Building Multimodal AI: Fusion Techniques, Architectures, and Challenges in LLM-Driven Systems

1st Andrew Gerstenslager
Dept. of Computer Science
University of Cincinnati
Cincinnati, United States
gerstead@mail.uc.edu

2nd Alex Johnson
Dept. of Computer Science
University of Cincinnati
Cincinnati, United States
johns9a4@mail.uc.edu

3rd Walaa Alshammari
Dept. of Computer Science
University of Cincinnati
Cincinnati, United States
alshamwm@mail.uc.edu

***Abstract*—Multimodal AI systems integrate information from diverse sources—such as text, images, audio, video, and sensor data—to create more robust, context-aware models. As Large Language Models (LLMs) evolve into Multimodal LLMs (MM-LLMs), they increasingly serve as a central foundation for cross-modal reasoning. This paper systematically reviews core components of modern multimodal AI, including representation, tokenization, fusion strategies, architectural paradigms, and training methodologies. We trace the field's evolution from early vision-language encoders like CLIP and BLIP to unified any-to-any architectures such as GPT-4o, Gemini 2.5, and Qwen 2.5-Omni. Key challenges—such as missing modalities, cross-modal hallucination, computational cost, adversarial vulnerabilities, and ethical risks—are critically examined. We further present a comparative evaluation across major benchmarks using VLMEvalKit, summarize state-of-the-art models, and highlight emerging trends like embodied AI and proactive governance. Together, this paper serves as both a historical survey and a discussion of the open problems that persist in this area of research.**

***Index Terms*—Multimodal AI, Large Language Models (LLMs), Multimodal Large Language Models (MM-LLMs), Cross-Modal Fusion, Cross-Modal Reasoning, Multimodal Alignment, Multimodal Benchmarking, Any-to-Any Models, Embodied AI**

I. Introduction

Multimodal Artificial Intelligence (AI) refers to AI systems that integrate and process information from multiple data sources or "modalities", such as text, images, audio, video, sensor data, and more. By integrating input from different modalities, these systems aim to mimic human-like understanding, robustness, and contextual awareness. Large Language Models (LLMs) have recently emerged as powerful foundational architectures for multimodal AI systems to create unified frameworks that can process and reason over diverse inputs. Through advanced representation learning and cross-modal alignment techniques, these models can perform tasks such as image captioning, visual question answering, and multimodal dialogue with improved accuracy and contextual reasoning across modalities.

A clear example of this technology in practice is the Google Nest Hub, which integrates the multimodal Gemini model to process audio and video inputs [1]. When paired with a security camera, the system allows the Google Assistant to answer questions about the video feed and past recorded events, demonstrating real-time integration of visual and auditory information.

Over the last several years, the development of LLMs has accelerated progress in multimodal AI. Early breakthroughs like CLIP [2] (Jan 2021) aligned text and image representations using contrastive learning. This was extended by models such as BLIP [3] (Jan 2022), which added captioning and retrieval objectives. These early systems paved the way for the first LLM-based multimodal models, which integrated vision and text, such as BLIP-2 [4] (Jan 2023), Kosmos-1 [5] (Feb 2023) and Flamingo [6] (Apr 2022). Building on this, frameworks like Unified-IO [7] (Jun 2022) and its extension Unified-IO 2 [8] (Dec 2023) emerged, offering more flexible architectures to process multiple modalities in both input and output. This progression advanced further with models incorporating more than two modalities, such as GPT-4o [9] (May 2024), and Gemini [10] (Dec 2023), which integrate text, images, video, and audio. The latest frontier includes multimodal models that are generalist agents like GPT-o3 [11] (Jan 2025), Gemini 2.5 [12] (Feb 2025), and Qwen2.5-Omni [13] (May 2025), designed to tackle a broad range of tasks across multiple domains and modalities.

This paper explores the evolution of multimodal AI systems, tracing their development from foundational research to current state-of-the-art (SOTA) models. The remainder of the paper is organized as follows: Section II reviews the foundational concepts of multimodal learning, while Section III surveys the major modalities and their implications for model design. Section IV details representation, tokenization, and embedding methods that bridge raw signals and language-model token streams. Section V discusses key multimodal techniques and architectural paradigms, including alignment and fusion. Section VI summarizes prevailing pre-training, fine-tuning, and instruction-alignment strategies. Section VII highlights the principal tasks and application domains enabled by multimodal AI, and Section VIII explains the datasets,

metrics, and leaderboards used for rigorous evaluation. Section IX provides concise case studies of recent frontier and state-of-the-art models, whereas Section X examines current limitations and open technical challenges. Section XI explores safety, ethical, and societal considerations, Section XII outlines promising future research directions, and Section XIII concludes the paper.

II. Background and Foundational Concepts

Multimodal AI integrates data from text, images, audio, video, and sensors into systems that can reason across them [14], [15]. Because each source is encoded differently, specialized representation and integration methods are required [16].

Modalities are distinct channels of information such as text, images, video, audio, or sensor data. Each modality has different features, such as dimensionality, sampling rate, and statistical structure. Text is a sequence of discrete symbols. Images are 2D or 3D grids of pixel values. Video is a sequence of images over time. Audio is a 1D waveform that can also be represented as a time-frequency spectrogram. Sensor data refers to numerical readings collected from physical devices such as accelerometers, gyroscopes, or LIDAR. Given each modality has its own distinct structure, they must be converted from raw signals into a common representation for models to process them together [14].

Tokenization and Embedding map raw modality specific inputs to numerical representations, such as vectors, that can be processed by models. Text is first split into tokens with methods such as Byte Pair Encoding, then embedded in a vector space [17]. Images may be divided into patches for a Vision Transformer encoder [18] or passed through a CNN to produce feature embeddings [2]. Once converted to numerical representations, information from different modalities can be aligned and processed jointly.

Fusion describes the process of combining multiple modalities into a single representation such that cross-modal relationships can be captured. Fusion strategies vary based on when and how data is combined, ranging from early fusion (integrating modalities before initial processing) to late fusion (combining independently processed modalities at the output stage), to hybrid approaches which use some combination of the two [19]–[21].

Transformer Architectures have revolutionized multimodal AI, starting from text-focused models like BERT [17] and ChatGPT [22] to multimodal variants such as ViLT [18], LXMERT [23], and SimVLM [24]. Transformers use self-attention mechanisms to capture long-range dependencies in data, which is ideal for processing sequences like text or image patches. Extending self-attention mechanisms across modalities allows for transformer based multimodal models to capture cross-modal dependencies, making them suitable for jointly processing and aligning multimodal inputs.

Contrastive Learning is a common training strategy, exemplified by CLIP [2] and ALIGN [25], where models learn to align related data pairs (e.g., images and their textual descriptions) closely in embedding space, while distancing unrelated pairs. This technique can be used with transformer architectures to enhance multimodal alignment.

Input vs. Output Modalities: One key distinction is whether a modality is used for *input*, *output*, or both. For instance, a model might take images, text, and audio as inputs, but only generate text outputs (e.g., captions). Some advanced systems aim for *any-to-any* transformations, where any combination of input modalities can generate different output modalities, such as text-to-image or image-to-audio.

