Article

Not peer-reviewed version

Addressing Challenges in Multimodal Large Language Model Development

Feidlimid Shyama *, Lucas Pereira, João Souza, Ana Costa

Posted Date: 22 December 2025

doi: 10.20944/preprints202512.1986.v1

Keywords: multimodal large language models; cross-modal alignment; contrastive learning; multimodal pretraining; multimodal fusion; image captioning; visual question answering; human-robot interaction; bias mitigation; computational efficiency; fairness; explainability; multimodal sentiment analysis; model generalization; transfer learning; artificial intelligence; multimodal data

 doi:10.20944/preprints202512.1986.v1

Article

Addressing Challenges in Multimodal Large Language Model Development

Feidlimid Shyama [1,*], Lucas Pereira [2], João Souza [3] and Ana Costa [4]

[1] National Technical University of Athens, Greece
[2] University of São Paulo, Brazil
[3] University of Porto, Portugal
[4] University of Coimbra, Portugal
* Correspondence: feidlimid.shyama@ntua.gr

Abstract

Multimodal Large Language Models (MLLMs) have emerged as a powerful paradigm in artificial intelligence, enabling systems to process and reason over data from multiple modalities, such as text, images, video, and audio. By combining the strengths of different data types, MLLMs offer the potential to tackle more complex and nuanced tasks than traditional unimodal models. This paper provides a comprehensive survey of the current state of MLLMs, examining their architecture, training strategies, applications, and the challenges that remain in scaling and deploying these models. We begin by reviewing the core components of MLLMs, including the integration of modality-specific encoders and the development of joint multimodal representations. The training strategies that support the learning of multimodal interactions, such as contrastive learning, early and late fusion, and self-supervised pretraining, are discussed in detail. Furthermore, we explore a wide range of applications where MLLMs have demonstrated success, including visual-language understanding tasks like image captioning and visual question answering, multimodal sentiment analysis, and human-robot interaction. Despite their impressive capabilities, MLLMs face a number of significant challenges, such as issues with cross-modal alignment, missing modalities, computational inefficiency, and the presence of bias in multimodal datasets. The ethical concerns associated with fairness, interpretability, and accountability are also highlighted. We conclude by exploring future research directions that could help address these challenges and advance the field, including improvements in cross-modal fusion, multimodal pretraining paradigms, model efficiency, and bias mitigation strategies. As MLLMs continue to evolve, they are poised to play a transformative role in various industries, from healthcare and education to robotics and entertainment, by enabling machines to understand and interact with the world in a more human-like and contextually aware manner. This survey aims to provide a comprehensive overview of the current landscape of MLLMs, offering insights into both their potential and the hurdles that remain for their widespread adoption.

Keywords: multimodal large language models; cross-modal alignment; contrastive learning; multimodal pretraining; multimodal fusion; image captioning; visual question answering; human-robot interaction; bias mitigation; computational efficiency; fairness; explainability; multimodal sentiment analysis; model generalization; transfer learning; artificial intelligence; multimodal data

1. Introduction

Large Language Models (LLMs) have rapidly emerged as a foundational technology in artificial intelligence, demonstrating remarkable capabilities in natural language understanding, reasoning, and generation [1]. Built upon large-scale neural architectures and trained on massive corpora of text, LLMs have achieved state-of-the-art performance across a wide range of tasks, including question answering, machine translation, code synthesis, and dialog systems. Despite these successes, traditional LLMs are inherently limited by their unimodal nature: they primarily operate on textual

 doi:10.20944/preprints202512.1986.v1

inputs and outputs, while much of human intelligence and real-world information is intrinsically multimodal, encompassing vision, speech, audio, video, and structured sensory data [2]. This discrepancy has motivated a growing body of research aimed at extending LLMs beyond text, giving rise to Multimodal Large Language Models (MLLMs) [3]. Multimodal Large Language Models seek to unify heterogeneous modalities within a single coherent framework, enabling models to jointly perceive, reason over, and generate content across multiple data types. By integrating modalities such as images, videos, audio signals, and sometimes even embodied sensor data, MLLMs aim to approximate more closely the way humans interact with and understand the world. For instance, humans naturally combine visual perception with linguistic reasoning when describing a scene, following instructions, or answering questions about images. MLLMs attempt to replicate this capability by aligning representations across modalities and leveraging the strong reasoning and generative abilities of LLMs to perform complex multimodal tasks. The rapid development of MLLMs has been driven by several converging trends. First, advances in representation learning for individual modalities—such as convolutional neural networks and vision transformers for images, as well as self-supervised learning for speech and audio—have produced powerful pretrained encoders capable of extracting rich semantic features [4]. Second, the success of transformer-based architectures and large-scale pretraining paradigms has demonstrated that scaling model size and data can lead to emergent abilities, including in-context learning and zero-shot generalization. Third, the availability of large multimodal datasets, often consisting of paired image-text, video-text, or audio-text data collected from the web, has enabled effective cross-modal alignment and joint training. Together, these factors have created fertile ground for the emergence of MLLMs as a central research direction in modern AI [5]. From a modeling perspective, MLLMs encompass a diverse set of architectural designs and training strategies. Early approaches focused on connecting pretrained unimodal encoders with LLM backbones through lightweight projection layers or adapters, allowing visual or auditory features to be mapped into the token embedding space of language models. More recent methods explore deeper integration, including joint pretraining of multimodal transformers, cross-attention mechanisms that dynamically fuse information across modalities, and unified tokenization schemes that treat different modalities in a more homogeneous manner. These design choices reflect trade-offs between computational efficiency, scalability, flexibility, and performance, and they continue to evolve as the field matures. The capabilities of MLLMs have expanded rapidly, enabling a wide range of applications that were previously difficult or impossible with unimodal models. In computer vision and vision-language research, MLLMs have demonstrated strong performance on tasks such as image captioning, visual question answering, image-grounded dialog, and visual reasoning. In audio and speech domains, multimodal models can transcribe, translate, and reason about spoken language while incorporating contextual cues from text or vision [6]. In more complex settings, MLLMs are being explored for video understanding, robotics and embodied AI, medical image analysis, scientific data interpretation, and assistive technologies. These applications highlight the potential of MLLMs to serve as general-purpose multimodal interfaces between humans and machines. Despite their promise, MLLMs also introduce significant challenges and open questions. Training large multimodal models is computationally expensive and often requires carefully curated datasets to avoid spurious correlations and modality imbalance [7]. Aligning representations across modalities remains non-trivial, particularly when modalities differ in temporal resolution, noise characteristics, or semantic granularity [8]. Moreover, MLLMs raise important concerns related to robustness, interpretability, bias, and safety [9]. For example, errors or biases in one modality may propagate to others, leading to misleading or harmful outputs [10]. Understanding and mitigating such risks is crucial for the responsible deployment of MLLMs in real-world systems. Given the rapid pace of progress and the diversity of approaches in this area, a comprehensive survey of Multimodal Large Language Models is both timely and necessary [11]. This survey aims to provide a systematic overview of the field, covering key concepts, architectural paradigms, training methodologies, datasets, evaluation benchmarks, and application domains. We also discuss emerging trends, open challenges, and future

 doi:10.20944/preprints202512.1986.v1

research directions, with the goal of offering researchers and practitioners a unified perspective on the current state and potential of MLLMs. By synthesizing insights across disciplines, this survey seeks to clarify the landscape of multimodal large language modeling and to facilitate further advances toward more general, capable, and human-aligned artificial intelligence systems.

2. Preliminaries and Problem Formulation

In this section, we establish the mathematical foundations and formal definitions underlying Multimodal Large Language Models (MLLMs). We begin by introducing notations for modalities, datasets, and model components, followed by a unified formulation of multimodal representation learning and generation. This formalism provides a common language for analyzing architectures, training objectives, and inference mechanisms across diverse MLLM designs [12].

2.1. Modalities and Data Representation

Let $\mathcal{M} = \{m_1, m_2, \ldots, m_K\}$ denote a finite set of modalities, where each modality m_k corresponds to a distinct data type such as text, image, audio, or video. For each modality m_k, we define an input space $\mathcal{X}^{(k)}$ and an output space $\mathcal{Y}^{(k)}$ [13]. For example, for text we may have $\mathcal{X}^{(\text{text})} = \mathcal{Y}^{(\text{text})} = \Sigma^*$, where Σ is a vocabulary and Σ^* denotes the set of all finite token sequences; for images, $\mathcal{X}^{(\text{img})} \subset \mathbb{R}^{H\times W\times C}$; and for audio, $\mathcal{X}^{(\text{aud})} \subset \mathbb{R}^T$ or its time-frequency representations. A multimodal dataset is typically composed of aligned samples across a subset of modalities. Formally, we define a dataset

$$\mathcal{D} = \left\{ \left(x_i^{(k_1)}, x_i^{(k_2)}, \ldots, x_i^{(k_r)} \right) \right\}_{i=1}^{N},$$

where $\{k_1, \ldots, k_r\} \subseteq \{1, \ldots, K\}$ indicates the modalities present in each sample [14]. In many practical scenarios, the alignment is weak or noisy, and not all modalities are available for every sample. This motivates models that can handle missing modalities and perform conditional generation given arbitrary subsets of inputs [15].

