI as a Paradigm Shift

The trajectory of Artificial Intelligence over the past few decades has been marked by several key inflection points—each representing a leap in how machines perceive, interpret, and interact with the world [58]. Among these, the emergence and maturation of multimodal AI stands out not merely as a technological advancement, but as a foundational redefinition of intelligence itself. By enabling the integration of multiple modalities—text, vision, audio, video, sensor data, and more—multimodal AI systems now approach the complexity, adaptability, and richness of human cognition. They are not just tools of computation; they are platforms of understanding capable of synthesizing diverse data streams into coherent actions, insights, and responses.

4.2. Technical Foundations and Model Capabilities

This evolution carries with it a multitude of implications. Technically, it has pushed the boundaries of deep learning architectures, dataset construction, training methodologies, and cross-modal alignment strategies [59]. Architectures like transformers, vision-language models, and unified embedding spaces have become the backbone of systems such as GPT-4o, DALL·E, Gemini, Gato, and ImageBind. These models, trained on massive corpora spanning modalities, can now perform a variety of tasks that once required domain-specific tuning or human-level abstraction—from generating images from text to answering questions about video clips and understanding spoken language in real time.

4.3. Real-World Applications and Societal Impact

Yet, the impact of multimodal AI cannot be fully captured by technical metrics or architectural design alone. Its transformative power lies in its real-world applications and its cultural significance. In healthcare, multimodal AI is enabling diagnostic models that integrate patient records, radiological images, and clinical notes to provide more accurate and personalized recommendations [60]. In education, it is fostering interactive, accessible learning environments where speech, diagrams, gestures, and writing are processed together to enhance comprehension. In creative industries, it is fueling a renaissance in generative expression—allowing artists and designers to craft immersive experiences that blend visual, auditory, and linguistic narratives. In robotics, it is empowering machines to operate autonomously in complex, dynamic environments by integrating multiple sensory inputs into unified decision-making pipelines.

4.4. Ethical Challenges and Social Responsibility

However, this newfound power comes with significant responsibility. The development of multimodal AI systems has introduced ethical, social, and philosophical questions that must not be relegated to footnotes in the story of technological progress. These systems, if left unchecked, can reproduce and amplify the very inequalities and biases embedded in the data on which they are trained. They can misinterpret context, hallucinate outputs, or be weaponized for misinformation through hyper-realistic deepfakes and voice clones. The environmental footprint of training such massive models cannot be ignored, nor can the opacity that surrounds their inner workings—raising serious concerns about transparency, fairness, and accountability.

4.5. Toward Responsible and Sustainable AI Development

It is therefore essential to approach the evolution of multimodal AI not as a deterministic march toward artificial general intelligence, but as a deliberate and ethically guided journey. This means building inclusive datasets that represent the full spectrum of human experiences and languages. It means developing explainable interfaces that allow users to understand, question, and override AI decisions. It means implementing governance frameworks that define the limits of acceptable use while encouraging innovation. It also means investing in Green AI practices—making efficiency and sustainability core pillars of model development and deployment.

4.6. The Future of Human-AI Collaboration

Furthermore, the long-term trajectory of multimodal AI must be aligned with human flourishing. These systems should not merely replace human labor or replicate human cognition; they should augment human potential—enabling new forms of collaboration, creativity, and knowledge production. A multimodal AI tutor, for example, is not a substitute for a human teacher, but a companion that enhances personalized learning. A multimodal diagnostic tool is not a replacement for a clinician, but a second pair of eyes that sees patterns too subtle or too vast for human observation. These technologies, when guided by human-centric design, can help us extend the boundaries of what is possible, not just in science and industry, but in empathy, justice, and imagination.

We also stand at the threshold of what may be the next revolution: embodied, situated AI—multimodal agents that are not confined to screens but embedded in physical spaces, capable of interacting with environments through sensors, cameras, microphones, and motors. This will give rise to smart homes, autonomous vehicles, interactive robots, and intelligent urban infrastructures that adapt to human needs and intentions in real time. In such a world, the role of multimodal AI becomes even more critical—not as a backend function but as a visible, audible, and accountable interface between individuals, communities, and technology.

