International Journal of Computer Applications (0975 - 8887)
Volume 187 - No.19, June 2025

Multimodal Deep Learning: A Survey of Models, Fusion Strategies, Applications, and Research Challenges

Sai Teja Erukude
Department of Computer Science, Kansas State University
Manhattan, 66506, KS, USA

Suhasnadh Reddy Veluru
College of Business Administration, Kansas State University
Manhattan, 66506, KS, USA

Viswa Chaitanya Marella
College of Business Administration, Kansas State University
Manhattan, 66506, KS, USA

ABSTRACT

Multimodal deep learning has become a primary methodological framework in artificial intelligence, allowing models to learn from (and reason over) many different types of data, such as text, images, audio, and video. By utilizing multiple modalities simultaneously, systems can enhance their contextual understanding, noise resilience, and generalization, all of which closely resemble human perception. This review offers a comprehensive overview of the field, taking a look at the basics of modality integration, fusion methods (early, late, and hybrid), and some of the main architectural advances in models like CLIP, Flamingo, GPT-4V, Gemini 1.5, and AudioCLIP. It also provides a primer on real-world applications in healthcare, autonomous systems, robotics, and education, including benchmarking datasets and evaluation metrics essential for evaluating performance. Notable challenges, such as modality imbalance, scalability, and interoperability, are highlighted, while also looking at growing areas of interest such as long-context modeling and embodied intelligence. As a review survey, the goal is to provide a map of options for researchers and practitioners who want to enhance their use of multimodal AI systems, both in research and in actual deployment.

Keywords

Cross-Modal Learning, Fusion Strategies, Vision-Language Reasoning, Multimodal Architectures, Foundation Models

1. INTRODUCTION

Multimodal deep learning refers to machine learning methods that can process and learn from multiple modalities at once (e.g., text, images, audio, video) to improve prediction and understanding. The motivation for multimodal models comes from the human brain, which naturally responds to multiple types of sensory inputs and generates a more complete picture of the world. Multimodal systems are rapidly advancing due to the growth of large-scale, multimodal data collection, model architectures, and self-supervised learning, and have become prominent in areas like vision-language, audio and video processing, and image and video classification. There are now realizable applications across sectors ranging from healthcare, robotics, education, entertainment, and more.

Multimodal learning is gaining traction as a core asset to building general-purpose intelligent systems. Unlike its unimodal counterparts that consume one data type and subsequently reduce many contextual cues, multimodal systems can contextualize multiple aspects of the environment. Multimodal systems inherit additional resiliency to noise from moment-to-moment environmental variability that single-modality models cannot. Further, multimodal systems can help perform reasoning tasks that require analyzing data from multiple sources at once to guide predictions. For example, consider how autonomous vehicles are designed to process visual and lidar data, GPS data, and auditory data simultaneously; this is vital for safe and accurate navigation. Similarly, a conversational AI assistant can take advantage of multimodality by audio input and facial expression to better infer user intent and meaning.

This research aims to provide both researcher and practitioner communities with a comprehensive review of multimodal deep learning and its evolution. The following sections in the review provide a summary of introductory concepts (section 2), architectures and major models (Section 3), applications (Section 4), datasets and evaluation metrics (Section 5), major challenges and thoughts on future research directions (Section 6).

2. CORE MULTIMODAL LEARNING CONCEPTS

2.1 Modalities and Heterogeneity

A modality is simply a separate source of data, commonly referred to as images, speech, text, and signals from senses, in the field of deep learning. Each of these data types also has different properties; e.g., text is sequential and symbolic, images are dense and spatial, and audio is continuous and temporal. This diversity in modality properties creates challenges for integrating one type of data with another, especially when the data has varying dimensions, rates of sampling, and structures [3].

Typically, modality-specific encoders are used to transform the raw input to a latent representation. Manufacturers typically provide encoders, such as convolutional neural networks (CNNs) for im-

age data, recurrent neural networks (RNNs) or Transformers (for text), and spectrogram-based CNNs or temporal convolution networks (for audio). After encoding the modalities with one or more modality-specific encoders, the representations generated by the encoders are then aligned and fused for joint reasoning.

2.2 Fusion strategies

A core aspect of multimodal learning is fusion, or the joining of representations obtained from several different modalities. There are broadly three strategies, or levels of fusion:

(1) **Early Fusion:** Early fusion combines or merges raw or lower-level features from multiple modalities and inputs to jointly model the different modalities. This allows cross-modal interaction at the early stages of processing, but is often faced with incompatible feature dimensions and noisy inputs.

