Gartner®

Innovation Insight: Multimodal AI Explained

Published 22 November 2023 - ID G00798532 - 13 min read

By Analyst(s): Leinar Ramos, Anthony Mullen, Ben Yan, Tong Zhang

Initiatives: Artificial Intelligence; Generative AI Resource Center

AI is quickly evolving toward models that combine multiple types of data such as text, images, video, audio and more. To drive more value from their data, IT leaders responsible for AI should explore multimodal AI models and start adding them into their capabilities.

Overview

Key Findings

- Artificial intelligence (AI) foundation models will increasingly become multimodal — they will be trained with different types of data (modalities) and be capable of handling more than one modality in their inputs and outputs. This is driven both by technical advances and by the performance, robustness and user experience improvements that multimodality unlocks.

- Multimodality opens up a broad set of new AI use cases that were not previously possible. These include: answering questions about videos and images, creating audio and video from text, and building agents that can dynamically interact via text, images, audio and video. Multimodal AI can be applied across the organization, particularly in areas like sales and marketing, media and entertainment, product design and customer service.

- Feasibility ranges widely across use cases. Multimodal AI is state of the art for image generation and other use cases that combine text and images. However, the technology is nascent for models that combine more than two modalities at once.

- Multimodality presents challenges in terms of data management, governance and integration, as well as new risks around data privacy and technical complexity in its use and customization.

Recommendations

- Educate your AI team on multimodality, introducing the concept, its benefits and risks. Break AI technical silos by encouraging AI experts to work on AI projects outside of their AI technical specialization, such as natural language processing and computer vision.

- Identify AI use cases in your organization where a model that combines multiple data modalities could generate business value above and beyond what is currently possible with unimodal AI foundation models. For example, using a mix of audio, text and video signals for sentiment analysis. Run pilots to demonstrate the business value that can be generated with these models.

- Start leveraging multimodal AI in use cases, such as generating images with complex text descriptions, where these models have already shown to be outperforming unimodal ones. Monitor the feasibility of emerging use cases, such as video generation, as the technology evolves.

- Experiment with new technical pipelines and workflows required for multimodal AI. Identify and mitigate risks around data privacy and security through robust data engineering and data governance.