III. Modalities in Multimodal AI

Multimodal Large Language Models (MM-LLMs) differ primarily in which modalities they accept or generate, and how those modalities are aligned within their architecture. This section covers common types of MM-LLMs, beginning with vision-text models, then expanding to models incorporating more specialized data types like video audio or sensor inputs.

A. Vision-Text MM-LLMs

Early multimodal large language models (MM-LLMs) focused on vision and text modalities due to the availability of large-scale paired datasets and pre-trained models [26]. Notable early examples include *Flamingo*, *PaLI*, and *BLIP-2* [4], [6], [27].

Flamingo uses both a frozen Vision Transformer (ViT) and Chinchilla language model for images and text representations respectively. To align the two modalities a Perceiver Resampler is used which turns the patch embeddings from the ViT into a set of 'visual tokens' that can be interleaved with text during generation. This approach leverages late fusion and modality specific encoders [6].

PaLI takes a much different approach, training a single encoder-decoder transformer on multilingual text and ViT image patches. This fusion approach creates fully joint embeddings, supporting over 100 languages. However, the multilingual dataset increases training overhead [27].

BLIP-2 uses a lightweight Q-Former to connect a *CLIP* vision encoder to a frozen LLM's embedding space, with the adapter comprising less than 10% of the model's parameters, an example of late fusion and modality specific embeddings [4].

While recent AI progress incorporates additional modalities like video and audio, many state-of-the-art models in 2025 focus on vision and text. Prominent examples include *OpenAI's o3*, *Anthropic's Claude 3.7 Sonnet*, and *DeepSeek AI's DeepSeek-VL2* (detailed further in Section IX). These models are reasoning models, generating Chain of Thought (CoT) outputs to address complex problems [11], [28], [29].

B. Expanding Beyond Vision Text

The development of MM-LLMs in recent years has evolved beyond the classic vision-text combination to integrate other modalities such as video and audio, which differ in structure compared to text and vision as they are temporally structured. Integrating these temporally structured modalities is a more

difficult problem, as it requires modeling temporal dependencies, aligning sequences across time, and compressing high-dimensional data into representations that can interact with static modalities like text and images.

Yet these modalities unlock new applications and use cases for MM-LLMs not possible with simple vision and text. One key capability that these modalities allow for is real time interaction whether through audio alone or both audio and video. For example, models that support audio can become voice assistants capable of having real time conversations. Additionally, models that incorporate video can respond to visual information in real time such as a phone camera feed or a desktop screen, allowing for entirely new interaction paradigms.

This subsection surveys four MM-LLM systems that incorporate text, image, video and audio modalities.

NExT-GPT is an any-to-any MM-LLM capable of both understanding and generating content across text, image, video and audio modalities. Its architecture integrates three stages: (1) modality-specific encoding stage where pretrained encoders (primarily ImageBend) extract features from input modalities are passed through transformer-based input projection layers to combine features into semantic concept tokens; (2) a core Vicuna 7B LLM interprets the concept tokens for cross modal reasoning and instruction following; (3) a decoding stage, where the LLM produces standard text and special modality signal tokens for pretrained diffusion decoders (Stable Diffusion, AudioLDM, Zeroscope). The model uses hybrid fusion and modality specific encoders which are projected into the shared LLM-compatible semantic space [30].

GPT-4o ("omni") utilizes a single end-to-end transformer trained across text, image, video, and audio, enabling deep, early fusion and real-time interaction capabilities (detailed further in Section IX) [9].

Gemini 2.5 incorporates text, image, video, and audio inputs, potentially using a Mixture-of-Experts (MoE) architecture to manage the diverse data streams, although this is not confirmed by Google (detailed further in Section IX) [12].

Qwen 2.5-Omni uses a dual-module architecture and introduces techniques like time-aligned Multimodal RoPE (TM-RoPE) specifically to handle temporal alignment between audio and vision (detailed further in Section IX) [13].

C. Embodied and Robotics MM-LLMs

This section focuses on MM-LLMs that are designed specifically for robotic and embodied systems. Unlike general purpose models, these systems can generate commands for physical devices, such as electric servos, to produce physical actions. These models leverage multimodal capabilities such as text and audio for commands, as well as video and sensor data for spatial awareness and proprioception. Additionally, these models must consider temporal grounding, closed-loop latency and safety constraints for real world applications.

PaLM-E is a generalist embodied language model developed by Google that extends the PaLM architecture to support robot perception and planning. It uses a decoder-only transformer with adapter modules to process multimodal input including camera images, 3D robot state, sensor data and text-based commands or questions. For images a Visual Transformer (ViT) adapter module is used and for robot state and sensor data a Multilayer Perceptron (MLP) is used. The output of these adapters is projected into the same embedding space as the language tokens and interleaved with text to form a sequence of multimodal tokens. These tokens are then jointly processed by the transformer from the very first layer, enabling early fusion of modalities. The model is trained to autoregressively generate text in the form of visual question answers, descriptions of the environment, or high-level plans for robotic control. These outputs allow for a semantic understanding of the robot's perception and context, however, are not used directly for robotic control, instead the outputs are interpreted by task-specific controllers [31].

RT-2 (Robotics Transformer 2) is a vision-language-action (VLA) model that extends PaLM-E by enabling direct robot control through token outputs that directly correspond to action outputs. Compared to PaLM-E, RT-2 offers improved generalization to unseen tasks, integrates chain-of-thought reasoning for complex instructions, and unifies perception, language, and control within a single model architecture (detailed further in Section IX)

Helix is a vision-language action model developed by Figure AI to enable humanoid robots to perform complex tasks using integrated perception, language understanding and precise motor control. The model uses dual transformer architecture consisting of System 2 (S2), used for high level planning and System 1 (S1), used for fast reactive motor manipulation (detailed further in Section IX) [32].

D. Strengths and Limitations

Adding more than two modalities increases the difficulty significantly. The model must handle very different types of data formats, deal with the timing and sequence in inputs like audio and video, and bridge larger semantic gaps between diverse inputs. This makes both alignment and fusion much more complicated compared to handling only text and images.

The advantage of incorporating additional modalities allows the model to develop a more complete understanding of the world. These models tend to perform better on complex tasks that require reasoning across different types of input, are more resilient when some input data is missing, and offer greater flexibility in handling a wider variety of formats for both input and output. However, there are trade-offs. Models with more modalities become more complex in both architecture and training. They require more computational resources, face bigger challenges in finding enough high-quality and well-aligned data, and can suffer from bias toward certain modalities. Evaluating them accurately is harder, and there is a greater risk of catastrophic forgetting, where the model might lose previously learned skills if it is not trained carefully [33], [34].

E. Trends Toward Any-to-Any Models

Modern research is pushing toward fully integrated, *unified* models that treat all inputs and outputs as sequences of tokens. Systems like Unified-IO [7], [8], NeXT-GPT [30], and GPT-4o [9] process diverse input modalities by first encoding them into a shared token space, and likewise generate outputs in a tokenized fashion. This approach greatly enhances flexibility, enabling models to perform a wide variety of tasks across different modality combinations within a single framework.