(2) **Late Fusion:** In late fusion, each modality gets processed independently, with the outputs fused at the decision level, typically through averaging, voting, or taking a weighted sum as an output value. Independent processing of each modality allows for robustness against missing modalities; however, it reduces potential interaction between modalities during learning.

(3) **Hybrid Fusion:** Hybrid fusion combines data at different fusion levels (early, mid, and late), establishing a balance between representational richness and robustness. More recent methods focus on using attention models and Transformers for dynamic fusion, which allow any modality to arbitrarily attend to relevant signals of other modalities [23].

2.3 Learning Challenges

Multimodal deep learning has introduced unique challenges in learning as listed below:

(1) **Alignment:** Meaningful associations between modalities require alignment, whether it be temporal or spatial. For instance, one could conceivably align oral words to mapped facial movements in audiovisual speech recognition.

(2) **Missing/noised modalities:** A real-world instance is one where data is missing (e.g., missing audio) or the input is perturbed. Models need to be conditioned on such scenarios effectively. Modal dropout, imputation, and conditional gating are some of the techniques used for model conditioning.

(3) **Representation Collapse and Dominance:** When a single modality is allowed to dominate the learning process (e.g., text in image-text datasets), the other modalities could be neglected in their potential. Strategies to mitigate this include regularization strategies, co-learning objectives, and attention normalization.

(4) **Cross-modal Transfer and Generalization:** A central purpose of foundation models and contrastive learning approaches [18] relates to the ability for one modality to inform another (e.g., learning visual grounding supervision on text only).

3. MAJOR ARCHITECTURES AND MODELS

Most modern multimodal AI architectures tend to be modular in design, where each modality of data (e.g., text, images, or audio) has its own encoder. Each encoder processes modality-specific features, which are then fused at the fusion layer, which encodes cross-modal relationships, thus producing the fused embedding for a richer and more contextualized representation. This is then decoded by a shared decoder to perform various tasks (e.g., captioning, retrieval, or question answering). This generalized pipeline, represented in Figure 1, serves as the basis for many recently developed models, including CLIP, Flamingo, and Gemini 1.5.

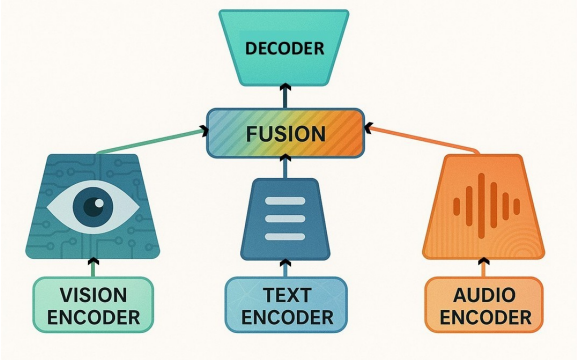

Fig. 1. Illustration of a typical multimodal deep learning architecture

3.1 CLIP (Contrastive Language-Image Pretraining)

CLIP was released by OpenAI in 2021 in the form of a dual-encoder model that is trained on 400 million image-text pairs [18]. It represents images and text into a common embedding space using a contrastive distributional loss function; the goal is to ensure that both matched image-text pairs have a higher joint probability than unmatched linkages. The model consists of (i) a Vision Transformer (or ResNet) to handle images and (ii) a Transformer to handle text. STRAP has demonstrated exceptional zero-shot performance on image classification and retrieval tasks without explicit fine-tuning.
Its training objectives scale well with noisy, internet-scale data and allow for generalizations across tasks such as object recognition, OCR, and visual entailment. The ease of use and efficiency of CLIP empowered an influx of multimodal research, while also providing an architecture backbone in many downstream pipelines such as DALL·E and Flamingo. Despite this exciting leap in performance, researchers are already exploring more general models, since CLIP's dual-encoder architecture introduces rigidities that limit fine-grained alignment between modalities.

3.2 DALL-E 2

DALL-E 2 is a generative model that can create high-resolution, realistic images from text descriptions [19]. The system first passes the text prompt to a prior model, which outputs a CLIP image embedding. The image embedding is then passed into a diffusion decoder, which builds a realistic image. The two-phase pipeline structure enables semantic adequacy while providing realistic image quality. The improvements in photorealism and text-image consistency are substantial compared to the first DALL-E. Also, DALL-E 2 is capable of inpainting (editing regions) and generating variations (different flexible outputs from a prompt). The model shows a flexible way of exhibiting visual creativity. The use of diffusion models is a notable departure from autoregressive pixel-level models, as DALL-E 2 generates images in latent space, which allows for better scalability. Ethical issues surrounding the misuse of large models and hallucinated content have spurred the continued development of guardrails and content filters in DALL-E 2 and other large language models.