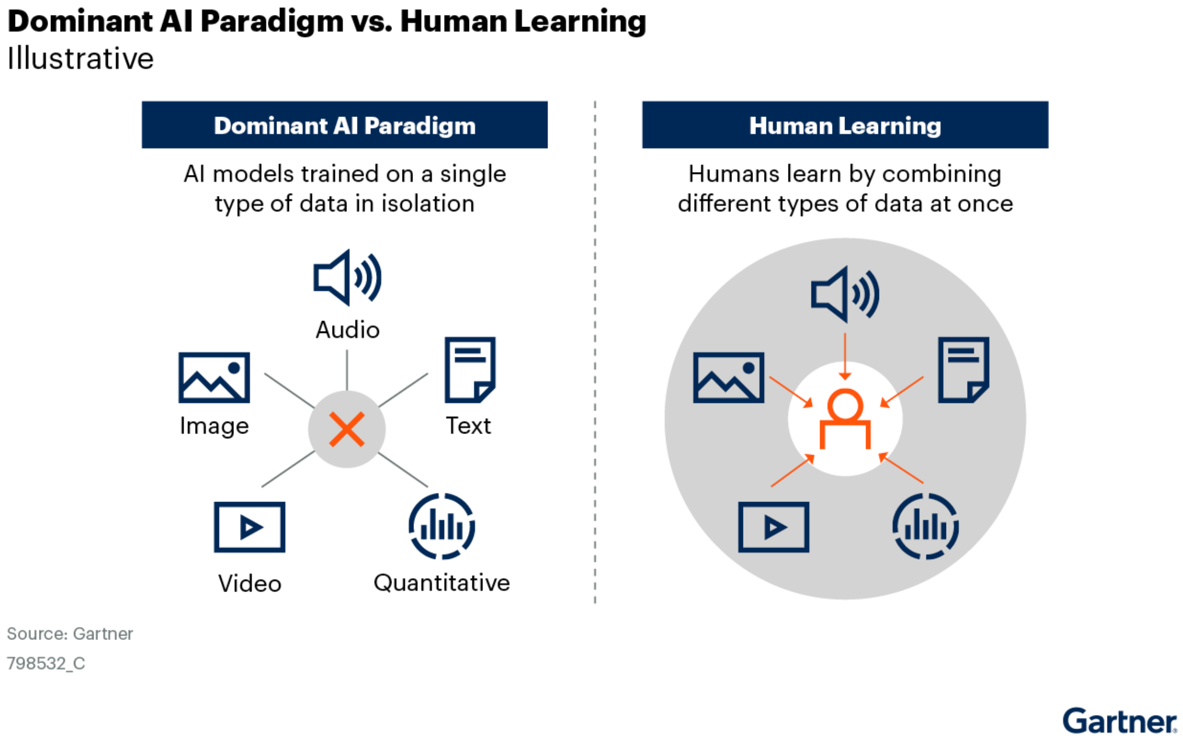
Strategic Planning Assumption

By 2025, AI for video, audio, vibration, text, emotion and other content analytics will trigger major innovations and transformations in 75% of Fortune 500 global enterprises.

Introduction

Current AI focuses on training models on a single type of data, also known as a modality — audio, documents, images, video, quantitative data and others (see Note 1). But this is not the way that we learn as humans. Instead, we simultaneously combine information from different types of data to make sense of the world (see Figure 1).

Figure 1: Dominant AI Paradigm vs. Human Learning


Dominant AI Paradigm vs. Human Learning
Illustrative
Dominant AI Paradigm
AI models trained on a single type of data in isolation
Audio
Image
Text
Video
Quantitative
Human Learning
Humans learn by combining different types of data at once
Gartner

Source: Gartner
798532_C

Increasingly, advanced AI foundation models are following the same path and are being trained on multiple modalities. For example, state-of-the-art large language models (LLMs), like GPT-4 Vision, are trained to take in both images and text, allowing users to ask questions about images and receive answers via text. This is only the beginning of this transformation.

Multimodality in AI is not limited to the senses available to humans. AI foundation models can be trained in other types of data, including infrared images, robotic actuator sensors and many more.

Multimodality is a key next step in generative AI, particularly for LLMs. These additional modalities will help to better ground LLMs in the real world and increase their capabilities beyond what is currently possible. IT leaders responsible for AI need to prepare their organizations for a multimodal AI future.

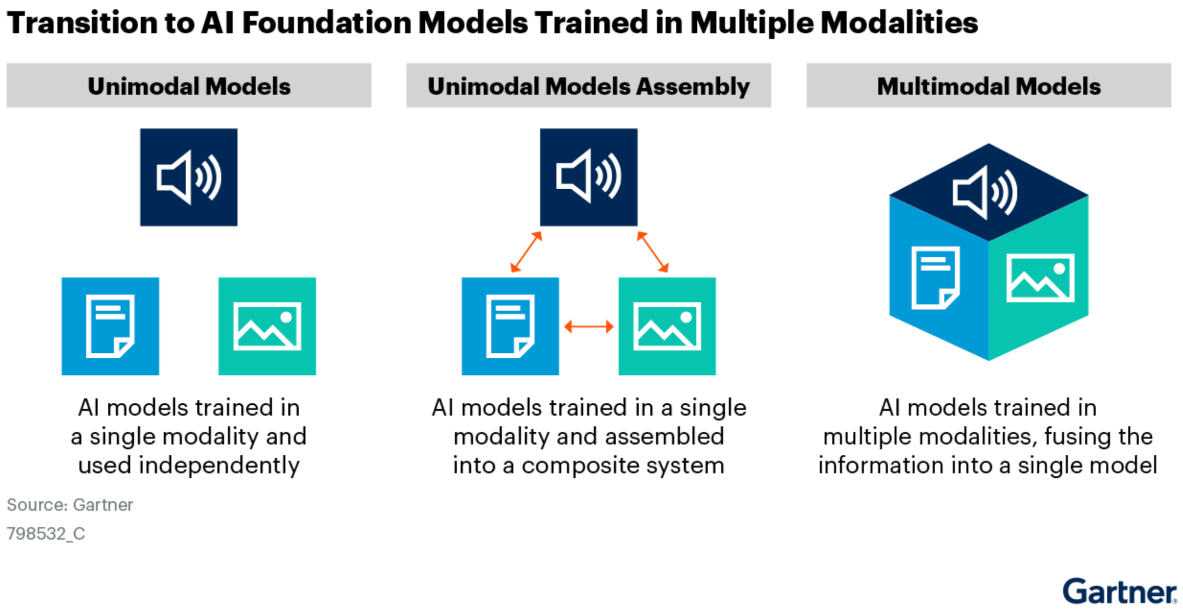
Description

> **Definition**
>
> Multimodal AI models are trained with multiple types of data (also known as modalities) simultaneously, such as images, video, audio and text. This enables them to create a shared data representation to improve performance in different tasks. At runtime, they can handle more than one modality, either in their inputs, their outputs or both.