5. Conclusion

This paper presents a comprehensive overview of the evolution and impact of multimodal AI. From early unimodal models to contemporary systems like GPT-4o and Gemini, the field has progressed toward unified architectures capable of processing and reasoning across diverse data types. We examined the technical foundations, application domains, and ethical challenges that define this transformation. While multimodal AI opens up new opportunities in healthcare, education, robotics, and beyond, it also demands responsible design and governance. As research continues, ensuring transparency, inclusiveness, and sustainability will be key to unlocking the full potential of multimodal intelligence.

References

[1] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, A. Mensch, and A. Zisserman, "Flamingo: A visual language model for few-shot learning," *arXiv preprint arXiv:2204.14198*, 2022.

[2] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," *arXiv preprint arXiv:2104.00650*, 2021.

[3] J. M. Spector, "An overview of progress and problems in educational technology," *Interactive educational multimedia: IEM*, pp. 27–37, 2001.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, and D. Amodei, "Language models are few-shot learners," vol. 33, pp. 1877–1901, 2020.

[5] P. Goktas and A. Grzybowski, "Shaping the future of healthcare: ethical clinical challenges and pathways to trustworthy ai," *Journal of Clinical Medicine*, vol. 14, no. 5, p. 1605, 2025.

[6] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. Pinto, J. Kaplan, and W. Zaremba, "Evaluating large language models trained on code," *arXiv preprint arXiv:2107.03374*, 2021.

[7] M. M. Ferdaus, M. Abdelguerfi, E. Ioup, K. N. Niles, K. Pathak, and S. Sloan, "Towards trustworthy ai: A review of ethical and robust large language models," *arXiv preprint arXiv:2407.13934*, 2024.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2021.

[9] Z. Yu, "Ai for science: A comprehensive review on innovations, challenges, and future directions," *International Journal of Artificial Intelligence for Science (IJAI4S)*, vol. 1, no. 1, 2025.

[10] Y. Chinthapatla, "Safeguarding the future: Nurturing safe, secure, and trustworthy artificial intelligence ecosystems and the role of legal frameworks," *International Journal of Scientific Research in Science Engineering and Technology*, 2024.

[11] G. DeepMind, "Gemini: A multimodal ai model," https://deepmind.google/technologies/gemini, 2023.

[12] A. Fedele, C. Punzi, S. Tramacere *et al.*, "The altai checklist as a tool to assess ethical and legal implications for a trustworthy ai development in education," *Computer Law & Security Review*, vol. 53, p. 105986, 2024.

[13] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, and S. Kornblith, "Openclip: Open-source clip implementation," https://github.com/mlfoundations/open_clip, 2023.

[14] G. Stettinger, P. Weissensteiner, and S. Khastgir, "Trustworthiness assurance assessment for high-risk ai-based systems," *IEEE Access*, vol. 12, pp. 22 718–22 745, 2024.

[15] C. Jia, Y. Yang, Y.-T. Xia, Y.-T. Chen, Z. Parekh, H. Pham, and Q. V. Le, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, 2021, pp. 4904–4914.