3.3 Flamingo

Flamingo [2] is a multimodal model created by DeepMind that builds off language models (LLMs) with visual input through gated cross-attention layers. Flamingo is a few-shot learning model that uses the vision encoder in a frozen state together with a pre-trained LLM, allowing the prompt to comprise alternating sequences of images and text. Gated cross-attention layers allow for hierarchical multimodal learning, where there is no full-tuning to accomplish few-shot learning.
An important innovation of Flamingo is their efficient learning paradigm, where they learn a small number of multimodal-specific parameters while keeping the vision backbone and language backbone frozen. The separate parameters are the learning focus that makes it quicker and cheaper to learn new tasks. Flamingo demonstrates state-of-the-art performance on tasks such as the VQA and Science QA benchmarks, demonstrating reasoning across modalities as a strong model. Flamingo has inspired successors such as Gemini, as well as new extensions to extend the competencies, well-being of the performance discovered in Flamingo.

3.4 GPT-4V

GPT-4V [1] is a multimodal extension of GPT-4 with vision capabilities, with image and text ability. In that domain of capabilities, GPT-4V gives rise to a new set of image and text processing tasks and forms of visual question answering, ie, multimodal reasoning, chart interpretation, etc. There are very little amounts of architectural details of the model, but they utilize a visual encoder that is embedded as part of the Transformer through token unification through learned embeddings.
In the proposed architecture of GPT-4V, they treat the image inputs as sequences of embedding tokens—so it's reasonable to then process the inputs as embedding tokens as text inputs either together or across a shared set of Transformer layers. This is viable because it circumvents the need for a separate modality, it allows for easy collaboration, and it more clearly preserves cross-modality attention between modalities across all layers of the network. GPT-4V is especially strong at few-shot visual tasks, and has strong capabilities for real-world examples of chart and web interface interpretation, diagrams, and handwritten text interpretation. In general, this work will serve to demonstrate the feasibility of unified architectures longer term and in a broader sense, generalist AI.

3.5 Gemini 1.5

Gemini 1.5 by Google DeepMind is unprecedented for long-context multimodal inputs (text, images, and audio) [22]. Gemini has a context window in millions of tokens and can handle cross-document and long-sequence tasks (e.g., multi-document QA, instructional video comprehension, and audio-visual analysis). The design builds upon Flamingo and PaLM-E and utilizes special adapters with sparse attention to handle larger input sequences. Gemini is promising for reasoning about multimodal data at scale, such as reading dense scientific papers, trying to follow a best workflow, and fine-grained entity identification across diagrams/tables/narration. Its training was focused on modularity and temporal coherence, which gives it a better grounding temporally when dealing with time-based data like video or speech.
3.6 AudioCLIP

3.6 AudioCLIP

AudioCLIP brings CLIP (Contrastive Language/Audio Pretraining) into the audio domain by applying a third encoder to learn to align audio embeddings to the shared vision-language space [12]. For example, tasks like audio-caption retrieval or classification of sounds using textual supervision. AudioCLIP shows one way transfer learning and shared latent spaces can enable models to generalize from one modality (ex., text) to another (ex., audio) and not require direct supervision. AudioCLIP comprises a pretrained audio encoder (i.e., VGGish or PANNs) and is aligned to the embeddings of images and text embeddings using a contrastive loss. The model has been applied successfully to musical genre classification, environmental sound tagging, and multilingual sound retrieval, to name a few. One of the contributions of AudioCLIP is showing one of the many low-barrier paths to extend existing multimodal models to novel sensory domains.