IV. Representation, Tokenization, and Embedding

This section explains the strategies used to convert different modalities (text, image, etc.) into tokens, and if those modalities are aligned together in an LLM or combined.

- **Text Tokenization** When working with text, the usual first step is to break it down into small parts called tokens—these could be sub-words or even characters, depending on the tokenizer used. Techniques like Byte Pair Encoding (BPE) or WordPiece help split the input into manageable chunks. Each token is then converted into a vector using an embedding matrix, which is how models like GPT process and understand language. This method is standard for text and helps highlight the difference when dealing with other input types like images or audio, where tokenization isn't as straightforward [14]
- **Image Tokenization** Images start off as raw pixels and need to be converted into meaningful inputs for models, especially those based on LLMs. One common way to do this is through a visual encoder like a CNN or a Vision Transformer (ViT). CNNs may produce pooled vectors or feature maps, while ViTs split the image into patches—say, 16×16 pixels—and convert them into patch embeddings. These embeddings act like the image version of word vectors. For instance, CLIP uses a ViT or ResNet to turn an image into a vector, and its text encoder does the same for captions. It then aligns both in a shared space, where a picture of a cat and the word "cat" appear close together. In LLM-based setups, like PaLM-E, image features are taken from a pre-trained encoder and projected into the same embedding format used by the language model. This makes it possible to treat visual information as if it were a sequence of text-like tokens, so the LLM can process everything together [14].
  An alternative method is to convert images into discrete tokens, kind of like turning visual content into a string of visual "words." This can be done using a model like VQ-VAE, which maps image patches to code book indices. Each index corresponds to a learned visual token. Unified-IO uses this approach to make all input types—text, images, even things like depth maps—fit into a shared token vocabulary. This way, a model can handle everything with the same architecture. While this method isn't as common in LLM-focused systems, it shows the extreme end of unifying different modalities into one format [8], [30].
- **Other Modalities (Audio, Video, etc.)** Other types of input, like audio or video, follow similar ideas. Audio can be turned into a spectrogram or passed through a model like CLAP to get a vector embedding. Video can be processed by analyzing frames individually or through specialized video encoders. Regardless of the type, the key idea is the same: everything must be converted into numbers—either as vectors or tokens—before the model can use it. [14]
  As a concrete example on a known LLMs, which presents the above mentioned strategies; **CLIP** uses separate encoders for image and text and aligns their outputs in a shared embedding space using contrastive learning [2], [14]. It doesn't use an LLM but set the stage for future multimodal work. **BLIP-2** uses a visual encoder and a Q-Former to convert images into 32 query tokens that can be processed by a frozen LLM [4]. This approach shows how to extend a language model to new modalities with minimal changes. **Unified-IO** converts everything into tokens using a shared vocabulary, which lets a single Transformer handle a wide range of input types [8]. It represents one of the most unified and general approaches. **PaLM-E** keeps things simple by projecting image data into the LLM's token space and prepending it to the text prompt. It shows that you can add multimodal capability to a massive language model without retraining it from scratch [31].

A. Joint Embedding Space vs. Modality-Specific Embeddings:

A key idea in multimodal representation is whether the model brings different types of inputs into the same space or keeps them separate. Some models like CLIP are trained in a way that both image and text encoders produce vectors that live in the same space. For example, if you have a picture of a cat and the word "cat," their representations end up close together in the same latent space. On the other hand, models like GPT-4V or PaLM-E don't always use a shared space [31], [35]. They feed the image features into the language model, and the alignment between vision and text happens later, inside the model itself. This alignment can be learned either directly—using a training objective that forces it, or just by training on enough paired data. For example, in BLIP-2, the first training stage teaches the Q-Former to output image embeddings that line up well with the corresponding text, which is similar to what CLIP does [2], [4]. Either way, the point is to make sure the model can compare or combine inputs from different modalities in a meaningful way [14].

B. Tokenization Strategies for LLM Integration

When combining non-text data with a large language model, there are a few ways to do it. One approach is to use a module that turns image features into tokens that look like text tokens. BLIP-2 does this using a Q-Former, which takes image embeddings and outputs a handful of query tokens that can be fed into the LLM alongside regular text [14]. These tokens match the expected dimensions, so they integrate smoothly.

PaLM-E uses a similar approach by projecting image features into the same token format as language input and simply placing them before the text in the prompt [31].

Another method is to handle fusion deeper inside the model using cross-attention. This is what DeepMind's Flamingo does—it takes visual features and lets the language model attend to them through dedicated attention layers [6]. Instead of turning the image into tokens right away, the model integrates visual data throughout the network. This can lead to more nuanced interactions between image and text but adds complexity and training challenges. In fact, some recent studies show that simpler fusion methods, like prepending visual tokens, can work just as well for many tasks and are easier to train [14].

V. Core Multimodal Techniques & Architectures

Building effective Multimodal Large Language Models (MM-LLMs) requires sophisticated methods for handling and integrating information from diverse data sources. This section details the fundamental techniques for aligning representations across modalities, strategies for fusing this information, and the dominant architectural paradigms used in modern systems.

A. Alignment Techniques

Alignment is the crucial process of establishing semantic correspondence between representations learned from different modalities (e.g., ensuring an image of a dog is linked meaningfully to the word "dog"). Key techniques include:

- **Mapping Modalities into LLM Space:** To leverage pre-trained LLMs, non-textual modalities are often mapped into the LLM's input embedding space. This typically involves:
  - *Projection Layers:* Simple linear layers or small MLPs can transform features from a unimodal encoder, like a Vision Transformer, into vectors compatible with the LLM's token embeddings [36].
  - *Querying Transformers / Adapters:* More sophisticated modules like the Q-Former in BLIP-2 [4] use learnable queries and cross-attention to distill salient features from a modality (like vision) into a fixed set of summary tokens that the LLM can process efficiently. PaLM-E [31] uses adapters for various sensor inputs.

  These transformed features act as "soft prompts" or special input tokens for the LLM.
- **Cross-Attention Mechanisms:** Models like Flamingo [6] integrate modalities more deeply by adding cross-attention layers within the LLM. These allow the LLM's representations to directly attend to features from other modalities at different processing stages, allowing more nuanced cross-modal reasoning capabilities.
- **Multimodal Instruction Tuning:** A critical fine-tuning stage is where models are trained on datasets pairing multimodal inputs with desired textual outputs (can be instructions or questions). This step aligns the model's generative behavior with user expectations across various modalities, significantly improving performance on interactive tasks [36], [37].

Successful alignment allows the model to ground concepts across modalities, forming the basis for complex cross-modal understanding and generation.