Multimodality goes beyond assembling separate unimodal models together — this is a valuable approach, but it is *not* multimodality. To qualify as multimodal, AI models need to be *trained* in different modalities at once, which allows them to fuse all of the information, creating a shared representation and unlocking new capabilities.

Figure 2 describes this evolution from unimodal models, which can either be used independently or assembled together, to true multimodal models which have been trained across modalities.

Figure 2: Transition to AI Foundation Models Trained in Multiple Modalities


Transition to AI Foundation Models Trained in Multiple Modalities
Unimodal Models
Unimodal Models Assembly
Multimodal Models
AI models trained in a single modality and used independently
AI models trained in a single modality and assembled into a composite system
AI models trained in multiple modalities, fusing the information into a single model
Gartner

Source: Gartner
798532_C

There are different types of multimodal AI depending on the modalities that the models can handle as inputs and the ones that they can generate as outputs at runtime (see Figure 3):

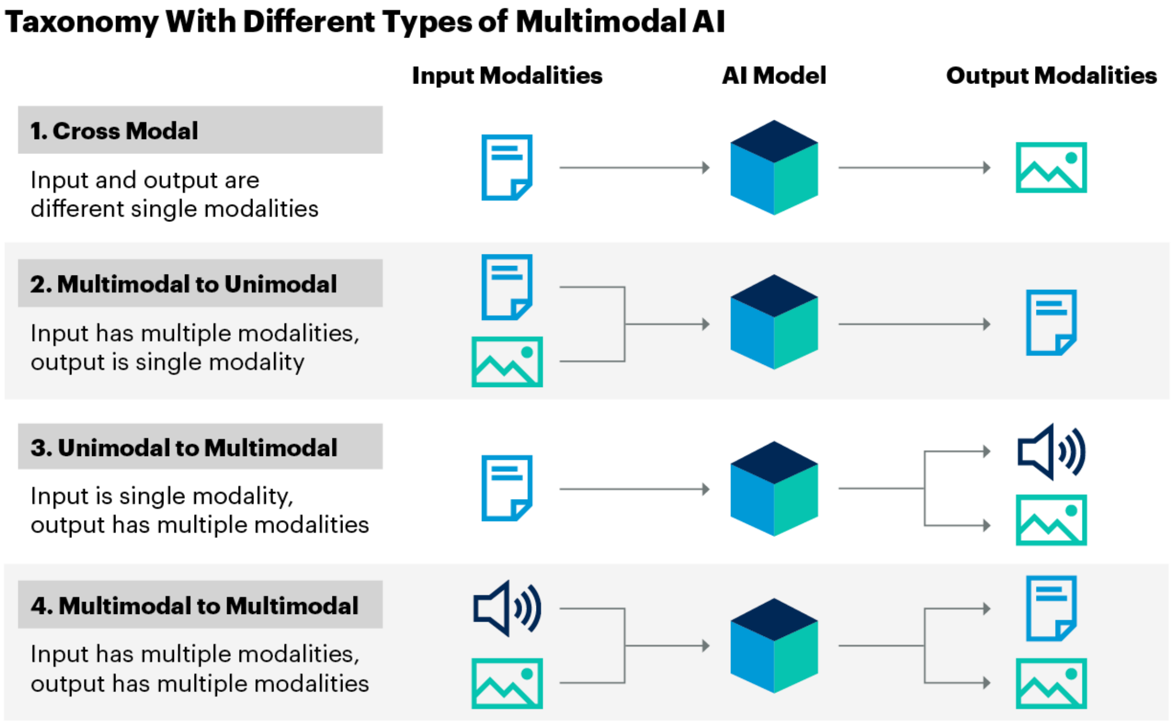
- **Cross-modal:** Input and outputs are from different single modalities. An example is image generation models, like DALL·E 3, which take descriptive text as input and generate images as outputs. [1]

- **Multimodal to unimodal:** Input has different modalities, but the output is from a single modality. An example is a model like GAIA-1 from Wayve, which takes video, text and action inputs to generate video as an output. [2]

- **Unimodal to multimodal:** Input is a single modality, whereas output has multiple modalities. An example is a model that takes text as input and generates both audio and text as outputs.