4. APPLICATIONS AND CASE STUDIES

4.1 Healthcare

Multimodal systems within healthcare integrate types of data (e.g., medical imaging (e.g., X-rays, MRIs), clinical notes, lab tests, genomics, patient history) to improve diagnoses and treatment decisions. For example, multimodal models that have been trained on the MIMIC-CXR dataset [14] can produce radiology reports from chest X-rays or classify findings like lung opacity or cardiomegaly. These systems improve diagnostic accuracy by contextualizing visual abnormalities with textual markers to make diagnostic decisions, which helps decrease false positive rates and also helps with participatory practice, especially important as an interpretation for radiologists. Alternatively, low-functioning models would help severe patient resource triaging and prioritization so clinicians may better respond to the most critical cases promptly. Recent efforts are also exploring the incorporation of patient data from wearable devices (e.g., heart rate, oxygen levels, etc.) along with the clinician-patient interaction from the visit and developing more complete models as needed. Federated learning methods, particularly those that are privacy-preserving, are also trending due to the sensitive nature of health data.

4.2 Autonomous Vehicles

Modalities are critical to autonomous driving, which relies on multimodal sensor fusion capabilities that utilize multiple sensor modalities such as vision from a camera, spatial depth from a lidar, motion from a radar, and multi-position based localization [5]. Models trained on the nuScenes datasets, for example, can utilize these modalities to achieve tasks such as 3D object detection, lane tracking, and semantic segmentation of the scene.
Bi-modal fusion is important for robustness under real-world environments where adverse weather or lighting conditions may degrade the usability of one of the modalities (e.g., vision). Fusion approaches may focus on the early stages of the fusion (i.e., early, late, and hybrid), and although innovation to the network is performed without post-hoc re-learning, modular designs are more sustainable for updates and facilitate re-usage. Recent efforts also use audio to detect emergency sirens and driver intention modeling that utilizes internal cabin cameras and voice commands. The use of multimodal reasoning in real-time allows for the robot's ability to understand a scene, reason about what to do, and generate an action plan. This helps a vehicle or robot safely and efficiently navigate a dynamic urban space.

Table 1. Summary of Key Multimodal Models.

Model	Modalities	Architecture	Highlights
CLIP [18]	Image + Text	Dual encoder, contrastive	Zero-shot classification and retrieval
DALL-E 2 [19]	Text → Image	CLIP prior + diffusion	High-quality generative synthesis
Flamingo [2]	Image + Text	LLM + vision cross-attn	Few-shot multimodal reasoning
GPT-4V [1]	Image + Text	Unified Transformer	General visual understanding
Gemini 1.5 [22]	Text + Vision + Audio	Long-context Transformer	Multimodal document and video QA
AudioCLIP [12]	Audio + Image + Text	Triple encoder	Cross-modal audio retrieval/classification

4.3 Robotics and Embodied AI

In robotics, multimodal models allow a machine to observe, reason, and act in real-world, dynamic environments. Agents like PaLM-E [7] use language commands, RGB-D vision, proprioceptive feedback, and maps of the environment to achieve tasks such as object retrieval or using a tool.

These agents have been able to ground language to actions and vice versa - for example, taking the meaning of "put the mug on the left shelf" through the combination of scene layout, perception of the object, and planning an arm trajectory. By using cross-modal Transformers, a person can control multiple robotic platforms using a minimum of fine-tuning. The advantage of having real-time feedback from different modalities provides the opportunity for adaptive learning that incorporates trial-and-error interactive experiences. For example, current designs for future robots will allow continuous memory and long-context reasoning to occur so the robot can robustly operate in household, industrial, and health care spaces with little human intervention.

4.4 Education and HCI

The use of multimodal AI systems in education uses various inputs such as speech, gaze, facial expressions, pen strokes, and typing behaviors to create personalized learning experiences. For example, intelligent tutoring systems can assess students' confusion depending on variations in the pitch of their voice and eye movement to dynamically adjust the level of difficulty for the presented content [6].

Visual dialog agents and virtual assistants can use facial tracking and recognition of gestures to recognize confusion and clarify ambiguities in user requests. In accessibility-oriented HCI, multimodal systems can integrate speech-to-text transcription, sign language interpretation, and customized interfaces for users with motor disabilities. Such applications not only add value for usability and inclusiveness but also help advance human-machine collaboration. As mixed-reality gains traction and AR/VR platforms, including spatial computing, emerge, the utilization of multimodal AI will only become more central to immersive educational simulation and remote collaborative environments.