2.2. Tokenization and Embedding Spaces

A central challenge in MLLMs is reconciling heterogeneous modalities within a unified representation space. Let $\mathcal{T}^{(k)}$ denote a tokenization function for modality m_k, mapping raw inputs to sequences of discrete or continuous tokens:

$$\mathcal{T}^{(k)} : \mathcal{X}^{(k)} \to \left(\mathbb{R}^{d_k} \right)^{n_k},$$

where n_k is the number of tokens and d_k is the dimensionality of modality-specific embeddings. For text, $\mathcal{T}^{(\text{text})}$ is typically a subword tokenizer followed by a learned embedding lookup; for images or audio, $\mathcal{T}^{(k)}$ often corresponds to a pretrained encoder (e.g., a Vision Transformer or a convolutional network) that outputs a sequence of patch- or frame-level embeddings. To enable multimodal fusion, these modality-specific embeddings are projected into a shared latent space $\mathbb{R}^d$ via projection functions

$$\phi^{(k)} : \mathbb{R}^{d_k} \to \mathbb{R}^d,$$

such that tokens from different modalities become comparable and combinable. The resulting unified token sequence for a multimodal input can be written as

$$\mathbf{Z} = \bigoplus_{k\in\mathcal{S}} \phi^{(k)} \left(\mathcal{T}^{(k)}(x^{(k)}) \right),$$

where $\mathcal{S} \subseteq \{1, \ldots, K\}$ is the set of observed modalities and $\oplus$ denotes sequence concatenation with optional modality-specific positional or type embeddings.

 doi:10.20944/preprints202512.1986.v1

2.3. Multimodal Transformer Backbone

Most contemporary MLLMs are built upon transformer architectures. Let $\mathbf{Z} \in \mathbb{R}^{n \times d}$ denote the concatenated multimodal token embeddings. A transformer layer applies self-attention and feed-forward operations:

$$\text{Attn}(\mathbf{Z}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V},$$

where $\mathbf{Q} = \mathbf{Z}W_Q$, $\mathbf{K} = \mathbf{Z}W_K$, and $\mathbf{V} = \mathbf{Z}W_V$. In the multimodal setting, attention can be unrestricted (allowing tokens from all modalities to attend to each other) or structured (e.g., via cross-attention), where text tokens attend to visual tokens but not vice versa. Stacking L such layers yields a multimodal encoder-decoder or decoder-only model parameterized by θ, which defines a conditional distribution over output token sequences:

$$p_\theta(\mathbf{y} \mid \mathbf{Z}) = \prod_{t=1}^{T} p_\theta(y_t \mid y_{<t}, \mathbf{Z}).$$

Here, $\mathbf{y}$ is typically a text sequence, reflecting the role of language as a universal interface for reasoning and communication in MLLMs [16].

2.4. Learning Objectives

Training objectives for MLLMs are designed to align representations across modalities and to endow the model with strong generative and reasoning capabilities [17]. A common objective is multimodal autoregressive language modeling, which minimizes the negative log-likelihood [18]:

$$\mathcal{L}_{\text{AR}}(\theta) = -\mathbb{E}_{(\mathbf{Z},\mathbf{y})\sim\mathcal{D}}\left[\sum_{t=1}^{T} \log p_\theta(y_t \mid y_{<t}, \mathbf{Z})\right].$$

In addition, contrastive objectives are often employed to enforce cross-modal alignment. Given paired samples $(x^{(i)}, x^{(j)})$ from modalities m_i and m_j, a contrastive loss such as InfoNCE can be written as

$$\mathcal{L}_{\text{CL}} = -\mathbb{E}\left[\log \frac{\exp\left(\text{sim}(\mathbf{z}^{(i)}, \mathbf{z}^{(j)})/\tau\right)}{\sum_{j'} \exp\left(\text{sim}(\mathbf{z}^{(i)}, \mathbf{z}^{(j')})/\tau\right)}\right],$$

where $\mathbf{z}^{(i)}$ and $\mathbf{z}^{(j)}$ are pooled representations, $\text{sim}(\cdot,\cdot)$ is a similarity function (e.g., cosine similarity), and τ is a temperature parameter. The total training objective is often a weighted combination:

$$\mathcal{L}(\theta) = \lambda_{\text{AR}}\mathcal{L}_{\text{AR}} + \sum_{(i,j)} \lambda_{ij}\mathcal{L}_{\text{CL}}^{(i,j)} + \mathcal{L}_{\text{aux}},$$

where $\mathcal{L}_{\text{aux}}$ includes auxiliary losses such as masked modeling, region-text alignment, or instruction-following objectives [19].

2.5. Inference and Multimodal Reasoning

At inference time, an MLLM performs conditional generation or prediction given a subset of modalities $\mathcal{S}$. The inference problem can be formalized as

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \; p_\theta(\mathbf{y} \mid \{x^{(k)}\}_{k\in\mathcal{S}}),$$

which is typically approximated using greedy decoding, beam search, or sampling-based methods. Crucially, the model's internal attention mechanisms implicitly perform multimodal reasoning by dynamically weighting information from different modalities across layers and time steps. This formal framework highlights that MLLMs can be viewed as probabilistic models over heterogeneous data sources, unified through shared representations and trained with objectives that balance alignment, generation, and reasoning [20]. The remainder of this survey builds upon these preliminaries to

 doi:10.20944/preprints202512.1986.v1

analyze concrete architectural instantiations, training pipelines, and evaluation methodologies for Multimodal Large Language Models.

3. Architectural Approaches in Multimodal Large Language Models

Multimodal Large Language Models (MLLMs) represent a significant shift from traditional unimodal approaches by introducing complex architectures that are capable of processing and generating information across multiple modalities. These models aim to capture intricate interactions between different data types, such as text, images, video, and audio, while maintaining the flexibility, scalability, and generalization capabilities that have made large language models so successful in the natural language processing (NLP) domain. In this section, we examine the diverse architectural paradigms that have emerged to address these challenges, from early fusion strategies to more advanced multimodal transformers and hybrid models. We focus on the key components of these architectures and how they have evolved to facilitate more effective multimodal learning. One of the foundational architectural approaches in MLLMs is early fusion, where individual modalities are processed independently and then combined at a later stage for joint representation learning [21]. Early fusion can take different forms, but the core idea is to use modality-specific encoders to extract features, which are then concatenated or otherwise combined to form a unified multimodal representation. This combined representation is then passed through further layers of the model to perform the downstream task, such as multimodal classification, reasoning, or generation. For instance, in early image-text fusion models, images might be processed by convolutional neural networks (CNNs) or vision transformers, while text is tokenized and processed through a language model. The features extracted from each modality are then concatenated and passed to a shared representation space, typically using simple projection layers that align the two modalities. While this approach is conceptually simple and effective for certain tasks, it suffers from the limitation that it does not allow the model to dynamically adjust the interaction between modalities during the learning process [22]. This can result in suboptimal performance when dealing with complex or highly correlated multimodal data [23]. In contrast to early fusion, late fusion architectures treat each modality as a separate stream, allowing the model to process and reason about each modality independently before combining them at a higher level. Late fusion architectures typically consist of modality-specific branches, where each modality is processed by its own set of encoders (e.g., a CNN for images, a transformer for text) to produce independent representations [24]. These representations are then combined at a later stage, often via a simple averaging or concatenation operation. While late fusion has the advantage of maintaining modality-specific processing, it may fail to fully capture the interactions between modalities, especially when the relationships between the different data types are complex [25]. Recent work has sought to overcome this limitation by introducing cross-attention mechanisms, which allow modalities to interact more effectively at the representation level. Cross-attention involves computing attention scores between the token embeddings of different modalities, facilitating the integration of information from one modality into the other. This mechanism has been shown to improve performance in tasks such as image captioning and visual question answering, where the relationship between the text and image is crucial for task success. More recently, transformer-based architectures have become the dominant framework for MLLMs, enabling more flexible and powerful ways to process and fuse multimodal data [26]. The transformer model, originally designed for natural language processing, is characterized by its use of self-attention mechanisms that allow tokens in a sequence to attend to all other tokens, regardless of their position. This mechanism has been shown to be highly effective at capturing long-range dependencies and contextual relationships in data, making it a natural fit for multimodal tasks where interactions between modalities may be complex and distant. In the context of MLLMs, transformers are often used to process multimodal token sequences that combine information from different modalities [27]. Each modality's tokenized inputs are embedded into a shared latent space using modality-specific encoders, and then the entire multimodal sequence is processed by a series of transformer layers. These layers allow the model to attend to and learn relationships between