- **Multimodal to multimodal:** Both inputs and outputs have different modalities. An example is Meta's SeamlessM4T model, which can take both audio and text as inputs, translating speech to a different language, with either text or audio as outputs. [3]

Figure 3: Taxonomy With Different Types of Multimodal AI

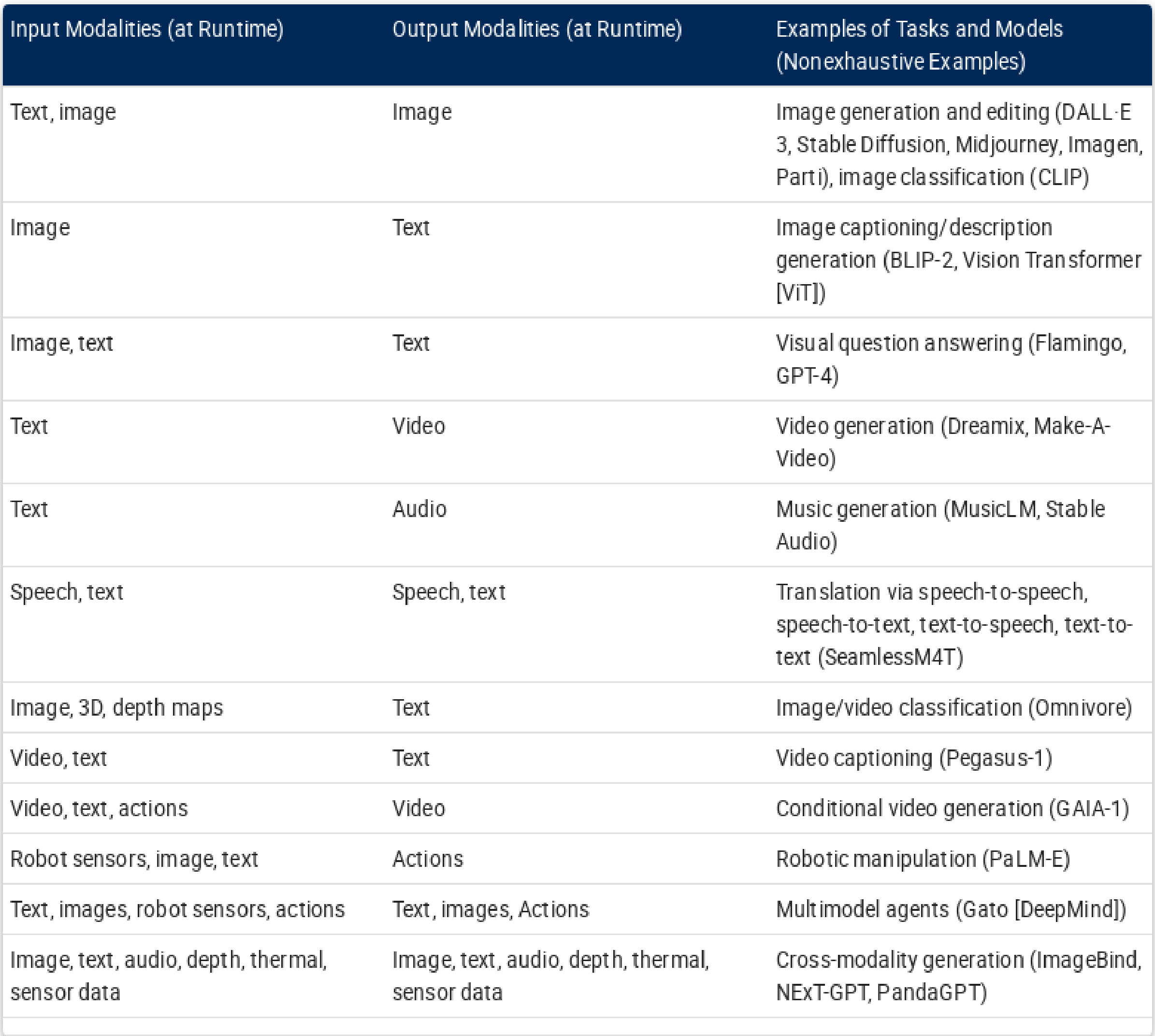

Taxonomy With Different Types of Multimodal AI
Input Modalities
AI Model
Output Modalities
1. Cross Modal
Input and output are different single modalities
2. Multimodal to Unimodal
Input has multiple modalities, output is single modality
3. Unimodal to Multimodal
Input is single modality, output has multiple modalities
4. Multimodal to Multimodal
Input has multiple modalities, output has multiple modalities

Source: Gartner
798532_C

Table 1 illustrates many more examples of multimodal tasks and existing models, classified according to their input and output modalities.