5. DATASETS AND EVALUATION METRICS

5.1 Benchmark Datasets

Multimodal datasets are fundamentally important for training, evaluating, and benchmarking AI systems that learn across multiple modalities. The key multimodal datasets that are used in the field are as follows:

Many of these datasets have millions of samples and span multiple modalities, which allows for large-scale pretraining of AI systems known as foundation models. For instance, LAION-5B is the dataset used to scale CLIP and Stable Diffusion, while HowTo100M helps to learn procedural tasks, for example, through narrated videos. Medical datasets such as MIMIC-CXR, where ground truth is structured, for instance, e-coding from radiology reports, allow for the prototyping of interpretable clinical systems. Ego4D provides a large collection of egocentric videos, providing an avenue for research in first-person activity recognition and multimodal memory. As multimodal benchmarks evolve, they will begin to incorporate different dimensions such as multilingual, interactive, and even synthetic modalities (e.g., 3D point clouds and generated speech), while expanding the dimensions of multimodal learning.

5.2 Evaluation Metrics

Evaluating multimodal models is difficult because of the wide variety of outputs and tasks. Some key metrics include:

(1) **Recall@K:** How many correct matches are in the top-K matches retrieved? Recall is typically used in cross-modal retrieval tasks such as text-to-image or audio-to-text matching [24].

(2) **BLEU/CIDEr/METEOR:** Common in image and video captioning. BLEU, CIDEr, and METEOR are used to extract generated text and compare it to human reference text by human metrics based on n-gram overlap, consensus, and precision/recall [4].

(3) **FID (Fréchet Inception Distance):** FID is a measure of the quality of generated images. It compares distributions of features extracted from real and generated samples [13].

(4) **F1-Score/Accuracy:** The standard for classification tasks. E.g., sentiment analysis or medical diagnosis is a classification task where you would be interested in a performance metric [17].

(5) **SPICE/ROUGE/BERTScore:** Commonly used to assess semantic similarity in language generation and question answering.

Relatively new trends are research that uses human-in-the-loop evaluation to evaluate complex outputs (e.g., the quality of dialogue or visual reasoning), and even uses the language model on its own as an evaluator (e.g., GPT-as-a-judge). There are multimodal benchmarks, including VQAv2 and ScienceQA, that provide not only automatic scoring but also human assessment, which better reflects real-world performance.

6. CHALLENGES AND FUTURE DIRECTIONS

Though much has been accomplished, multimodal deep learning still faces several challenges that prohibit robustness, scalability, and widespread deployment.

Table 2. Representative Multimodal Datasets.

Dataset	Modalities	Domain	Use Case
COCO [15]	Image + Text	General	Captioning, retrieval
VQA v2.0 [10]	Image + Text (Q&A)	General	Visual question answering
AudioSet [9]	Audio + Video	General	Audio-visual classification
MIMIC-CXR [14]	Image + Text	Medical	Diagnosis and report generation
LAION-5B [20]	Image + Text	Web-scale	Pretraining, retrieval
HowTo100M [16]	Video + Text	Instructional video	Pretraining, video-language tasks
Ego4D [11]	Video + Audio + Text	Egocentric vision	Action recognition, narration

6.1 Robustness to Missing Modalities

Unfortunately, in the real world, this often means that inputs are missing. For instance, a user could turn off the webcam, or a meeting's audio could be mostly background chatter or noise. Multimodal systems must work reliably in this scenario. Training techniques such as modality dropout have been developed where inputs can be masked at random during training to simulate robustness.
Methods such as dynamic fusion models or conditional networks allow the model to adaptively ignore or reweight inputs when a modality is missing. There is even research that explores the notion of cross-modal prediction (i.e., hallucinating a missing input modality, say the audio input, based on an available modality, such as lip movement), which enables performance to gracefully degrade when a modality is missing.

6.2 Scalability and Efficiency

People are beginning to train massive multimodal models like Gemini and GPT-4V. These models consume staggering amounts of compute and memory to train, making them all but impossible to reproduce or enable accessibility. People have begun to investigate modular training (i.e., freezing vision/language backbones), low-rank adapters, and mechanisms for efficient attention (e.g., sparse attention, or linear attention) to circumvent the scalability problem. Some petabytes of useful model size reduction research have focused on knowledge distillation and quantization. As the world continues to build out edge-computing capabilities and cloud-optimized design, multimodal models will play an increasing role in human-machine interfacing in robotics, AR/VR applications at the edge, or simply for real-time deployment for oil and gas, construction, or energy-efficient building features.