 doi:10.20944/preprints202512.1986.v1

tokens from different modalities, effectively fusing the information across modalities at every layer [28]. The power of transformer-based models lies in their ability to scale with large datasets and model architectures, which has been a key factor in the success of recent MLLMs [29]. One specific approach that has gained significant traction in multimodal research is the multimodal transformer (MMT) architecture, which integrates cross-modal attention and encoder-decoder frameworks to jointly process multiple modalities. In these architectures, each modality is encoded into a shared space, and a set of attention layers is applied to enable dynamic interactions between modalities. The encoder part of the model generates a joint multimodal representation by attending to and fusing features from each modality, while the decoder generates the output (e.g., a text description or a prediction) based on this fused representation [30]. The attention mechanism in MMTs allows for the dynamic fusion of information based on the relevance of each modality at every step of the process, leading to more accurate and contextually aware outputs. This attention-driven fusion is especially useful in tasks like image captioning, where the text output depends heavily on the visual content, and vice versa [31]. One key design decision in multimodal transformers is the choice of cross-modal attention mechanisms [32]. In cross-attention, the model learns to attend to tokens from one modality based on the relevance of tokens from another modality. This can be done in several ways, such as using the output from one modality's encoder as a query for the attention mechanism while using the other modality's embeddings as keys and values. This allows the model to selectively focus on certain parts of the input, such as a specific region in an image or a key phrase in a text, that are most relevant to the task at hand. Cross-attention mechanisms have been shown to significantly improve performance in tasks like visual question answering (VQA), where the model must integrate both image content and textual questions to generate an appropriate answer [33]. In addition to the encoder-decoder and cross-attention models, hybrid architectures have also been explored in recent work. These hybrid models often combine transformer-based architectures with other deep learning techniques, such as convolutional networks or recurrent neural networks (RNNs), to process different modalities in a more specialized manner. For example, vision transformers (ViTs) are often used in conjunction with language transformers to process image and text data simultaneously [34]. These hybrid approaches are designed to leverage the strengths of different model types, enabling more efficient and accurate processing of multimodal data. While multimodal transformers and hybrid models have demonstrated impressive capabilities, they are not without challenges. One major hurdle is the computational cost associated with processing large multimodal datasets, especially when dealing with high-dimensional data such as images and videos. Transformers are known for their high memory and computational requirements, and these costs are exacerbated in multimodal settings where multiple modalities must be processed simultaneously. Techniques like sparse attention, multi-scale processing, and model distillation are being explored to address these challenges and make multimodal models more scalable. Furthermore, the issue of data alignment and modality imbalance remains a significant challenge in multimodal learning. In practice, not all modalities are available for every sample, and even when they are, they may not be perfectly aligned [35]. For example, in a video, the spoken dialogue might be misaligned with the visual content, or in a multi-modal medical dataset, the image and text annotations might not match up perfectly. Dealing with this misalignment is a key research question, and approaches such as multimodal contrastive learning, alignment loss functions, and semi-supervised learning are being developed to mitigate the impact of misalignment on model performance. To summarize, the architectural landscape of MLLMs is highly diverse and rapidly evolving. From early fusion to late fusion, cross-attention mechanisms, and multimodal transformers, these architectures provide a variety of ways to process and combine multimodal data [36]. While significant progress has been made, challenges such as computational efficiency, data alignment, and modality imbalance remain active areas of research. Future work will likely focus on optimizing these architectures to handle even larger and more complex multimodal datasets, improving the robustness and generalization of MLLMs across different tasks, and addressing the computational bottlenecks that arise in large-scale multimodal learning.

 doi:10.20944/preprints202512.1986.v1

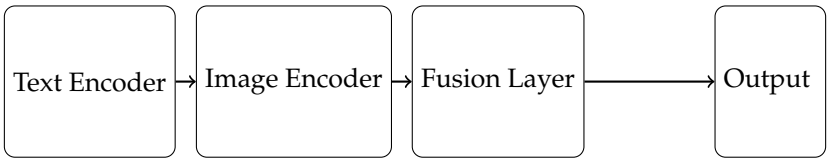

Figure 1. Overview of a Multimodal Large Language Model architecture with separate modality encoders (text and image) and a fusion layer combining the modalities for joint processing.

4. Training Strategies for Multimodal Large Language Models

Training Multimodal Large Language Models (MLLMs) is a complex and computationally demanding process that requires careful consideration of the unique challenges posed by multiple data modalities. Unlike unimodal models, which only deal with textual or visual inputs, MLLMs must simultaneously handle and align inputs from various domains such as text, image, video, and audio [37]. This introduces several critical issues that must be addressed during training, including cross-modal alignment, modality imbalance, missing modality handling, and the computational cost associated with processing large multimodal datasets [38]. In this section, we explore the primary strategies employed in the training of MLLMs, highlighting key methods, techniques, and challenges that arise when training models to perform tasks that involve multiple modalities. The first major consideration when training MLLMs is the alignment between modalities. Since data from different modalities often have different formats, temporal resolutions, and noise characteristics, it is crucial to design training objectives that ensure meaningful relationships between these modalities. Cross-modal alignment aims to ensure that representations from different modalities correspond to the same underlying semantic concepts [20]. This can be achieved by training the model to learn joint representations of the modalities that align at both the feature level and the task-specific level. The most commonly used approach to enforce alignment is via contrastive learning, where the model is trained to bring representations of the same instance (e.g., an image-text pair) closer together in the shared latent space, while pushing representations of unrelated instances apart. In this framework, the model learns to distinguish between aligned and non-aligned pairs by minimizing a contrastive loss function, such as InfoNCE or triplet loss. Another important training strategy involves modality-specific pretraining. For example, separate pretrained encoders for text (e.g., BERT or GPT) and vision (e.g., Vision Transformers or ResNet) can be used to process each modality before they are integrated into a multimodal model [39]. This strategy leverages the large amounts of available data for each modality, allowing each encoder to develop specialized representations that capture modality-specific features [40]. These modality-specific encoders can then be fine-tuned in a multimodal setting, enabling the model to learn both the individual modality representations and their interactions. This pretraining strategy is especially useful for tasks where each modality has its own unique structure (e.g., textual data consisting of words and sentences, and visual data consisting of pixel grids or regions in images), as it allows each modality to be represented in its most natural form before fusion [41]. Once the model has learned meaningful representations for each modality, the next step is to combine them effectively. The fusion strategy typically involves a multimodal fusion layer that integrates features from each modality into a joint multimodal representation. This fusion layer can operate at different stages in the model. One common approach is late fusion, where features from each modality are independently processed and then combined at the final stages of the model [42]. Another approach is early fusion, where modality-specific features are merged at the initial layers, enabling the model to learn cross-modal relationships early in the processing pipeline [43]. The decision between early and late fusion often depends on the task at hand and the complexity of the interactions between the modalities [44]. In tasks that require fine-grained reasoning between modalities, such as visual question answering (VQA) or image captioning, early fusion may be more effective as it allows the model to directly learn the relationships between modalities. On the other hand, in tasks where modalities are less interdependent, late fusion may be sufficient and computationally more efficient. A critical challenge in MLLM training is dealing with missing or incomplete modalities. In many