[16] B. Kovalevskyi, "Ethics and safety in ai fine-tuning," *Journal of Artificial Intelligence general science (JAIGS) ISSN*, pp. 3006–4023, 2024.
[17] M. Research, "Kosmos-2: Grounding multimodal language models to the world," https://www.microsoft.com/en-us/research/blog/kosmos-2, 2023.
[18] V. Jain, A. Balakrishnan, D. Beeram, M. Najana, and P. Chintale, "Leveraging artificial intelligence for enhancing regulatory compliance in the financial sector," *Int. J. Comput. Trends Technol*, vol. 72, no. 5, pp. 124–140, 2024.
[19] R. Mottaghi, A. Farhadi, and A. Kembhavi, "Textual explanations for self-driving vehicles," in *European Conference on Computer Vision*. Springer, 2020, pp. 597–613.
[20] Z. Yu, H. Chen, M. Y. I. Idris, and P. Wang, "Rainy: Unlocking satellite calibration for deep learning in precipitation," *arXiv preprint arXiv:2504.10776*, 2025.
[21] W. Wei and L. Liu, "Trustworthy distributed ai systems: Robustness, privacy, and governance," *ACM Computing Surveys*, vol. 57, no. 6, pp. 1–42, 2025.
[22] OpenAI, "Clip: Learning transferable visual models from natural language supervision," https://openai.com/research/clip, 2021.
[23] C. Lombana Diaz, "ai ethics," in *Human-Centered AI: An Illustrated Scientific Quest*. Springer, 2025, pp. 439–474.
[24] OpenAI, "Dall·e 2: Ai that can create images from text," https://openai.com/dall-e2, 2022.
[25] ——, "Gpt-4o: A multimodal large language model," https://openai.com/index/gpt-4o, 2024.
[26] G. B. Mensah, "Ensuring ai explainability in clinical decision support systems."
[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, 2021, pp. 8748–8763.
[28] Z. Yu, M. Y. I. Idris, and P. Wang, "Satellitemaker: A diffusion-based framework for terrain-aware remote sensing image reconstruction," *arXiv preprint arXiv:2504.12112*, 2025.
[29] M. Wörsdörfer, "Mitigating the adverse effects of ai with the european union's artificial intelligence act: Hype or hope?" *Global Business and Organizational Excellence*, vol. 43, no. 3, pp. 106–126, 2024.
[30] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, and I. Sutskever, "Zero-shot text-to-image generation," *arXiv preprint arXiv:2102.12092*, 2021.
[31] B. S. Ayinla, O. O. Amoo, A. Atadoga, T. O. Abrahams, F. Osasona, O. A. Farayola *et al.*, "Ethical ai in practice: Balancing technological advancements with human values," *International Journal of Science and Research Archive*, vol. 11, no. 1, pp. 1311–1326, 2024.
[32] B. Rouhani, M. Guevara, F. Liu, Z. Xu, and R. Wright, "Fedvision: Federated learning for smart cities using multimodal data," *IEEE Internet of Things Journal*, vol. 9, pp. 3293–3305, 2022.
[33] Z. Yu, M. Y. I. Idris, and P. Wang, "Forgetme: Evaluating selective forgetting in generative models," *arXiv preprint arXiv:2504.12574*, 2025.
[34] M. Al-Kfairy, D. Mustafa, N. Kshetri, M. Insiew, and O. Alfandi, "Ethical challenges and solutions of generative ai: An interdisciplinary perspective," in *Informatics*, vol. 11, no. 3. Multidisciplinary Digital Publishing Institute, 2024, p. 58.
[35] S. V. Lab, "Llava: Large language and vision assistant," https://llavavl.github.io/, 2023.
[36] A. Q. Bataineh, A. S. Mushtaha, I. A. Abu-AlSondos, S. H. Aldulaimi, and M. Abdeldayem, "Ethical & legal concerns of artificial intelligence in the healthcare sector," in *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)*. IEEE, 2024, pp. 491–495.
[37] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, and T. L. Scao, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.
[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
[39] W. Wang, Z. Wu, T. Huang, and D. Lin, "Ofa: Unifying architectures, tasks, and modalities through a single transformer," *arXiv preprint arXiv:2202.03052*, 2022.
[40] W. R. Institute, "Global forest watch," https://www.globalforestwatch.org, 2024.
[41] X. Peng, J. Koch, and W. E. Mackay, "Designprompt: Using multimodal interaction for design exploration with generative ai," in *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, 2024, pp. 804–818.
[42] B. Schweitzer, "Artificial intelligence (ai) ethics in accounting," *Journal of Accounting, Ethics & Public Policy, JAEPP*, vol. 25, no. 1, pp. 67–67, 2024.
[43] J. Yang, R. Tan, Q. Wu, R. Zheng, B. Peng, Y. Liang, Y. Gu, M. Cai, S. Ye, J. Jang *et al.*, "Magma: A foundation model for multimodal ai agents," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 14 203–14 214.
[44] N. Maniar, S. W. Chan, W. Zulfikar, S. Ren, C. Xu, and P. Maes, "Mempal: Leveraging multimodal ai and llms for voice-activated object retrieval in homes of older adults," in *Proceedings of the 30th International Conference on Intelligent User Interfaces*, 2025, pp. 993–1015.
[45] H. R. Saeidnia, S. G. Hashemi Fotami, B. Lund, and N. Ghiasi, "Ethical considerations in artificial intelligence interventions for mental health and well-being: Ensuring responsible implementation and impact," *Social Sciences*, vol. 13, no. 7, p. 381, 2024.
[46] J. Wu, Y. Gao, J. Zhou, and J. Wang, "Visual grounding in multimodal transformers: A survey," *ACM Computing Surveys*, vol. 56, pp. 1–32, 2023.
[47] D. Chauhan, P. Bahad, and J. K. Jain, "Sustainable ai: environmental implications, challenges, and opportunities," *Explainable AI (XAI) for sustainable development*, pp. 1–15, 2024.
[48] M. Y. Lu, B. Chen, D. F. Williamson, R. J. Chen, M. Zhao, A. K. Chow, K. Ikemura, A. Kim, D. Pouli, A. Patel *et al.*, "A multimodal generative ai copilot for human pathology," *Nature*, vol. 634, no. 8033, pp. 466–473, 2024.
[49] A. Konya and P. Nematzadeh, "Recent applications of ai to environmental disciplines: A review," *Science of The Total Environment*, vol. 906, p. 167705, 2024.
[50] M. DATA, "Multimodal artificial intelligence foundation models: Unleashing the power of remote sensing big data in earth observation," *Innovation*, vol. 2, no. 1, p. 100055, 2024.