B. Fusion Strategies

Fusion refers to the mechanism and timing of combining information from different (aligned) modalities within the model architecture [19], [20]. Common strategies include:

- **Early Fusion (Input/Feature-Level):** Information is combined at the beginning of the processing pipeline. This often involves concatenating raw or extracted features/token embeddings from different modalities into a single input representation processed by a unified model component [21]. Considering projected visual tokens to text embeddings before the first layer of an LLM, as in PaLM-E [31], is a form of early fusion. It allows for capturing low-level interactions early but can be sensitive to missing modalities or structural differences.
- **Late Fusion (Decision/Score-Level):** Each modality is processed through independent pathways initially. The outputs or high-level representations from these pathways are combined only near the end, often for a final prediction or decision. This approach is modular and robust but may miss complex inter-modal dependencies learned through joint processing.
- **Hybrid Fusion (Intermediate):** This strategy combines elements of both early and late fusion, allowing interaction between modalities at intermediate stages of the model. Examples include using cross-attention layers within the main model (as in Flamingo [6]) or having separate encoders feed into a central reasoning module (like the LLM in NExT-GPT [30]) which then drives separate decoders. Many modern, complex MM-LLMs utilize hybrid approaches to balance integration depth and modularity [8].

The choice depends on task requirements, modalities, and desired trade-offs between depth and architectural complexity.

C. Architectural Paradigms

MM-LLMs are implemented using several overarching architectural designs:

- **Modular (Encoder-Adapter-LLM):** A highly popular and parameter-efficient approach. It combines pre-trained unimodal encoders (like CLIP or ViT) with a pre-trained LLM (like Llama or Vicuna), freezing both. A lightweight 'adapter' or 'projector' module is trained to map the encoder's outputs into the LLM's input space [2], [4]. This facilitates leveraging powerful existing models with relatively little multimodal training data. These are often decoder-only.
- **Integrated with Architectural Modifications:** Some models modify a base LLM architecture to intrinsically

handle multimodal inputs. Flamingo's [6] addition of gated cross-attention layers is an example of this, modality fusion directly into the transformer blocks. This allows deeper integration but requires more complex training and architectural design.

- **End-to-End Unified Models:** The goal here is a single, cohesive model (often a transformer) trained jointly across multiple modalities, processing diverse inputs typically tokenized into a common sequence format. Unified-IO [7], [8] and the design philosophy behind models like GPT-4o [9] lean towards this paradigm. While offering maximum flexibility for 'any-to-any' tasks, they usually demand significant computational resources and large-scale multimodal pre-training datasets. Both encoder-decoder (like PaLI [27]) and decoder-only variants exist.

While modular designs are currently prevalent due to efficiency, the trend explores increasingly integrated and unified architectures for greater capability and flexibility.

VI. Training Strategies

A. Pre-training on Multimodal Data

Most of the multimodal models today begin with a pre-training stage. This is where the model works on large-scale datasets that include paired information, such as images with captions. Pre-training is important because it builds a foundation, giving the model a general understanding of how images and text relate. Two common strategies are described below.

Contrastive Learning: This is how CLIP was trained, where the model sees many image–text pairs and learns to bring correct pairs closer together in the representation space, while pushing unrelated ones apart. This method requires a lot of data (CLIP used 400 million pairs), but it gives the model a strong sense of cross-modal similarity [2]. ALIGN [25] took a similar approach and showed that scaling to one billion noisy image–text pairs can compensate for lower-quality data.

Captioning or Generative Objectives: Other models focus on generating descriptions from images. For example, the model might be shown an image and asked to produce a caption, which teaches it to interpret visual content and express it in language. Models like BLIP [3] combine both contrastive pre-training and a captioning objective, first aligning the modalities and then learning to generate captions. Similarly, BLIP-2 [4] extends this approach by freezing certain large encoders and introducing a lightweight query module for improved efficiency.

Masked Modeling: Beyond contrastive and captioning, some frameworks also use masked modeling on both text and images. For instance, SimVLM [24] employs a "PrefixLM" objective, treating images and text tokens uniformly in a language-modeling manner. Others, like Unified-IO [7], [8], rely on masking and reconstruction for vision, language, and more. This helps the model learn deeper, context-aware representations.

Data Quality and Filtering: While large public datasets such as COCO, Conceptual Captions, or LAION provide the scale needed for robust training, they often contain noise. BLIP [3] and related methods show that strategies like filtering out weak image–text pairs or re-generating improved captions can significantly boost performance. Balancing quantity and quality remains a crucial aspect of multimodal pre-training.

B. Fine and Instruction Tuning

Once a model is pre-trained, it can be fine-tuned for specific tasks. A major trend lately is instruction tuning, where the model is adapted to follow prompts, often resembling how conversational agents are fine-tuned. For instance, LLaVA [36] uses a pre-trained vision–language model and refines it with synthetic Q&A pairs—many generated by GPT-4—to better interpret visual inputs and answer open-ended questions about them. This style of tuning is akin to ChatGPT's conversational fine-tuning approach, but extended to images. InstructBLIP and other similar setups [3], [4] have shown improved performance in tasks like visual question answering and captioning with minimal additional data requirements.

C. Alignment with Human Feedback

A more advanced direction is refining models via human-in-the-loop feedback, similar to Reinforcement Learning from Human Feedback (RLHF). Although it requires substantial resources, RLHF has been successfully used in text-only systems (e.g., ChatGPT). Extending the same paradigm to multimodal models can help align outputs with human preferences or correctness. While full RLHF pipelines are still uncommon for vision–language models, they represent a logical next step toward building large multimodal systems that respond accurately and ethically [5], [30].

D. Handling Missing Modalities

A practical challenge in multimodal AI systems is the scenario where a user provides only a subset of the expected inputs. For example, a model may be designed to accept both an image and a textual prompt, but is only given text at runtime. To handle such cases, many approaches use *modality dropout* or *special tokens* to indicate missing inputs, allowing the model to learn from incomplete data during training [26], [38]. This strategy often involves randomly masking or dropping one modality on a portion of the training data, thereby encouraging the model to maintain robust performance with partial input. Our earlier discussion regarding joint embedding spaces and modality-specific embedding spaces is important here.

In some systems, where the model has separate modality-specific inputs and the image encoder output is missing (e.g., the image encoder is not invoked), the model still produces an output from the remaining modality. BLIP-2 [4], for instance, connects a frozen visual encoder to a language model via a Q-Former; when no image is available, the Q-Former receives a special embedding that effectively simulates a "no image" condition.

Models that rely on unified representations handle missing modalities naturally as the modalities are already fused together in the same space, meaning that any number of modalities can be represented/missing. During training, a subset of examples intentionally omits one or more modalities, ensuring that the model's fusion mechanism can gracefully handle partial inputs. As a result, these architectures can fall back on unimodal cues when data is missing, while still leveraging cross-modal features whenever available. This flexibility is especially useful for real-world applications such as conversational agents, where users might include or omit images at will [8], [18].

VII. Key Tasks and Applications

This section highlights some of the primary tasks and applications in which multimodal AI excels:

A. Retrieval & Matching

Methods like CLIP or ALIGN can retrieve images based on text queries and vice versa.

B. Description & Understanding

Image captioning, visual question answering, and audio-based understanding all leverage multimodal fusion to provide richer context.