 doi:10.20944/preprints202512.1986.v1

real-world scenarios, one or more modalities may be unavailable or incomplete during training. For example, in a video question answering (VideoQA) task, some video frames may be missing, or in a medical diagnosis task, only the textual report may be available without the corresponding images [45]. To address this, models need to be robust to missing modalities. One approach to handle missing data is the use of masking techniques, where missing modalities are explicitly masked during training, and the model is trained to predict the missing modality or to perform the task with the available data. Another strategy is to leverage multimodal contrastive loss functions, where the model is trained to optimize the representations for the modalities that are available, while ignoring the missing ones. These approaches ensure that the model remains effective even in the absence of some modalities, which is crucial for real-world applications where data is often incomplete [46]. Training MLLMs is computationally intensive due to the large number of parameters involved and the need to process large multimodal datasets [47]. The sheer size of the model architecture and the datasets used for training mean that the model requires significant hardware resources, including high-performance GPUs or TPUs. As a result, many MLLMs employ techniques to reduce computational costs while maintaining performance. One common strategy is the use of model distillation, where a smaller, more efficient model is trained to approximate the behavior of a larger, more powerful model [48]. This distillation process helps reduce the memory and computation requirements of the model, making it more feasible to deploy in resource-constrained environments [49]. Additionally, techniques such as sparse attention and multi-scale processing are often employed to reduce the computational complexity of the attention mechanisms in transformer-based models. Sparse attention techniques limit the number of tokens that can attend to one another, making the attention mechanism more computationally efficient, while multi-scale processing allows the model to focus on different levels of granularity for different modalities [50]. Finally, the issue of modality imbalance is another key challenge when training MLLMs. In many multimodal datasets, one modality may be more abundant or more informative than others. For example, in a video-captioning task, the text data (captions) may be much shorter and more limited in quantity than the video data. To address this, researchers have explored different ways of balancing the contributions of each modality during training. One common approach is to weight the loss functions for each modality according to the number of available examples or the quality of the data [51]. Alternatively, techniques such as modality-specific regularization or adversarial training have been proposed to ensure that the model does not become overly reliant on one modality at the expense of others. The effectiveness of these training strategies depends on the task and the specific multimodal model being used [52]. For example, contrastive learning-based approaches are highly effective for tasks that involve joint reasoning across multiple modalities, such as image-text retrieval or multimodal classification. On the other hand, models trained with modality-specific pretraining and early fusion may be more suitable for tasks that involve intricate reasoning within each modality, such as image captioning or visual question answering. In either case, the challenge lies in ensuring that the model learns to effectively integrate and reason over multimodal inputs while avoiding overfitting to any single modality. In summary, training Multimodal Large Language Models involves a complex interplay of strategies designed to address the unique challenges posed by multimodal data [53]. These strategies include pretraining on modality-specific encoders, fusion at various stages of the model, contrastive learning for cross-modal alignment, handling missing modalities, and computational optimizations such as model distillation and sparse attention. As the field of MLLMs continues to evolve, these training techniques will be refined and expanded to enable even more powerful and efficient models that can handle the complexities of multimodal data across a wide range of real-world applications.

 doi:10.20944/preprints202512.1986.v1

Table 1. Summary of common training strategies for Multimodal Large Language Models.

Training Strategy	Description	Advantages	Challenges
Contrastive Learning	Optimizes joint representations of aligned modality pairs	Improves cross-modal alignment	Requires large amounts of aligned data
Modality-Specific Pretraining	Pretrain individual modality encoders before multimodal fine-tuning	Leverages modality-specific features	Requires extensive pretraining data for each modality
Early Fusion	Combines modality-specific features at the initial layers of the model	Enables fine-grained multimodal interactions early in processing	Computationally expensive due to large input size
Late Fusion	Combines modality-specific features at the final layers of the model	Simpler architecture and easier to train	May fail to capture complex modality interactions
Model Distillation	Transfers knowledge from a large model to a smaller, more efficient one	Reduces memory and computation requirements	Distilled models may lose some accuracy
Sparse Attention	Limits the number of tokens that can attend to each other to reduce complexity	Reduces computational cost	May decrease model performance if too aggressive

5. Applications of Multimodal Large Language Models

The advent of Multimodal Large Language Models (MLLMs) has opened up a wide range of possibilities in various application domains, allowing for a more seamless interaction between multiple data modalities. By leveraging the power of combining and reasoning over text, images, video, audio, and even sensor data, MLLMs are bridging the gap between how machines and humans perceive and interact with the world. This section delves into the most impactful applications of MLLMs across diverse fields, ranging from natural language processing (NLP) and computer vision to robotics, healthcare, and entertainment. We examine the potential of MLLMs to revolutionize these domains and discuss how their capabilities are shaping real-world use cases [54,55]. One of the most notable areas where MLLMs have made significant strides is in the realm of visual-language understanding, particularly in tasks such as image captioning, visual question answering (VQA), and image-text retrieval [56]. In image captioning, an MLLM is tasked with generating a natural language description of an image, a task that requires a deep understanding of the visual content as well as the ability to describe it coherently in words. Traditional image captioning models relied heavily on CNN-based image encoders and LSTMs or transformers for generating captions. However, MLLMs, through the use of joint training on both image and text data, have enabled the generation of more accurate and contextually relevant captions by capturing the interactions between the visual and textual components of the input [57]. Similarly, in VQA, models are trained to answer questions about images, which may range from simple queries (e.g., "What is the color of the car?") to more complex ones (e.g., "What is the person in the image doing?"). MLLMs, by integrating text and vision through cross-modal attention mechanisms, allow the model to reason about the question and the image in parallel, ensuring that the generated answers are grounded in both the visual content and the textual input. Image-text retrieval, another important application, involves searching for images based on a textual query (or vice versa). MLLMs have excelled in this task by learning to align image and text representations in a shared multimodal space, enabling more accurate retrieval based on semantic meaning rather than mere keyword matching [58]. In the field of audio and speech processing, MLLMs are revolutionizing tasks such as speech-to-text, audio-visual speech recognition, and multimodal sentiment analysis. Speech-to-text models have long been a key area of research, and while traditional models focused on converting spoken language into text, MLLMs take this a step further by incorporating additional modalities to enhance recognition accuracy and contextual understanding. For example, audio-visual speech recognition systems use both audio (speech) and visual (lip movement) signals to improve transcription accuracy, especially in noisy environments where audio alone may not suffice [59]. The integration of vision and speech allows for a more robust model that can disambiguate homophones or detect nuances in speech that would otherwise be missed [60]. Furthermore, MLLMs are being used in multimodal sentiment analysis, where the

 doi:10.20944/preprints202512.1986.v1

goal is to determine the sentiment (e.g., positive, negative, neutral) of a speaker based on both their speech (tone, pitch, cadence) and facial expressions or gestures. By jointly analyzing auditory and visual cues, these models offer a more nuanced understanding of sentiment, which is valuable in applications ranging from customer service chatbots to social media monitoring. In robotics and embodied AI, MLLMs have the potential to vastly improve task execution, human-robot interaction (HRI), and visual navigation [61]. Robotics tasks often involve multimodal sensor data, such as camera images, depth maps, and tactile sensors. MLLMs can integrate these data sources to enable robots to better understand their environment and interact with humans [62]. For example, in human-robot interaction, an MLLM can enable a robot to understand both spoken commands and visual cues from the environment, allowing it to respond to instructions in a more human-like manner [63]. This is particularly important in settings where the robot needs to interpret ambiguous or incomplete information, as it can use both linguistic and visual context to make more informed decisions. In visual navigation, robots equipped with cameras and other sensors must interpret their surroundings to navigate through complex environments. MLLMs can fuse the visual input with other sensory data, such as LiDAR or GPS, to improve the robot's ability to move autonomously, avoid obstacles, and interact with humans effectively. Furthermore, MLLMs can be used to facilitate task planning and reasoning, where a robot must understand high-level instructions and break them down into actionable steps, using both visual and textual information. In healthcare, the potential applications of MLLMs are equally transformative. One of the most promising areas is medical image analysis, where multimodal models are being used to assist in diagnosing and interpreting medical images, such as X-rays, MRIs, and CT scans [64]. Traditional models often focus on single-modal analysis, such as classifying medical images based solely on pixel data. However, MLLMs can combine medical images with associated text, such as doctor's notes, patient histories, or pathology reports, to improve diagnostic accuracy. For instance, in the case of radiology, a model could simultaneously analyze an X-ray image and interpret the accompanying radiologist's report to identify inconsistencies or potential diagnoses [65]. Furthermore, MLLMs are being used to develop multimodal decision support systems, which combine clinical data (such as lab results, patient demographics, and medical histories) with medical imaging to provide more comprehensive and accurate clinical decision-making support [66]. This capability is particularly valuable in areas like oncology, where the ability to fuse imaging data with textual data can improve the detection of early-stage cancers or predict patient outcomes. In entertainment and media, MLLMs are enabling more immersive experiences in content generation, recommendation systems, and interactive media. In content generation, MLLMs can be used to create storylines, scripts, and even artwork based on input from multiple modalities. For instance, a model might take in a text prompt, generate a detailed visual scene, and provide corresponding audio (e.g., sound effects or dialogue) to create a complete multimedia experience. This capability has significant implications for industries such as gaming, animation, and virtual reality, where creating cohesive and engaging environments often requires the integration of various media types [67]. Additionally, MLLMs are being incorporated into recommendation systems, where they analyze user preferences across multiple modalities—such as past viewing history (video), user reviews (text), and user ratings (audio or visual)—to provide more accurate and personalized recommendations. For example, a movie recommendation system might take into account both textual reviews and visual cues from trailers to suggest films that align with a user's tastes [68]. Lastly, MLLMs are playing a crucial role in natural language processing (NLP) tasks by enhancing models' ability to reason across modalities. Multimodal translation, for instance, involves translating both text and images together, such as when translating a webpage that contains both text and embedded images. MLLMs can handle the translation of the text while ensuring that the images are contextually and visually coherent with the translated content. Additionally, multimodal commonsense reasoning and multimodal dialog systems are benefiting from MLLMs, as they allow machines to answer questions, participate in conversations, and generate responses that take into account both the visual and linguistic context of the situation. This is particularly useful in conversational AI, where understanding user queries often requires