Table 1: Types of Multimodal AI Models (Nonexhaustive Examples)

(Enlarged table in Appendix)

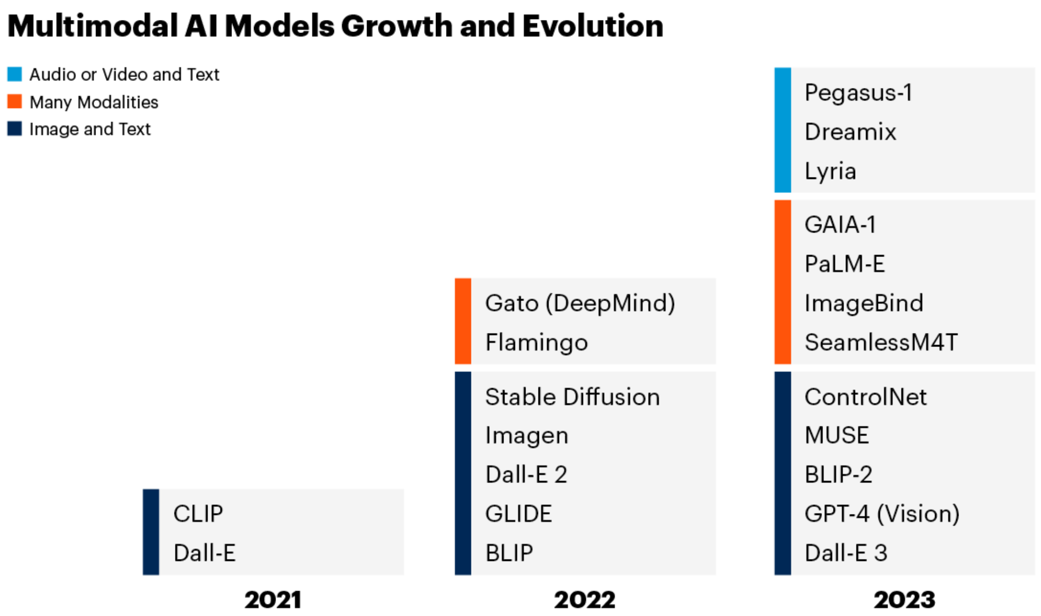


Input Modalities (at Runtime)	Output Modalities (at Runtime)	Examples of Tasks and Models (Nonexhaustive Examples)
Text, image	Image	Image generation and editing (DALL·E 3, Stable Diffusion, Midjourney, Imagen, Parti), image classification (CLIP)
Image	Text	Image captioning/description generation (BLIP-2, Vision Transformer [ViT])
Image, text	Text	Visual question answering (Flamingo, GPT-4)
Text	Video	Video generation (Dreamix, Make-A-Video)
Text	Audio	Music generation (MusicLM, Stable Audio)
Speech, text	Speech, text	Translation via speech-to-speech, speech-to-text, text-to-speech, text-to-text (SeamlessM4T)
Image, 3D, depth maps	Text	Image/video classification (Omnivore)
Video, text	Text	Video captioning (Pegasus-1)
Video, text, actions	Video	Conditional video generation (GAIA-1)
Robot sensors, image, text	Actions	Robotic manipulation (PaLM-E)
Text, images, robot sensors, actions	Text, images, Actions	Multimodel agents (Gato [DeepMind])
Image, text, audio, depth, thermal, sensor data	Image, text, audio, depth, thermal, sensor data	Cross-modality generation (ImageBind, NExT-GPT, PandaGPT)

Source: Gartner (November 2023)

The early examples of multimodality have been models that leverage two modalities, particularly images and text. Increasingly, multimodal models will learn from more than two modalities at once. Figure 4 provides an illustration of this evolution toward many modalities.

Figure 4: Multimodal AI Models Growth and Evolution



Source: Gartner
798532_C

Crucially, multimodal models can be used for generative AI tasks, but they can also be used for more traditional AI tasks like classification or regression. For example, a multimodal model can use customer service text reports, call audio and video of local weather conditions as inputs to better predict the next best action for a customer (a classification task). In practice, multimodal AI can be used across many tasks and use cases.