6.3 Interpretability

Unfortunately, multimodal models act as black boxes, which makes it challenging to understand the rationale of how actions were made by the model. For example, in medical diagnosis tasks, it is crucial to know the affected regions of an X-RAY or the influenced clinical terms that inform a prediction, especially as part of a social contract with the user's confidence for future engagements.
Based on research, human-readable explanations are beginning to explore visual attention maps, cross-modal attention visualization, and chain-of-thought for natural language generation. The rationale explicitly focused on multimodal ensembles of human-generated text explanations that are grounded in images and videos, providing another mode of interpretability. Establishing interpretability will also protect against bias and maintain ethical sensibilities, too, particularly in sensitive applications of AI.

6.4 Multilingual and Cross-Cultural Learning

At present, the majority of multimodal models that exist are trained on primarily English data without consideration of cultural similarities or differences. Despite their shortcomings, this will mean that these models would only be effective in the West for global deployments, including health care, education, and social services. More recently, researchers are working on multilingual multimodal pretraining. While models like M3P and UC2 are quickly making strides to bridge the divide through alignment of visual representations and semantics in multiple languages [8], it remains important to consider and create data credits across highly variable datasets of various forms of media. It should be a consideration of other modes of representation to realize the goal of building more equitable AI systems.

6.5 Embodied Intelligence and AGI

The promise of multimodal extending into embodied, intelligent agents aligns with building avatars and robots that perceive their environments as they reason and act. Services will need to adhere to vision and language grounded in long-horizon memory to assess actions embedded in all four modalities, including unmixed motor control [7, 21].
Models such as PaLM-E and Gemini represent the first significant advances toward this vision of embodied progress. These multimodal models demonstrate reasoning with sequences of observations and can make high-level action plans through modality coupling of visual and textual sequences. Future multi-agent environments could leverage wearable sensors and headsets, spatialized audio-generated audio, and dynamic, real-world dialogues, creating temporally aware and safe forms of interaction with humans in homes, factories, and public including public transportation. Ultimately, capabilities will expand through some virtually personalized agent, where the development of multimodal learning will represent a sizable piece of the road-map toward general-purpose AI.

7. CONCLUSION

Multimodal deep learning has emerged as an important domain for advancing AI, providing ways to connect otherwise isolated data modalities to allow for sophisticated contextual reasoning. By using several modalities of visual images, language, audio, and other types of sensor data, multimodal models can generalize and provide robustness and flexibility that was not possible with unimodal systems. This review of the substantive literature in multimodal AI covers what multimodal learning is, various augmentation aspects and fusion strategies, describes system architectures, and surveys some of the most prominent multimodal systems to date, including CLIP, Flamingo, GPT-4V, and Gemini 1.5. Several instances of real-world applications of multimodal AI are summarized, span-

ning domains such as healthcare, robotics, education, autonomous vehicles, human-in-the-loop environments, and recommender systems. Existing benchmarking datasets and evaluation approaches that are shaping research and deployment are also discussed. However, despite the advances and momentum growing in the field, there are still many important limitations to overcome, including robust treatment of missing modalities, scalable and efficient training, interpretability, multilingual alignment, and the potential for embodied interaction. These considerations will be critical to overcome if multimodal AI advances to further support the development of capable and trustworthy general-purpose AI.

Looking to the future, the directions in which multimodal AI may develop include embodied agents, long context learning, cross-cultural generalization, and ethical alignment. These systems will not only recognize an image or understand a sentence, but there will be reasoning across modalities, retaining memories across timeframes or states, adapting across tasks, and interacting in rich real-world environments. Thus, multimodal deep learning represents not just a new area of research but also a new paradigm for the design and deployment of AI systems. The next evolution of intelligent technology, as they integrate context and augment the physical and social world, will rely on multimodal AI.

8. REFERENCES

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019.

[4] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[5] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[6] Daniele Di Mitri, Jan Schneider, and Hendrik Drachsler. The rise of multimodal tutors in education: insights from recent research. *Handbook of open, distance and digital education*, pages 1037–1056, 2023.

[7] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. Palm-e: An embodied multimodal language model. 2023.

[8] Hongliang Fei, Tan Yu, and Ping Li. Cross-lingual cross-modal pretraining for multimodal retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3644–3650, 2021.

[9] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[10] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[11] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022.

[12] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022.

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[14] Alistair E W Johnson, Tom J Pollard, Lu Shen, et al. Mimic-cxr: A large publicly available dataset of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

[16] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019.

[17] David MW Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*, 2020.

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[19] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[20] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

[21] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Perceiver-actor: A multi-task transformer for robotic manipulation. In *Conference on Robot Learning*, pages 785–799. PMLR, 2023.

[22] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

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

[24] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1386–1393, 2014.