 doi:10.20944/preprints202512.1986.v1

not just interpreting text but also understanding visual cues (such as facial expressions or gestures) that accompany spoken language [69]. MLLMs have thus made significant strides in pushing the boundaries of what machines can understand and generate, by moving beyond language alone and incorporating a more holistic understanding of the world. The diverse range of applications of MLLMs underscores the transformative potential of this technology across industries. From visual-language tasks like image captioning and VQA, to multimodal sentiment analysis, medical image interpretation, and entertainment content generation, MLLMs are reshaping how we interact with machines and how machines perceive the world. As the field continues to evolve, it is likely that even more innovative applications will emerge, enabling more seamless, intuitive, and human-like interactions between machines and their environments.

6. Challenges and Open Issues in Multimodal Large Language Models

While the field of Multimodal Large Language Models (MLLMs) has made remarkable progress in recent years, there remain several critical challenges and open issues that must be addressed to further advance the technology and make it more effective, scalable, and generalizable across a range of real-world applications. These challenges span various aspects of model design, training, data, and deployment, and require deep research into new methodologies, frameworks, and approaches to overcome the limitations that currently exist. In this section, we delve into some of the most pressing challenges in the development of MLLMs, discussing their implications and potential directions for future research. One of the most fundamental challenges in MLLMs is cross-modal alignment [70]. Cross-modal alignment refers to the ability of a model to learn representations where different modalities (such as text, images, audio, and video) share a common semantic space, allowing the model to effectively relate them to one another [71]. Despite recent advances in contrastive learning, which aligns representations by bringing together corresponding image-text pairs, this problem remains far from solved. In many real-world scenarios, the alignment between modalities can be weak, noisy, or incomplete. For example, in video captioning tasks, the alignment between spoken language, visual content, and action in the video can be difficult to establish, especially when the video contains ambiguous scenes or overlapping activities. Furthermore, different modalities may have different granularity or levels of abstraction, which can further complicate the task of finding a common space for them to interact. Text, for instance, has a discrete structure (e.g., sentences, words, and tokens), while images are continuous in nature (e.g., pixel grids, regions, and visual features). Developing techniques that can bridge this gap and ensure meaningful alignment between these diverse data types is a major area of ongoing research. To address these issues, recent work has focused on more sophisticated fusion techniques, including hierarchical modeling, where multimodal inputs are processed at different levels of abstraction (e.g., word, sentence, paragraph, image region), and dynamic alignment strategies, where the model can adjust its attention based on the relevance of each modality at different stages of processing. Another significant challenge in MLLMs is the efficient handling of missing modalities [72]. In many practical applications, some modalities may be missing or unavailable for certain instances [73]. For example, in a multimodal medical diagnosis system, a doctor's notes might be available, but the corresponding medical image (e.g., X-ray) might be absent, or in a multimodal conversational agent, some users might only provide text while others provide both speech and visual inputs. Current MLLMs often struggle with these missing modality scenarios, as they are typically trained on datasets where each modality is available for every sample [74]. Dealing with missing modalities requires models to be flexible and robust enough to make accurate predictions or generate meaningful responses even when part of the input is missing. Solutions to this problem typically involve masking or imputation strategies, where missing modalities are either replaced with placeholder values (e.g., a special "mask" token) or inferred from the available data. Another approach is to incorporate multimodal attention mechanisms that can selectively focus on the available modalities while ignoring the missing ones [75]. However, these techniques are still in their early stages, and more advanced methods need to be developed to improve the model's ability to

 doi:10.20944/preprints202512.1986.v1

handle missing modalities gracefully, especially in real-world, dynamic settings where the availability of modalities can fluctuate. The computational cost of training and deploying MLLMs is another major hurdle. The training process for these models is extremely resource-intensive due to the large number of parameters involved, the need for multimodal data, and the complexity of processing and fusing multiple modalities simultaneously [76]. Transformer-based architectures, which are commonly used for MLLMs, are notorious for their quadratic complexity with respect to the sequence length, making them computationally expensive when dealing with high-dimensional input data like images and videos. Furthermore, as MLLMs often require the processing of large-scale datasets to achieve state-of-the-art performance, the amount of required computational power increases significantly. This issue is compounded by the need for specialized hardware such as GPUs or TPUs to handle the vast amount of data and computations involved. The financial and environmental costs associated with training large-scale multimodal models are considerable, raising concerns about the sustainability of such approaches in the long run. To address these issues, several techniques are being explored to reduce the computational load, such as sparse attention, where the attention mechanism is restricted to a subset of tokens, and model distillation, where a smaller, more efficient model is trained to mimic the performance of a larger one [77]. Furthermore, multi-task learning and transfer learning strategies are being used to reduce the amount of training data and computational effort required, by leveraging pre-trained models and fine-tuning them for specific multimodal tasks. Another pressing challenge is the issue of bias and fairness in MLLMs. Like all large AI models, MLLMs are susceptible to the biases present in the data they are trained on [78]. Given that multimodal datasets often consist of real-world data collected from the web or other sources, they can reflect societal biases related to race, gender, ethnicity, or socioeconomic status. These biases can be inadvertently learned by the model, leading to harmful or discriminatory outcomes when the model is deployed in real-world applications [79]. For example, a multimodal model trained on biased image-caption pairs may generate captions that reinforce stereotypes or make inappropriate assumptions based on visual content [80]. Additionally, multimodal models may amplify biases in specific modalities, such as text or image, and propagate them through the model's decision-making process. Addressing these biases requires careful attention during the data collection, preprocessing, and model training stages [81]. Techniques like data augmentation, adversarial training, and bias mitigation can be employed to reduce bias, but these methods often need to be customized for specific tasks and modalities. More importantly, rigorous evaluation metrics that account for fairness and bias are needed to assess the performance of MLLMs in real-world settings, ensuring that the models are not only accurate but also ethical and inclusive. The interpretability and explainability of MLLMs also remain significant challenges [82]. As these models grow in complexity, it becomes increasingly difficult to understand how they arrive at their predictions or decisions, especially when they operate across multiple modalities [83]. This lack of transparency is a serious issue for critical applications such as healthcare, finance, and law enforcement, where stakeholders must be able to trust and understand the model's outputs. In multimodal settings, the challenge is even more pronounced, as the model must process and integrate information from several data types that each have their own characteristics and structures. For instance, how can a multimodal model explain why it chose a particular image-text pair in a retrieval task, or why it generated a specific response in a multimodal dialog system? To address this, recent work has focused on attention visualization and explanation generation, where models are designed to highlight the parts of the input (e.g., regions in an image, words in a text) that contributed most to the model's decision. However, these techniques are still evolving, and much more work is needed to improve the interpretability of multimodal models in a way that is both accurate and accessible to non-experts. Finally, generalization is another critical challenge for MLLMs. While these models have shown impressive performance on the tasks they are trained on, they often struggle to generalize to new, unseen modalities or domains. For instance, an MLLM trained on image-captioning tasks might not perform as well when tasked with generating captions for medical images, or an MLLM trained on English-language data might not generalize to other languages or cultural contexts. Generalization

 doi:10.20944/preprints202512.1986.v1

is particularly challenging because multimodal models must learn to reason not just within a single modality, but across multiple modalities, each with its own distribution and variation [84]. To improve generalization, researchers are exploring domain adaptation, few-shot learning, and unsupervised learning approaches, where models are trained to be more flexible and adaptable to new contexts with limited data. Additionally, techniques like multimodal transfer learning and zero-shot learning are being explored, where models can apply knowledge learned from one domain or modality to a different one, without requiring extensive retraining. In conclusion, while Multimodal Large Language Models hold great promise for transforming a wide range of industries and applications, significant challenges remain in their development and deployment [85]. These challenges include issues related to cross-modal alignment, missing modalities, computational efficiency, bias, fairness, interpretability, and generalization [86]. As the field continues to evolve, addressing these open issues will be crucial to unlocking the full potential of MLLMs and ensuring that they can be deployed safely, effectively, and equitably in real-world applications. The solutions to these challenges will likely involve a combination of novel algorithms, better data management strategies, and more robust evaluation frameworks, as well as greater collaboration between researchers, industry professionals, and policymakers.