Benefits and Uses

Benefits

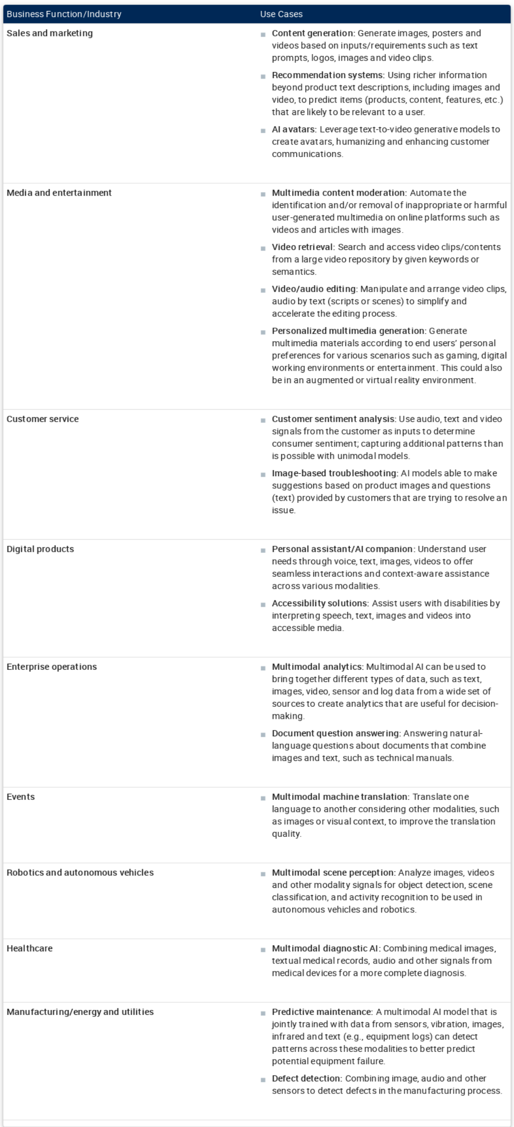
- **Increased robustness:** Multimodality can make AI foundation models more robust by reducing the dependency on a single modality; enabling them to work with noisy or missing data, and disambiguating information across modalities.
- **New use cases:** The ability to process data across modalities opens up many new AI use cases that were previously not possible, such as visual question-answering, multimodal data retrieval, and complex audio and video generation. These use cases are broadly applicable across many different industries and business functions (see Table 2).
- **New and improved user experience:** Multimodal AI enables richer forms of AI-human interaction across different modalities, combining voice, text, images and video. For instance, it is increasingly possible to interact with virtual assistants, like ChatGPT, via audio and images.
- **Improved performance:** By integrating information across modalities, multimodal models can improve performance for some use cases. For example, a customer sentiment model that leverages audio, text and video will be able to capture additional patterns in the data and outperform unimodal models.
- **Increased scalability:** Multimodal AI can be trained with orders-of-magnitude more data than AI foundation models trained on a single modality, such as LLMs. This increase in training data makes multimodal models more scalable as they can be trained with information in video and audio form, going beyond text data.

Uses

Multimodal models have applications across a broad set of business functions, with many potential use cases. Table 2 enumerates examples of existing use cases for multimodal AI.

Table 2: Multimodal Use Cases by Business Function

(Enlarged table in Appendix)



Business Function/Industry	Use Cases
Sales and marketing	**Content generation:** Generate images, posters and videos based on inputs/requirements such as text prompts, logos, images and video clips. **Recommendation systems:** Using richer information beyond product text descriptions, including images and video, to predict items (products, content, features, etc.) that are likely to be relevant to a user. **AI avatars:** Leverage text-to-video generative models to create avatars, humanizing and enhancing customer communications.
Media and entertainment	**Multimedia content moderation:** Automate the identification and/or removal of inappropriate or harmful user-generated multimedia on online platforms such as videos and articles with images. **Video retrieval:** Search and access video clips/contents from a large video repository by given keywords or semantics. **Video/audio editing:** Manipulate and arrange video clips, audio by text (scripts or scenes) to simplify and accelerate the editing process. **Personalized multimedia generation:** Generate multimedia materials according to end users' personal preferences for various scenarios such as gaming, digital working environments or entertainment. This could also be in an augmented or virtual reality environment.
Customer service	**Customer sentiment analysis:** Use audio, text and video signals from the customer as inputs to determine consumer sentiment; capturing additional patterns than is possible with unimodal models. **Image-based troubleshooting:** AI models able to make suggestions based on product images and questions (text) provided by customers that are trying to resolve an issue.
Digital products	**Personal assistant/AI companion:** Understand user needs through voice, text, images, videos to offer seamless interactions and context-aware assistance across various modalities. **Accessibility solutions:** Assist users with disabilities by interpreting speech, text, images and videos into accessible media.
Enterprise operations	**Multimodal analytics:** Multimodal AI can be used to bring together different types of data, such as text, images, video, sensor and log data from a wide set of sources to create analytics that are useful for decision-making. **Document question answering:** Answering natural-language questions about documents that combine images and text, such as technical manuals.
Events	**Multimodal machine translation:** Translate one language to another considering other modalities, such as images or visual context, to improve the translation quality.
Robotics and autonomous vehicles	**Multimodal scene perception:** Analyze images, videos and other modality signals for object detection, scene classification, and activity recognition to be used in autonomous vehicles and robotics.
Healthcare	**Multimodal diagnostic AI:** Combining medical images, textual medical records, audio and other signals from medical devices for a more complete diagnosis.
Manufacturing/energy and utilities	**Predictive maintenance:** A multimodal AI model that is jointly trained with data from sensors, vibration, images, infrared and text (e.g., equipment logs) can detect patterns across these modalities to better predict potential equipment failure. **Defect detection:** Combining image, audio and other sensors to detect defects in the manufacturing process.