C. Generation & Synthesis

Models can produce text, images, or even audio from multimodal inputs—e.g., generating detailed captions or images from textual descriptions.

D. Domain-Specific Applications

- **Healthcare**: Multimodal diagnosis combining text (doctor's notes), images (X-rays), and sensor data.
- **Robotics & Embodied AI**: Integrating camera, LiDAR, and textual instructions for tasks like navigation or manipulation.
- **Autonomous Driving**: Fusing multiple sensors for environment understanding and decision-making.
- **Creative Industries & Entertainment**: Automated multimedia production.
- **Accessibility**: Speech-to-text or text-to-audio for visually/hearing-impaired users.
- **Conversational AI**: Integrates text, voice, and even camera input for more interactive virtual assistants.

VIII. Evaluation: Datasets, Metrics, and Benchmarks

To compare multimodal LLMs across perception, reasoning, and safety, we chose **six** core dimensions recommended by the latest evaluation survey [39]. All scores are computed with **VLMEvalKit** [40], which standardizes prompt formatting, deterministic decoding and metric calculation.

1) **Avg Score** – mean of task-normalised percentages used by the OpenCompass leaderboard.
2) **MMBench_V1** – multi-choice perception & grounding QA (accuracy).
3) **MMMU_VAL** – graduate-level interdisciplinary reasoning (accuracy).
4) **MathVista** – maths & diagram comprehension (exact-match).
5) **OCRBench** – text-in-image understanding (character accuracy).
6) **HallusionBench** – robustness/safety; percentage of responses *without* hallucination (higher better).

This subset still spans the three capability layers identified in the *MME-Survey*—foundational perception (MMBench), multimodal reasoning (MMMU and MathVista), and robustness/safety (HallusionBench)—while keeping results compact.

VLMEvalKit provides loaders for all six datasets, ensuring reproducibility across open-weight (Llama 4, Qwen 2.5-Omni) and proprietary models (GPT-4o, Gemini 2.5). [40]

Instead of displaying every entry from the HuggingFace leaderboard, we select the highest-ranked model for each major lab—Google, Alibaba, OpenAI, Anthropic, XAI, DeepSeek, and Meta. Table I highlights the results of these publicly available benchmarks.

TABLE I
OpenCompass snapshot, Apr 2025 (higher = better).

Model	Rank	Avg	MMBench	MMMU	MathVista	OCR	Hallu
Gemini 2.5-Pro	1	79.6	88.3	74.7	80.9	86.2	64.1
Qwen 2.5-Omni-72B	2	78.2	87.7	72.2	79.0	90.8	59.1
GPT-4o (latest)	13	74.4	86.0	72.9	71.6	82.2	57.0
Claude 3.5 Sonnet	37	69.8	81.7	66.4	65.3	79.8	55.5
Grok-2-Vision	67	64.8	82.9	67.1	66.6	55.5	51.7
DeepSeek-VL 2	70	64.2	81.2	54.0	63.9	80.9	45.3
Llama-3.2-90B-Vis-Inst	74	63.7	77.3	60.3	58.2	78.3	44.1

Snapshot analysis.:

- **Google Gemini 2.5-Pro** leads the board with **Avg 79.6** and the top MathVista score, reflecting robust early-fusion training.
- **Alibaba Qwen 2.5-Omni-72B** follows closely, dominating OCRBench (90.8) thanks to heavy document-centric pre-training.
- **OpenAI GPT-4o** remains competitive in MMMU reasoning (72.9) and HallusionBench (57.0) despite a lower overall rank.
- **Anthropic Claude 3.5 Sonnet** is balanced but trails on mathematics, echoing known arithmetic gaps.
- **XAI Grok-2-Vision** shows strong perception (MMBench 82.9) yet weaker OCR and robustness, hinting at limited document data.
- **DeepSeek-VL 2** excels in Chinese OCR (80.9) but lags in cross-domain reasoning (MMMU 54.0).
- **Meta Llama-3.2-90B** is the top fully open-weight model but still trails proprietary systems on higher-order reasoning and safety.

All scores are reproduced directly from the OpenCompass VLM leaderboard and computed with the open-source evaluation suite `VLMEvalKit` [40]. Metric selection mirrors the capability taxonomy outlined in *MME-Survey* [39].

IX. Case Studies: Frontier & State-of-the-Art Models

This section provides brief case studies of selected frontier and state-of-the-art multimodal models, highlighting their key characteristics, modality support, and notable capabilities. Table II provides a comparative overview of input/output modalities for many of these models.

A. OpenAI o3

OpenAI's o3 represents a SOTA vision-language model focused on complex reasoning. It can handle image and text input and generate text output, although it is missing additional modalities such as video and audio, it performs extraordinarily well in tasks grounded in text and vision that require complex reasoning. Notably, its high-compute version achieved a record 87.5% on the ARC-AGI-PUB benchmark, a test requiring novel problem-solving with visual grid-based puzzles and limited prior exposure, indicating strong abstract visual reasoning capabilities. This result is remarkable because it demonstrates o3's ability to generalize and tackle unfamiliar challenges with human-like intuition, surpassing previous models that struggled with the ARC-AGI-PUB's demand for abstract reasoning, with the previous best result being 56%.

B. Anthropic Claude 3.7 Sonnet

Claude 3.7 Sonnet is an upgrade to the previous 3.5 version, being Anthropic's first hybrid reasoning model. It can understand text and various visual inputs such as images, charts and diagrams. The model is only capable of generating text; however, it includes a feature that allows the model to take advantage of its visual abilities to create diagrams using code. Additionally, the model has a 200k-token context window, with the underlying capability to extend this to 1M-tokens, although this version is not widely available. It also features computer-use, where the model can interact with a desktop environment, by feeding desktop screenshots in addition to text prompts, the model is able to follow instructions and interact via a virtual mouse and keyboard. A highly modified version of this feature has been developed, that allows the model to play the Gameboy game Pokémon Red, by allowing the model analyze game images, execute actions (via virtual button presses) and store long-term information in a knowledge base. These features, in addition to its reasoning capabilities, has allowed Claude to defeat three of the eight gym leaders in the game. This is an interesting new benchmark for MM-LLM models because it tests the models abilities to understand visual information, make game decisions and handle long context over an extended time. [28], [41].

C. DeepSeek AI DeepSeek-VL 2

DeepSeek-VL2 is an open-source Mixture-of-Experts (MoE) vision-language model. It accepts vision and text inputs, generating text outputs. It uses a SigLIP-based vision encoder with a dynamic tiling to process high-resolution images, a vision-language adaptor, and an MoE-based language model with Multi-head Latent Attention (MLA). This is a hybrid fusion approach, integrating visual and language embeddings through both early and late fusion [29].

D. OpenAI GPT-4o ("omni")

GPT-4o, developed by OpenAI, can accept any combination of text, image, video and audio, then generate any combination of text, image and audio outputs. It achieved real time speech-to-speech interaction with latency as low as 232ms and can handle multi-turn, cross-modal conversations. It uses a single end-to-end autoregressive transformer, trained across all modalities. It employs early fusion by processing all modality inputs as a unified token sequence and uses a shared embedding space for cross-modal attention and reasoning [9].