7. Future Directions and Research Trends in Multimodal Large Language Models

The field of Multimodal Large Language Models (MLLMs) is still in its early stages, yet it has already shown tremendous potential across a wide array of applications, from natural language processing and computer vision to robotics and healthcare. However, as discussed in previous sections, several challenges remain, including issues related to cross-modal alignment, model interpretability, computational efficiency, and bias. To address these challenges and unlock the full potential of MLLMs, ongoing research will need to explore a variety of novel techniques and paradigms [87]. This section looks ahead to the future of MLLMs, outlining key research trends, emerging methodologies, and exciting opportunities for further development. One of the most promising research directions in the MLLM field is the continued improvement of cross-modal representation learning. As noted earlier, one of the most fundamental challenges in MLLMs is learning meaningful representations that can align and integrate information across diverse modalities, such as text, image, video, and audio. While contrastive learning-based approaches, such as CLIP (Contrastive Language-Image Pretraining), have made significant progress in learning joint representations for image and text, this approach is still limited when it comes to incorporating other modalities or dealing with more complex, nuanced relationships between modalities. Future research will likely focus on developing multimodal contrastive learning strategies that extend beyond image-text pairs to include multimodal scenarios involving audio, video, or even sensor data. Additionally, approaches that combine supervised, unsupervised, and semi-supervised learning strategies will become more critical as the need for more robust, scalable, and data-efficient models grows [88]. Research may explore how to effectively transfer knowledge learned in one modality to another, such as using text descriptions to improve the performance of models trained for visual tasks or vice versa [89]. In parallel with improving multimodal representation learning, the development of dynamic fusion mechanisms will play a crucial role in advancing MLLMs [90]. Most current MLLMs rely on fixed fusion strategies, such as early fusion (concatenating modality-specific features early in the model) or late fusion (combining modality features in later stages). However, these approaches often fail to capture the highly dynamic and interdependent nature of multimodal data. For instance, in tasks like video captioning or multimodal question answering (VQA), the relationship between text and video content is often fluid and can vary depending on the specific context. As such, there is a growing interest in dynamic fusion strategies, where the interaction between modalities can be adapted during both training and inference. For example, recent developments in cross-attention mechanisms allow the model to attend to different parts of the input depending on the specific question or context at hand, ensuring that the fusion process is more flexible and context-sensitive. Further advancements in adaptive fusion—where the model learns which modalities to prioritize based on the task—will be key to improving the efficiency and

 doi:10.20944/preprints202512.1986.v1

effectiveness of MLLMs [91]. A particularly exciting area for future research lies in the development of multimodal pretraining paradigms. Much like how the success of large language models like GPT, BERT, and T5 has been fueled by pretraining on massive text corpora, MLLMs will benefit from pretraining on large-scale, diverse multimodal datasets [92]. However, due to the heterogeneous nature of multimodal data, designing effective pretraining tasks that can jointly optimize across multiple modalities is no trivial feat. Pretraining tasks such as image-text matching, image captioning, visual question answering, and masked modality prediction are commonly used in current MLLMs, but future work may explore more complex pretraining objectives that can simultaneously optimize the alignment and fusion of all available modalities. For example, a model might be pretrained to predict both the textual caption and the visual content from an image, while also predicting audio data associated with the same image. Additionally, self-supervised learning methods, where models can learn from unlabeled multimodal data, are expected to become more prevalent [93]. The ability to leverage large amounts of unlabeled data—especially in settings where multimodal data is abundant but labeled examples are scarce—will dramatically reduce the dependency on expensive labeled datasets and help scale MLLMs to a broader range of tasks and domains. Another critical research direction involves improving the computational efficiency of MLLMs. As highlighted in previous sections, training and deploying multimodal models requires substantial computational resources, often resulting in high memory and processing costs [94]. The problem is particularly acute for large-scale models like GPT-3 or Vision Transformers (ViTs), which require hundreds of gigabytes of memory and days or weeks of training on powerful hardware. While techniques like sparse attention and model distillation have been proposed to reduce the computational burden, further advancements are necessary to make MLLMs more efficient and scalable. One promising avenue is efficient transformer architectures, which aim to reduce the quadratic complexity of attention mechanisms by utilizing sparse, local, or low-rank approximations. Research in memory-efficient models, such as those that use techniques like memory compression and parameter sharing, will also become critical as the scale of multimodal data continues to grow. Moreover, the development of hardware accelerators specifically optimized for multimodal tasks—such as custom GPUs or TPUs—could provide significant improvements in both training time and energy efficiency, making MLLMs more feasible for real-world deployment. In the context of data efficiency, few-shot learning and zero-shot learning are likely to become key techniques for future MLLMs. One of the major limitations of current multimodal models is their need for large amounts of labeled data to perform well [95]. However, few-shot learning methods, which allow models to generalize from only a handful of labeled examples, have shown promising results in recent years, particularly in the NLP domain with models like GPT-3 [96]. For MLLMs, few-shot learning can help mitigate the challenges posed by the scarcity of high-quality multimodal datasets. In particular, zero-shot learning, where a model is asked to generalize to tasks or domains it has not seen during training, offers the potential for MLLMs to be deployed in more flexible and adaptive ways. Research in multimodal meta-learning and transfer learning will be essential in developing MLLMs that can quickly adapt to new tasks with limited data, enabling them to perform a wide range of applications without the need for extensive retraining [97]. As MLLMs continue to advance, ethics, fairness, and bias mitigation will remain some of the most important concerns. Given the complexity and diversity of data used in training MLLMs—spanning text, images, video, and audio—these models are highly susceptible to inheriting biases present in the data. Biases related to gender, race, culture, and socioeconomic status can be unintentionally learned by MLLMs, leading to harmful or discriminatory outputs. Future research will focus on the development of bias mitigation techniques that can be incorporated into the training pipeline to reduce harmful biases in the learned representations. These techniques may involve fairness constraints during training, where the model's predictions are regularly evaluated for bias, or the use of adversarial training, where biased outputs are penalized to ensure more equitable outcomes [98]. Furthermore, research into interpretability and explainability will be crucial for ensuring that MLLMs operate transparently and are accountable for their decisions. The ability to explain how a model arrived at a particular

 doi:10.20944/preprints202512.1986.v1

decision or output, particularly when processing multimodal inputs, will be vital for high-stakes applications like healthcare, criminal justice, and finance [99]. Human-AI collaboration is another promising avenue for future research in MLLMs. One of the most exciting aspects of MLLMs is their potential to facilitate more natural and intuitive interactions between humans and machines [100]. By understanding and reasoning across multiple modalities, MLLMs could significantly enhance areas such as human-robot interaction (HRI), augmented reality (AR), and virtual assistants. Future research may focus on interactive multimodal systems that can engage with users across multiple channels, such as text, voice, and gesture. These systems could be used in a variety of settings, from personal assistants that understand both spoken commands and visual gestures, to collaborative robots (cobots) that work alongside humans in industrial or healthcare environments [101]. In such systems, the model would need to adapt to dynamic interactions, learn from continuous input, and provide contextually appropriate responses that consider both visual and linguistic signals. This would also require a more robust understanding of intent recognition, where the system can discern the user's goal based on multimodal cues, and affective computing, where the system can recognize and respond to emotions expressed through both text and visual cues. Finally, the integration of multimodal learning with multimodal reasoning is poised to be a key focus for future research [102]. As MLLMs progress in terms of representation and integration across modalities, they will need to enhance their ability to reason over these diverse data sources [103]. This could include tasks such as multimodal commonsense reasoning, where a model needs to combine knowledge from both textual and visual inputs to make inferences about the world [104]. For example, in an autonomous driving scenario, a model might need to process visual inputs (e.g., images of the road, traffic signals, pedestrians) and combine them with contextual text-based inputs (e.g., navigation instructions, road conditions) to make safe and efficient driving decisions. Multimodal reasoning is also crucial for improving the generalization of models across different modalities and domains. Future MLLMs must be able to apply learned knowledge to unseen modalities or novel tasks with minimal supervision, enabling them to effectively operate in dynamic, real-world environments [105]. In conclusion, the future of Multimodal Large Language Models holds immense promise across multiple domains, driven