Source: Gartner (November 2023)

Risks

Security risks for multimodal AI:

- **New security attack surface:** New modalities can also bring in new attack surfaces. Malicious users could leverage different modality inputs to jailbreak the models and carry out prompt injection — for example, by inserting hard-to-detect text in images in order to manipulate LLMs to execute a malicious instruction.
- **Data exposure:** Multimodal AI increases the exposure to a wider range of sensitive data. Examples of particularly sensitive data types include: maps or geolocation data, biometric data or health data.

Complexity of multimodal AI:

- **Data management:** Multimodal data needs to be aligned and integrated. Multimodal data is more complex as it has varying degrees of quality and formats compared to unimodal data.
- **Deviation/misalignment across modalities:** The nuances and logic in one modality can be challenging for multimodal models to reflect in other modalities (e.g., some subtle objects in images may not be well perceived and described in text).
- **Training complexity:** Multimodal models use complex techniques to train, integrate and analyze data sourced from the multiple modalities. This is not a risk for the use of multimodal AI, but rather a challenge for training or fine-tuning these models.

Readiness for multimodal AI:

- **Vendor maturity:** The vendor landscape is nascent, and current enterprise adoption is low. These factors could present a challenge when you are running initial multimodal pilots.
- **Cost and computational efficiency:** Given the use of multiple types of data, multimodal models will likely be more expensive to train and to run.
- **Talent:** A lack of awareness about multimodal AI and a lack of technical skills to leverage and embed these models could prevent organizations from realizing value from this technology.

Adoption Rate

Multimodal AI is still nascent and we believe that fewer than 1% of businesses have deployed it in production. However, we have seen an increased proliferation of multimodal foundation models since 2021, which has been accelerating over the last year. Technology vendors are investing in creating models that can work across many different modalities.

This is a fast-moving field, and the adoption rate is likely to accelerate in the future. Enterprises are likely to adopt multimodal models once they become more widely available in the market given the benefits discussed in this research.

Recommendations

- Educate your AI team on multimodality, introducing the concept, its benefits and risks. Break AI technical silos by encouraging AI experts to work on AI projects outside of their AI technical specialization, such as natural language process and computer vision. Include multimodality as a key topic in your AI community of practice. Expose AI teams to vendors that focus on multimodal models as part of this education process.

- Identify AI use cases in your organization where a model that combines multiple data modalities could generate business value above and beyond what is currently possible with unimodal AI foundation models. Run pilots with off-the-shelf multimodal models to demonstrate, not only their technical feasibility, but the business value that can be generated with these models.

- Start leveraging multimodal AI in use cases, such as image generation, where these models are already performing better than unimodal ones. Monitor the feasibility of emerging use cases, such as video generation, as the technology evolves.

- Assess the technical complexities of processing and integrating data inputs and outputs from diverse multimodal sources. Identify and mitigate risks around data privacy and security through robust data engineering and data governance.

Representative Providers

- Google
- Huawei
- Jina AI
- Meta
- Midjourney
- Nvidia
- OpenAI
- Stability AI
- TwelveLabs
- Wayve

Evidence

[1] Improving Image Generation With Better Captions, OpenAI.

[2] GAIA-1: A Generative World Model for Autonomous Driving, Wayve.

[3] Introducing SeamlessM4T, a Multimodal AI Model for Speech and Text Translations, Meta.

Note 1: Definition of Modality

A technical definition of modality is that two data sources come from two different modalities if they cannot be mapped unambiguously onto each other by an algorithm without losing information. For example, a picture in two different image formats (e.g., PNG, JPEG) does not represent different modalities since they contain the same information. On the other hand, an audio file from a conversation and its text transcription represent different modalities since there is information that is not present in the text transcription (e.g., intonation, emotional cues, etc.). See What Is Multimodality?, Heidelberg University.