E. Google Gemini 2.5

Gemini 2.5 is the latest model from Google DeepMind's Gemini family of multimodal models "from the ground up". The Gemini 2.5 version is capable of understanding text, image, video and audio, however it can only generate text. It builds upon the previous models by introducing a 1 million token context window as well as an explicit "think-then-respond" decoding strategy for chain-of-thought (CoT) reasoning capabilities. The model architecture specifics are not publicly available; however, a public statement suggests that the model uses a new Mixture-of-Experts (MoE) architecture, although this has not been directly confirmed by Google. Additionally, Gemini has recently taken on a similar challenge to Claude, with a playthrough of Pokemon Blue, currently surpassing Claude attaining all eight of the gym badges in the game as of May 25th, 2025, only needing to defeat the final four, the last challenge in the game. This is a remarkable achievement, demonstrating extremely strong reasoning abilites especially over time with long context. To reach this game state the model had to make over 85,000 individual actions. [12], [42], [43].

F. Alibaba Qwen 2.5-Omni

Qwen 2.5-Omni is an open-source MM-LLM released by Alibaba Cloud, that can understand text, images, video and audio with the ability to generate both text and audio. It uses a dual-module Thinker-Talker architecture where the thinker handles multimodal understanding and test generation, while the talker generates speech. It utilizes a novel time-aligned Multimodal RoPE (TMRoPE) which allows for precise temporal alignment between audio and vision inputs. The model uses a hybrid fusion strategy and modality specific encoders with shared attention in the thinker to integrate all inputs into a unified embedding space [13].

G. DeepMind RT-2 (Robotics Transformer 2)

RT-2 is a vision-language-action (VLA) model that extends PaLM-E by enabling direct robot control through token outputs that directly correspond to action outputs. While PaLM-E generates text-based plans that are interpreted by separate controllers, RT-2 modifies the model architecture and training objective to produce low-level robot actions directly. It uses the

same early fusion transformer architecture with ViT and MLP adapters. To allow for end-to-end control, RT-2 discretizes robot actions, such as 6-DoF end-effector movements, into bins that are assigned a unique token within the model's vocabulary. During inference, the model generates a sequence of these tokens which are decoded into control commands that directly move the robot to the correct position. The model was co-fine-tuned on language-vision data sets and robot trajectories, allowing it to generalize to unseen objects, tasks and environments. In practice the model exhibits emergent capabilities such as symbolic reasoning, spatial inference, and visual grounding [44].

H. Figure AI Helix

Helix is a vision-language action model developed by Figure AI to enable humanoid robots to perform complex tasks using integrated perception, language understanding and precise motor control. Additionally, when two Figure robots are within close proximity, they are capable of collaborating on tasks. The model uses a dual-system transformer architecture, consisting of System 2 (S2), a pretrained 7B-parameter Vision Language Model (VLM) that operates at 7-9Hz and System 1 (S1) an 80M-parameter cross-attention encoder-decoder transformer operating at 200Hz. S2 operates at a slower frequency and gives the system high-level visual understanding of the environment and natural language, while S1 translates these into continuous actions for full control of the entire upper body (wrist, finger, torso, and head) of the robot with a much faster frequency. This design decision allows the S2 system to 'think slow' and consider high level goals within environmental context while S1 can act fast and execute actions in real time. The model was trained end-to-end on a dataset derived from around 500 hours of high-quality, multi-robot, teleoperated demonstrations, processed by an auto-labeling vision-language model to generate natural language-conditioned training pairs from segmented video clips. This model architecture and training has allowed Helix to achieve zero-shot generalization in novel and complex tasks. One example of this is a demonstration where two figure robots were placed within a kitchen and given common grocery items such as an apple, eggs trash bags and ketchup, with the instruction of working together to put the items away. The robots then collaborated to place the items within the correct locations, apple on the counter, trash bags in the cabinet, eggs and ketchup in the fridge [32].

I. Meta Llama 4

Llama 4 is Meta's first *natively multimodal* Llama generation, described publicly in an April 2025 blog post [45]. Unlike Llama 3's detachable vision head, Llama 4 embeds text, images, video frames, and audio spectrogram tokens into a single transformer sequence, allowing early-fusion cross-modal attention at every layer. Meta reports three variants—*Scout* (17 B MoE), *Maverick* (57 B), and *Behemoth* (310 B)—all pre-trained from scratch on trillions of mixed-modality tokens. Early-fusion design yields stronger joint reasoning but entails higher pre-training cost compared with the late-fusion Llama 3 vision add-on. Also a notable point about this model is that this along with all of the other LLAMA series models are fully open-weights and are publicly available.

J. XAI Grok 3

Grok 3, unveiled by xAI in mid-February 2025, is the first Grok model to pair image understanding with its established text pipeline while still producing text only; it features a retrieval-augmented transformer with a built-in "Think" drafting loop, a reported 1M-token context window, and roughly ten times the training compute of Grok 2. Early benchmarks cited by xAI and third-party outlets place it ahead of GPT-4o on MMLU-Pro and GPQA, and the model has already been wired into the X mobile app for real-time camera Q&A, putting it in direct competition with Gemini 2.5 and GPT-4o vision. Independent coverage also notes lighter guard-rails than rival systems, prompting xAI to enlarge its red-team effort after high-profile NSFW and bias incidents. [46]

K. Model Capability Comparison Summary

In table II on the next page, we also summarize these case studies to show the modality capabilities of each model. We show the input and outputs for text, image, video, audio, code. We also show the input for sensors and the outputs for actions for robotic systems for easier comparison.

X. Limitations and Open Challenges

Despite significant progress, multimodal AI still faces many open challenges:

- **Data Imbalance & Missing Modalities:** Training data may be abundant for text and images but sparse for audio or specialized sensor data.
- **Hallucinations & Factual Inconsistency:** Generative models can produce incorrect or nonsensical outputs, especially when modalities misalign.
- **Robustness:** Real-world noise, occlusion in images, or background sounds can degrade performance.
- **Interpretability & Explainability:** Transformer-based architectures remain largely black boxes, complicating trust in safety-critical applications.
- **Computational Cost:** Handling multiple modalities in large-scale transformers is GPU-intensive and environmentally costly.
- **Compositionality & Reasoning:** Systems often struggle with tasks requiring complex reasoning over multiple modalities, especially if the training data has not demonstrated such compositional tasks.

XI. Safety, Ethics, and Societal Impact

Multimodal AI, like all AI technologies, can have far-reaching societal impacts:

- **Bias and Fairness:** Models can encode and amplify biases present in training data, potentially across multiple modalities.