8. Conclusions

In this survey, we have explored the current state of Multimodal Large Language Models (MLLMs), a rapidly evolving area of artificial intelligence that seeks to integrate and reason over multiple modalities such as text, images, audio, and video [17]. These models represent a significant leap forward from traditional unimodal systems, offering the potential to understand and generate more complex, human-like outputs by leveraging the complementary nature of various data types. Through an examination of their architecture, training strategies, applications, challenges, and future directions, we have highlighted both the promise and the hurdles that MLLMs face as they continue to mature. One of the key strengths of MLLMs lies in their ability to bridge the gap between different sensory inputs, enabling systems to generate richer, more accurate representations of the world. By learning joint representations across multiple modalities, MLLMs have shown impressive performance in a wide range of tasks, from image captioning and visual question answering to multimodal sentiment analysis and autonomous robotics [106]. These capabilities open up exciting possibilities in diverse fields such as healthcare, education, entertainment, and human-robot interaction, where the ability to process and reason over multiple types of data is essential [35]. Despite these advancements, there are still several significant challenges that need to be addressed in order for MLLMs to reach their full potential. Cross-modal alignment remains a difficult problem, particularly as the complexity and heterogeneity of data increase. Dealing with missing modalities, handling modality imbalance, and ensuring the robustness of models to noisy data are critical challenges for practical deployment. Moreover, the computational cost associated with training large multimodal models is a barrier to scalability, particularly in resource-constrained environments. Advances in efficient model architectures, such as sparse attention mechanisms and model distillation, are helping to mitigate these issues, but

 doi:10.20944/preprints202512.1986.v1

further innovations are needed. In addition to technical challenges, there are also important ethical considerations surrounding the deployment of MLLMs. Bias in multimodal datasets, which can lead to unfair or discriminatory outcomes, remains a major concern. Ensuring that MLLMs are interpretable and explainable is equally crucial, particularly for applications in high-stakes areas like healthcare, finance, and law enforcement [107]. As the field progresses, it will be essential to develop fairer and more transparent models that prioritize inclusivity and accountability.

Looking to the future, we identified several promising research directions for advancing MLLMs. These include improving cross-modal alignment through more sophisticated fusion and contrastive learning strategies, developing multimodal pretraining techniques, and enhancing the scalability of MLLMs through more efficient architectures and computational strategies. The integration of multimodal reasoning and the ability to generalize across domains with minimal supervision are also areas that will drive significant progress. Additionally, as MLLMs become more integrated into real-world applications, their ability to interact seamlessly with humans across multiple modalities will become increasingly important, paving the way for more intuitive and adaptive AI systems.

Ultimately, the journey of Multimodal Large Language Models is one that is just beginning. While we have made significant strides in both research and application, there is still much work to be done in order to unlock the full potential of these models. As the field evolves, continued interdisciplinary collaboration—spanning machine learning, linguistics, computer vision, ethics, and human-computer interaction—will be crucial to ensuring that MLLMs are not only powerful but also responsible and beneficial to society. As we look forward, it is clear that MLLMs will play a transformative role in shaping the future of artificial intelligence, offering new ways for machines to understand, interpret, and interact with the rich, multimodal world we live in.

References

1. Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM* **2021**, *65*, 99–106.
2. Hong, Y.; Zhen, H.; Chen, P.; Zheng, S.; Du, Y.; Chen, Z.; Gan, C. 3d-llm: Injecting the 3d world into large language models. *NeurIPS* **2023**.
3. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* **2020**, *33*, 9459–9474.
4. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the CVPR, 2022.
5. Srinivasan, K.; Raman, K.; Chen, J.; Bendersky, M.; Najork, M. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In Proceedings of the Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 2443–2449.
6. Bustos, A.; Pertusa, A.; Salinas, J.M.; De La Iglesia-Vaya, M. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis* **2020**, *66*, 101797.
7. Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.H.; Li, Z.; Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the International conference on machine learning. PMLR, 2021, pp. 4904–4916.
8. Bu, H.; Du, J.; Na, X.; Wu, B.; Zheng, H. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In Proceedings of the O-COCOSDA, 2017.
9. Hu, X.; Gan, Z.; Wang, J.; Yang, Z.; Liu, Z.; Lu, Y.; Wang, L. Scaling up vision-language pre-training for image captioning. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 17980–17989.
10. Zhang, W.; Wang, X.; Nie, W.; Eaton, J.; Rees, B.; Gu, Q. MoleculeGPT: Instruction Following Large Language Models for Molecular Property Prediction. In Proceedings of the NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development, 2023.
11. Rae, J.W.; Potapenko, A.; Jayakumar, S.M.; Lillicrap, T.P. Compressive transformers for long-range sequence modelling. *arXiv preprint arXiv:1911.05507* **2019**.

 doi:10.20944/preprints202512.1986.v1

12. Lipping, S.; Sudarsanam, P.; Drossos, K.; Virtanen, T. Clotho-aqa: A crowdsourced dataset for audio question answering. In Proceedings of the 2022 30th European Signal Processing Conference (EUSIPCO). IEEE, 2022, pp. 1140–1144.
13. Wang, Y.; Ma, X.; Zhang, G.; Ni, Y.; Chandra, A.; Guo, S.; Ren, W.; Arulraj, A.; He, X.; Jiang, Z.; et al. MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. *arXiv preprint arXiv:2406.01574* **2024**.
14. He, X.; Zhang, Y.; Mou, L.; Xing, E.; Xie, P. Pathvqa: 30000+ questions for medical visual question answering. *arXiv:2003.10286* **2020**.
15. Liang, H.; Li, J.; Bai, T.; Chen, C.; He, C.; Cui, B.; Zhang, W. KeyVideoLLM: Towards Large-scale Video Keyframe Selection. *arXiv preprint arXiv:2407.03104* **2024**.
16. Lu, K.; Yuan, H.; Yuan, Z.; Lin, R.; Lin, J.; Tan, C.; Zhou, C.; Zhou, J. # InsTag: Instruction Tagging for Analyzing Supervised Fine-tuning of Large Language Models. In Proceedings of the The Twelfth International Conference on Learning Representations, 2023.
17. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, 2019.
18. Hsu, C.; Verkuil, R.; Liu, J.; Lin, Z.; Hie, B.; Sercu, T.; Lerer, A.; Rives, A. Learning inverse folding from millions of predicted structures. In Proceedings of the International Conference on Machine Learning. PMLR, 2022, pp. 8946–8970.
19. Shu, F.; Zhang, L.; Jiang, H.; Xie, C. Audio-Visual LLM for Video Understanding. *arXiv preprint arXiv:2312.06720* **2023**.
20. Liu, H.; Li, C.; Li, Y.; Lee, Y.J. Improved baselines with visual instruction tuning. *arXiv:2310.03744* **2023**.
21. He, X.; Zhao, K.; Chu, X. AutoML: A survey of the state-of-the-art. *Knowledge-based systems* **2021**, *212*, 106622.
22. Xu, R.; Yao, Y.; Guo, Z.; Cui, J.; Ni, Z.; Ge, C.; Chua, T.S.; Liu, Z.; Sun, M.; Huang, G. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. *arXiv preprint arXiv:2403.11703* **2024**.
23. Kafle, K.; Price, B.; Cohen, S.; Kanan, C. Dvqa: Understanding data visualizations via question answering. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5648–5656.
24. Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T.Y. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558* **2020**.
25. Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; Xiao, J. 3d shapenets: A deep representation for volumetric shapes. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1912–1920.
26. Ye, J.; Hu, A.; Xu, H.; Ye, Q.; Yan, M.; Xu, G.; Li, C.; Tian, J.; Qian, Q.; Zhang, J.; et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. In Proceedings of the EMNLP, 2023.
27. Jiang, A.Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D.S.; Casas, D.d.l.; Hanna, E.B.; Bressand, F.; et al. Mixtral of experts. *arXiv:2401.04088* **2024**.
28. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *NeurIPS* **2017**.
29. Gadre, S.Y.; Ilharco, G.; Fang, A.; Hayase, J.; Smyrnis, G.; Nguyen, T.; Marten, R.; Wortsman, M.; Ghosh, D.; Zhang, J.; et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems* **2024**, *36*.
30. Johnson, A.E.W.; Pollard, T.J.; Greenbaum, N.R.; Lungren, M.P.; ying Deng, C.; Peng, Y.; Lu, Z.; Mark, R.G.; Berkowitz, S.J.; Horng, S. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs, 2019, [arXiv:cs.CV/1901.07042].
31. Sun, Z.; Shen, S.; Cao, S.; Liu, H.; Li, C.; Shen, Y.; Gan, C.; Gui, L.Y.; Wang, Y.X.; Yang, Y.; et al. Aligning large multimodal models with factually augmented rlhf. *arXiv:2309.14525* **2023**.
32. Han, J.; Zhang, R.; Shao, W.; Gao, P.; Xu, P.; Xiao, H.; Zhang, K.; Liu, C.; Wen, S.; Guo, Z.; et al. Imagebind-llm: Multi-modality instruction tuning. *arXiv:2309.03905* **2023**.
33. Yang, Z.; Gan, Z.; Wang, J.; Hu, X.; Lu, Y.; Liu, Z.; Wang, L. An empirical study of gpt-3 for few-shot knowledge-based vqa. In Proceedings of the AAAI, 2022.