Recommended by the Authors

Some documents may not be available as part of your current Gartner subscription.

Use Multistructured Analytics for Complex Business Decisions

Innovation Guide for Generative AI in Computer Vision

Table 1: Types of Multimodal AI Models (Nonexhaustive Examples)

Input Modalities (at Runtime)	Output Modalities (at Runtime)	Examples of Tasks and Models (Nonexhaustive Examples)
Text, image	Image	Image generation and editing (DALL·E 3, Stable Diffusion, Midjourney, Imagen, Parti), image classification (CLIP)
Image	Text	Image captioning/description generation (BLIP-2, Vision Transformer [ViT])
Image, text	Text	Visual question answering (Flamingo, GPT-4)
Text	Video	Video generation (Dreamix, Make-A-Video)
Text	Audio	Music generation (MusicLM, Stable Audio)
Speech, text	Speech, text	Translation via speech-to-speech, speech-to-text, text-to-speech, text-to-text (SeamlessM4T)
Image, 3D, depth maps	Text	Image/video classification (Omnivore)
Video, text	Text	Video captioning (Pegasus-1)
Video, text, actions	Video	Conditional video generation (GAIA-1)
Robot sensors, image, text	Actions	Robotic manipulation (PaLM-E)
Text, images, robot sensors, actions	Text, images, Actions	Multimodel agents (Gato [DeepMind])
Image, text, audio, depth, thermal, sensor data	Image, text, audio, depth, thermal, sensor data	Cross-modality generation (ImageBind, NExT-GPT,

	PandaGPT)

Source: Gartner (November 2023)

Table 2: Multimodal Use Cases by Business Function

Business Function/Industry	Use Cases
Sales and marketing	■ **Content generation:** Generate images, posters and videos based on inputs/requirements such as text prompts, logos, images and video clips. ■ **Recommendation systems:** Using richer information beyond product text descriptions, including images and video, to predict items (products, content, features, etc.) that are likely to be relevant to a user. ■ **AI avatars:** Leverage text-to-video generative models to create avatars, humanizing and enhancing customer communications.
Media and entertainment	■ **Multimedia content moderation:** Automate the identification and/or removal of inappropriate or harmful user-generated multimedia on online platforms such as videos and articles with images. ■ **Video retrieval:** Search and access video clips/contents from a large video repository by given keywords or semantics. ■ **Video/audio editing:** Manipulate and arrange video clips, audio by text (scripts or scenes) to simplify and accelerate the editing process. ■ **Personalized multimedia generation:** Generate multimedia materials according to end users' personal preferences for various scenarios such as gaming, digital working environments or entertainment. This could also be in an augmented or virtual reality environment.

Customer service	■ **Customer sentiment analysis:** Use audio, text and video signals from the customer as inputs to determine consumer sentiment; capturing additional patterns than is possible with unimodal models. ■ **Image-based troubleshooting:** AI models able to make suggestions based on product images and questions (text) provided by customers that are trying to resolve an issue.
Digital products	■ **Personal assistant/AI companion:** Understand user needs through voice, text, images, videos to offer seamless interactions and context-aware assistance across various modalities. ■ **Accessibility solutions:** Assist users with disabilities by interpreting speech, text, images and videos into accessible media.
Enterprise operations	■ **Multimodal analytics:** Multimodal AI can be used to bring together different types of data, such as text, images, video, sensor and log data from a wide set of sources to create analytics that are useful for decision-making. ■ **Document question answering:** Answering natural-language questions about documents that combine images and text, such as technical manuals.
Events	■ **Multimodal machine translation:** Translate one language to another considering other modalities, such as images or visual context, to improve

	the translation quality.
Robotics and autonomous vehicles	- **Multimodal scene perception:** Analyze images, videos and other modality signals for object detection, scene classification, and activity recognition to be used in autonomous vehicles and robotics.
Healthcare	- **Multimodal diagnostic AI:** Combining medical images, textual medical records, audio and other signals from medical devices for a more complete diagnosis.
Manufacturing/energy and utilities	- **Predictive maintenance:** A multimodal AI model that is jointly trained with data from sensors, vibration, images, infrared and text (e.g., equipment logs) can detect patterns across these modalities to better predict potential equipment failure. - **Defect detection:** Combining image, audio and other sensors to detect defects in the manufacturing process.

Source: Gartner (November 2023)