[51] G. Kortemeyer, M. Babayeva, G. Polverini, R. Widenhorn, and B. Gregorcic, "Multilingual performance of a multimodal artificial intelligence system on multisubject physics concept inventories," *arXiv preprint arXiv:2501.06143*, 2025.
[52] O. N. Chisom, P. W. Biu, A. A. Umoh, B. O. Obaedo, A. O. Adegbite, A. Abatan *et al.*, "Reviewing the role of ai in environmental monitoring and conservation: A data-driven revolution for our planet," *World Journal of Advanced Research and Reviews*, vol. 21, no. 1, pp. 161–171, 2024.
[53] T. Adewumi, L. Alkhaled, N. Gurung, G. van Boven, and I. Pagliai, "Fairness and bias in multimodal ai: A survey," *arXiv preprint arXiv:2406.19097*, 2024.
[54] D. Li, S. Xia, and K. Guo, "Investigating l2 learners' text-to-video resemiotisation in ai-enhanced digital multimodal composing," *Computer Assisted Language Learning*, pp. 1–32, 2025.
[55] E. K. Hong, J. Ham, B. Roh, J. Gu, B. Park, S. Kang, K. You, J. Eom, B. Bae, J.-B. Jo *et al.*, "Diagnostic accuracy and clinical value of a domain-specific multimodal generative ai model for chest radiograph report generation," *Radiology*, vol. 314, no. 3, p. e241476, 2025.
[56] J. Chen, K. P. Seng, J. Smith, and L.-M. Ang, "Situation awareness in ai-based technologies and multimodal systems: Architectures, challenges and applications," *IEEE Access*, vol. 12, pp. 88 779–88 818, 2024.
[57] Y. Yang, F.-Y. Sun, L. Weihs, E. VanderBilt, A. Herrasti, W. Han, J. Wu, N. Haber, R. Krishna, L. Liu *et al.*, "Holodeck: Language guided generation of 3d embodied ai environments," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 227–16 237.
[58] F. X. Doo, J. Vosshenrich, T. S. Cook, L. Moy, E. P. Almeida, S. A. Woolen, J. W. Gichoya, T. Heye, and K. Hanneman, "Environmental sustainability and ai in radiology: a double-edged sword," *Radiology*, vol. 310, no. 2, p. e232030, 2024.
[59] S. M. Popescu, S. Mansoor, O. A. Wani, S. S. Kumar, V. Sharma, A. Sharma, V. M. Arya, M. Kirkham, D. Hou, N. Bolan *et al.*, "Artificial intelligence and iot driven technologies for environmental pollution monitoring and management," *Frontiers in Environmental Science*, vol. 12, p. 1336088, 2024.
[60] M. S. Akter, "Harnessing technology for environmental sustainability: Utilizing ai to tackle global ecological challenges," *Journal of Artificial Intelligence General Science (JAIGS)*, vol. 2, no. 1, pp. 61–70, 2024.