TABLE II
INPUT/OUTPUT MODALITY SUPPORT FOR RECENT FRONTIER MM-LLMS

Model	Text		Image		Video		Audio		Code		Sensor	Action
	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out
OpenAI **o3**	Y	Y	Y	Y	N	N	N	N	Y	Y	N	N
Anthropic **Claude 3.7 Sonnet**	Y	Y	Y	N	N	N	N	N	Y	Y	N	N
DeepSeek AI **DeepSeek-VL 2**	Y	Y	Y	N	N	N	N	N	Y	Y	N	N
OpenAI **GPT-4o ("omni")**	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	N	N
Google **Gemini 2.5**	Y	Y	Y	N	Y	N	Y	N	Y	Y	N	N
Alibaba **Qwen 2.5-Omni**	Y	Y	Y	N	Y	N	Y	Y	Y	Y	N	N
DeepMind **RT-2**	Y	N	Y	N	L	N	N	N	N	N	Y	Y
Figure AI **Helix**	Y	N	Y	N	Y	N	N	N	N	N	Y	Y
Meta **Llama 4**	Y	Y	Y	Y	?	?	?	?	Y	Y	N	N
xAI **Grok 3**	Y	Y	Y	N	N	N	N	N	Y	Y	N	N

Legend: Y = supported; L = limited/pipeline; N = not supported; ? = not yet disclosed. **Sensor In** refers to proprioceptive or other robot-state inputs; **Action Out** denotes direct robot-control tokens or continuous control vectors.

- **Misinformation & Malicious Use:** Deepfakes become more sophisticated when AI can generate both fake audio and visuals.
- **Robustness & Safety Critical Applications:** Failures in medical or autonomous systems can have life-threatening consequences.
- **Transparency & Accountability:** As these systems become black boxes, attributing responsibility for errors is challenging.
- **Accessibility:** Ensuring accessible outputs (e.g., text descriptions for images) is crucial for inclusive technologies.
- **Alignment:** Ensuring the model's objectives are consistent with human values and minimizing harmful or deceptive outputs.

XII. FUTURE DIRECTIONS

Multimodal AI research is rapidly evolving, and several directions are emerging:

- **Towards "Any-to-Any" Models:** Unified frameworks like Unified-IO and Qwen-Omni that can handle any input-output modality pair.
- **Scaling Laws & Emergent Abilities:** Investigating how larger models exhibit new capabilities and emergent behaviors.
- **Enhanced Reasoning & Grounding:** Addressing compositional reasoning and grounding textual outputs in real-world sensory input.
- **Efficiency and Optimization:** Reducing the massive computational and memory overhead of large multimodal transformers.
- **Openness & Reproducibility:** Encouraging open-source models and transparent benchmarks to foster collaborative progress.
- **Embodied Multimodal AI:** Integrating robotics and sensor data to enable continuous learning in real-world environments.
- **Personalization & Continual Learning:** Adapting models to user-specific data over time without catastrophic forgetting.
- **Advancements in Specific Modalities:** Audio, video, and sensor-based modeling improvements.
- **Proactive Ethics & Governance:** Anticipating ethical pitfalls and embedding safeguards in model training and deployment.
- **Multimodal Safety:** Research on multimodal jailbreak strategies can help protect these models against such attacks.

XIII. CONCLUSION

Multimodal AI systems, particularly those driven by large language models, have shown remarkable progress in recent years. By fusing data from multiple domains—text, images, audio, video, and beyond—these systems achieve more robust, context-aware reasoning and generation. Nonetheless, important limitations remain, including data imbalance, hallucination, interpretability, and the high computational costs associated with large-scale training.

As research continues, the field is likely to trend toward unified "any-to-any" architectures, more transparent training protocols, and deeper integration with real-world environments (e.g., robotics, autonomous vehicles, healthcare). Ensuring safety, fairness, and alignment will be critical to harnessing the full potential of multimodal AI for societal benefit.

REFERENCES

[1] Google, "Unlock a whole new level of google home," 2024, accessed: 2025-04-18. [Online]. Available: https://home.google.com/get-inspired/unlock-a-whole-new-level-of-google-home/

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[3] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*. PMLR, 2022, pp. 12 888–12 900. [Online]. Available: https://proceedings.mlr.press/v162/li22n.html

[4] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proceedings of the 40th International Conference on Machine Learning*. PMLR, 2023, pp. 19 730–19 742. [Online]. Available: https://proceedings.mlr.press/v202/li23q.html

[5] S. Huang, L. Dong, W. Wang, Y. Hao, S. Singhal, S. Ma, T. Lv, L. Cui, O. K. Mohammed, B. Patra, Q. Liu, K. Aggarwal, Z. Chi, J. Bjorck, V. Chaudhary, S. Som, X. Song, and F. Wei, "Language is not all you need: Aligning perception with language models," *arXiv preprint arXiv:2302.14045*, 2023.

[6] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, "Flamingo: a visual language model for few-shot learning," in *Advances in Neural Information Processing Systems*, 2022.

[7] J. Lu, C. Clark, R. Zellers, R. Mottaghi, and A. Kembhavi, "Unified-io: A unified model for vision, language, and multi-modal tasks," in *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023. [Online]. Available: https://openreview.net/forum?id=E01k9048soZ

[8] J. Lu, C. Clark, S. Lee, Z. Zhang, S. Khosla, R. Marten, D. Hoiem, and A. Kembhavi, "Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024, pp. 26 439–26 455.

[9] OpenAI, "GPT-4o System Card," *arXiv preprint arXiv:2410.21276*, Oct. 2024. [Online]. Available: https://arxiv.org/abs/2410.21276

[10] S. Pichai, "Introducing gemini: our largest and most capable ai model," https://blog.google/technology/ai/google-gemini-ai/, 2023, accessed: 2025-03-30.

[11] OpenAI, "OpenAI o3 and o4-mini System Card," Apr. 2025, accessed: 2025-04-18. [Online]. Available: https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf

[12] K. Kavukcuoglu, "Gemini 2.5: Our most intelligent ai model," https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/, March 2025, accessed: 2025-03-30.

[13] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang, B. Zhang, X. Wang, Y. Chu, and J. Lin, "Qwen2.5-omni technical report," 2025, available at https://arxiv.org/abs/2503.20215.

[14] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," *arXiv preprint arXiv:2306.13549*, 2023.

[15] D. Huang, C. Yan, Q. Li, and X. Peng, "From large language models to large multimodal models: A literature review," *Applied Sciences*, vol. 14, no. 12, p. 5068, 2024. [Online]. Available: https://www.mdpi.com/2076-3417/14/12/5068

[16] D. Caffagni, F. Cocchi, L. Barsellotti, N. Moratelli, S. Sarto, L. Baraldi, M. Cornia, and R. Cucchiara, "The revolution of multimodal large language models: A survey," *arXiv preprint arXiv:2402.12451*, 2024.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[18] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5583–5594.

[19] M. Pawłowski, A. Wróblewska, and S. Sysko-Romańczuk, "Effective techniques for multimodal data fusion: A comparative analysis," *Sensors*, vol. 23, no. 5, p. 2381, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/5/2381

[20] F. Zhao, C. Zhang, and B. Geng, "Deep multimodal data fusion," *ACM Computing Surveys*, vol. 56, no. 9, pp. 1–36, Apr. 2024. [Online]. Available: https://dl.acm.org/doi/10.1145/3649447

[21] J. Li, R. R. Selvaraju, A. D. Gotmare, S. Joty, C. Xiong, and S. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *Advances in Neural Information Processing Systems*, vol. 34, 2021, pp. 9694–9705.

[22] OpenAI, "Chatgpt: Optimizing language models for dialogue," 2022, https://openai.com/blog/chatgpt.

[23] H. Tan and M. Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019, pp. 5099–5110.

[24] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "Simvlm: Simple visual language model pretraining with weak supervision," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=GUrhfTuf_3

[25] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021, pp. 4904–4916. [Online]. Available: https://proceedings.mlr.press/v139/jia21b.html

[26] J. Wang, H. Jiang, Y. Liu, C. Ma, X. Zhang, Y. Pan, M. Liu, P. Gu, S. Xia, W. Li, Y. Zhang, Z. Wu, Z. Liu, T. Zhong, B. Ge, T. Zhang, N. Qiang, X. Hu, X. Jiang, X. Zhang, W. Zhang, D. Shen, T. Liu, and S. Zhang, "A comprehensive review of multimodal large language models: Performance and challenges across different tasks," *arXiv preprint arXiv:2408.01319*, 2024. [Online]. Available: https://arxiv.org/abs/2408.01319

[27] X. Chen, X. Wang, S. Changpinyo, A. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer, A. Kolesnikov, J. Puigcerver, N. Ding, K. Rong, H. Akbari, G. Mishra, L. Xue, A. Thapliyal, J. Bradbury, W. Kuo, M. Seyedhosseini, C. Jia, B. K. Ayan, C. Riquelme, A. Steiner, A. Angelova, X. Zhai, N. Houlsby, and R. Soricut, "Pali: A jointly-scaled multilingual language-image model," in *Proceedings of the 11th International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=mWVoBz4W0u

[28] Anthropic, "Claude 3.7 sonnet system card," https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-3-7-sonnet-system-card.pdf, 2025, accessed: 2025-04-15.

[29] Z. Wu, X. Chen, Z. Pan, X. Liu, W. Liu, D. Dai, H. Gao, Y. Ma, C. Wu, B. Wang, Z. Xie, Y. Wu, K. Hu, J. Wang, Y. Sun, Y. Li, Y. Piao, K. Guan, A. Liu, X. Xie, Y. You, K. Dong, X. Yu, H. Zhang, L. Zhao, Y. Wang, and C. Ruan, "Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding," *arXiv preprint arXiv:2412.10302*, 2024. [Online]. Available: https://arxiv.org/abs/2412.10302

[30] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua, "NExT-GPT: Any-to-any multimodal LLM," *arXiv preprint arXiv:2309.05519*, 2023.

[31] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu *et al.*, "Palm-e: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.

[32] Figure AI, "Introducing Helix: A Generalist Vision-Language-Action Model for Humanoid Robots," https://www.figure.ai/news/helix, 2025, accessed: 2025-04-23.

[33] H. T. Songtao Li, "Multimodal alignment and fusion: A survey," *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 2024.

[34] J. Huang and J. Zhang, "A survey on evaluation of multimodal large language models," *arXiv preprint arXiv:2408.15769*, 2024. [Online]. Available: https://arxiv.org/abs/2408.15769

[35] OpenAI, "GPT-4V(ision) System Card," https://cdn.openai.com/papers/GPTV$_{}system_{}Card.pdf$, 2023, $accessed: 2025-04-25$.

[36] H. Liu, C. Zhang, Y. Xu, Z. Zhang, X. Zhang, X. Hu, J. Wang, and J. Yang, "Visual instruction tuning," *arXiv preprint arXiv:2304.08485*, 2023. [Online]. Available: https://arxiv.org/abs/2304.08485

[37] L. Li, G. Chen, H. Shi, J. Xiao, and L. Chen, "A survey on multimodal benchmarks: In the era of large ai models," *arXiv preprint arXiv:2409.18142*, 2024. [Online]. Available: https://arxiv.org/abs/2409.18142

[38] Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma, and F. Wei, "Kosmos-2: Grounding multimodal large language models to the world," *arXiv preprint arXiv:2306.14824*, 2023. [Online]. Available: https://arxiv.org/abs/2306.14824

[39] C. Fu, Y.-F. Zhang, S. Yin, B. Li, X. Fang, S. Zhao, H. Duan, X. Sun, Z. Liu, L. Wang, C. Shan, and R. He, "MMESurvey: A comprehensive survey on evaluation of multimodal llms," *arXiv preprint arXiv:2411.15296*, 2024, version 2, 8 Dec 2024. [Online]. Available: https://arxiv.org/abs/2411.15296

[40] H. Duan, J. Yang, Y. Qiao, X. Fang, L. Chen, Y. Liu, X. Dong, Y. Zang, P. Zhang, J. Wang *et al.*, "Vlmevalkit: An open-source toolkit for evaluating large multi-modality models," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 11 198–11 201.

[41] L. Space, "How claude 3.7 plays pokémon," Mar. 2025, accessed: 2025-04-25. [Online]. Available: https://www.latent.space/p/how-claude-plays-pokemon-was-made

[42] N. Malaviarachchi, "Gemini 2.5: A leap forward in ai reasoning and contextual understanding," https://www.linkedin.com/pulse/gemini-25-leap-forward-ai-reasoning-contextual-malaviarachchi-of1lc/, 2025, accessed: 2025-04-22.

[43] waylaidwanderer, "Gemini plays pokémon," https://www.twitch.tv/gemini$_p lays_p okemon$, 2025, $twitch livestream featuring Gemini Pro 2.5 playing Pokémon Blue$.

[44] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, Q. Vuong, V. Vanhoucke, H. Tran, R. Soricut, A. Singh, J. Singh, P. Sermanet, P. R. Sanketi, G. Salazar, M. S. Ryoo, K. Reymann, K. Rao, K. Pertsch, I. Mordatch, H. Michalewski, Y. Lu, S. Levine, L. Lee, T.-W. E. Lee, I. Leal, Y. Kuang, D. Kalashnikov, R. Julian, N. J. Joshi, A. Irpan, B. Ichter, J. Hsu, A. Herzog, K. Hausman, K. Gopalakrishnan, C. Fu, P. Florence, C. Finn, K. A. Dubey, D. Driess, T. Ding, K. M. Choromanski, X. Chen, Y. Chebotar, J. Carbajal, N. Brown, A. Brohan, M. Gonzalez Arenas, and K. Han, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Proceedings of The 7th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 229. PMLR, 2023, pp. 2165–2183. [Online]. Available: https://proceedings.mlr.press/v229/zitkovich23a.html

[45] Meta AI, "Llama 4: Multimodal Intelligence," https://ai.meta.com/blog/llama-4-multimodal-intelligence/, 2025, accessed: 2025-04-19.

[46] xAI. (2025, Feb.) Grok 3 beta — the age of reasoning agents. Accessed 25 Apr 2025. [Online]. Available: https://x.ai/blog/grok-3