 doi:10.20944/preprints202512.1986.v1

34. Lai, Z.; Zhang, H.; Wu, W.; Bai, H.; Timofeev, A.; Du, X.; Gan, Z.; Shan, J.; Chuah, C.N.; Yang, Y.; et al. From scarcity to efficiency: Improving clip training via visual-enriched captions. *arXiv preprint arXiv:2310.07699* **2023**.
35. Xia, M.; Malladi, S.; Gururangan, S.; Arora, S.; Chen, D. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333* **2024**.
36. Nakano, R.; Hilton, J.; Balaji, S.; Wu, J.; Ouyang, L.; Kim, C.; Hesse, C.; Jain, S.; Kosaraju, V.; Saunders, W.; et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv:2112.09332* **2021**.
37. Hong, W.; Wang, W.; Lv, Q.; Xu, J.; Yu, W.; Ji, J.; Wang, Y.; Wang, Z.; Dong, Y.; Ding, M.; et al. Cogagent: A visual language model for gui agents. *arXiv:2312.08914* **2023**.
38. Panagopoulou, A.; Xue, L.; Yu, N.; Li, J.; Li, D.; Joty, S.; Xu, R.; Savarese, S.; Xiong, C.; Niebles, J.C. X-InstructBLIP: A Framework for aligning X-Modal instruction-aware representations to LLMs and Emergent Cross-modal Reasoning. *arXiv preprint arXiv:2311.18799* **2023**.
39. Yu, Q.; Sun, Q.; Zhang, X.; Cui, Y.; Zhang, F.; Wang, X.; Liu, J. Capsfusion: Rethinking image-text data at scale. *arXiv preprint arXiv:2310.20550* **2023**.
40. Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL* **2014**.
41. Suárez, P.J.O.; Sagot, B.; Romary, L. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In Proceedings of the 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7). Leibniz-Institut für Deutsche Sprache, 2019.
42. Liu, Z.; Li, S.; Luo, Y.; Fei, H.; Cao, Y.; Kawaguchi, K.; Wang, X.; Chua, T.S. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. *arXiv preprint arXiv:2310.12798* **2023**.
43. Wu, P.; Xie, S. V*: Guided Visual Search as a Core Mechanism in Multimodal LLMs. *arXiv:2312.14135* **2023**.
44. Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; Christiano, P.F. Learning to summarize with human feedback. *NeurIPS* **2020**.
45. Zhou, T.; Chen, Y.; Cao, P.; Liu, K.; Zhao, J.; Liu, S. Oasis: Data curation and assessment system for pretraining of large language models. *arXiv preprint arXiv:2311.12537* **2023**.
46. Wang, Z.; Zhong, W.; Wang, Y.; Zhu, Q.; Mi, F.; Wang, B.; Shang, L.; Jiang, X.; Liu, Q. Data management for large language models: A survey. *arXiv preprint arXiv:2312.01700* **2023**.
47. Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J.M.; Parikh, D.; Batra, D. Visual dialog. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 326–335.
48. Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; Rohrbach, M. Towards vqa models that can read. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 8317–8326.
49. Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K.V.; Joulin, A.; Misra, I. Imagebind: One embedding space to bind them all. In Proceedings of the CVPR, 2023.
50. Gu, J.; Meng, X.; Lu, G.; Hou, L.; Minzhe, N.; Liang, X.; Yao, L.; Huang, R.; Zhang, W.; Jiang, X.; et al. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems* **2022**, *35*, 26418–26431.
51. Shu, Y.; Dong, S.; Chen, G.; Huang, W.; Zhang, R.; Shi, D.; Xiang, Q.; Shi, Y. Llasm: Large language and speech model. *arXiv preprint arXiv:2308.15930* **2023**.
52. Gao, J.; Sun, C.; Yang, Z.; Nevatia, R. Tall: Temporal activity localization via language query. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2017, pp. 5267–5275.
53. Longpre, S.; Yauney, G.; Reif, E.; Lee, K.; Roberts, A.; Zoph, B.; Zhou, D.; Wei, J.; Robinson, K.; Mimno, D.; et al. A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity. *arXiv preprint arXiv:2305.13169* **2023**.
54. Kembhavi, A.; Salvato, M.; Kolve, E.; Seo, M.; Hajishirzi, H.; Farhadi, A. A diagram is worth a dozen images. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. Springer, 2016, pp. 235–251.
55. Pham, V.T.; Le, T.L.; Tran, T.H.; Nguyen, T.P. Hand detection and segmentation using multimodal information from Kinect. In Proceedings of the 2020 International Conference on Multimedia Analysis and Pattern Recognition (MAPR). IEEE, 2020, pp. 1–6.
56. Rogers, V.; Meara, P.; Barnett-Legh, T.; Curry, C.; Davie, E. Examining the LLAMA aptitude tests. *Journal of the European Second Language Association* **2017**, *1*, 49–60.

 doi:10.20944/preprints202512.1986.v1

57. Kombrink, S.; Mikolov, T.; Karafiát, M.; Burget, L. Recurrent Neural Network Based Language Modeling in Meeting Recognition. In Proceedings of the Interspeech, 2011, Vol. 11, pp. 2877–2880.
58. Askell, A.; Bai, Y.; Chen, A.; Drain, D.; Ganguli, D.; Henighan, T.; Jones, A.; Joseph, N.; Mann, B.; DasSarma, N.; et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861* **2021**.
59. Chen, C.; Qin, R.; Luo, F.; Mi, X.; Li, P.; Sun, M.; Liu, Y. Position-enhanced visual instruction tuning for multimodal large language models. *arXiv preprint arXiv:2308.13437* **2023**.
60. Shen, S.; Hou, L.; Zhou, Y.; Du, N.; Longpre, S.; Wei, J.; Chung, H.W.; Zoph, B.; Fedus, W.; Chen, X.; et al. Mixture-of-experts meets instruction tuning: A winning combination for large language models. *arXiv:2305.14705* **2023**.
61. Piczak, K.J. ESC: Dataset for environmental sound classification. In Proceedings of the Proceedings of the 23rd ACM international conference on Multimedia, 2015, pp. 1015–1018.
62. Roberts, J.; Lüddecke, T.; Sheikh, R.; Han, K.; Albanie, S. Charting new territories: Exploring the geographic and geospatial capabilities of multimodal llms. *arXiv preprint arXiv:2311.14656* **2023**.
63. Hu, A.; Xu, H.; Ye, J.; Yan, M.; Zhang, L.; Zhang, B.; Li, C.; Zhang, J.; Jin, Q.; Huang, F.; et al. mPLUG-DocOwl 1.5: Unified Structure Learning for OCR-free Document Understanding. *arXiv:2403.12895* **2024**.
64. Byeon, M.; Park, B.; Kim, H.; Lee, S.; Baek, W.; Kim, S. COYO-700M: Image-Text Pair Dataset. https://github.com/kakaobrain/coyo-dataset, 2022.
65. Zhao, H.; Andriushchenko, M.; Croce, F.; Flammarion, N. Long is more for alignment: A simple but tough-to-beat baseline for instruction fine-tuning. *arXiv preprint arXiv:2402.04833* **2024**.
66. Liu, Z.; Sun, M.; Zhou, T.; Huang, G.; Darrell, T. Rethinking the value of network pruning. *arXiv preprint arXiv:1810.05270* **2018**.
67. Feng, J.; Sun, Q.; Xu, C.; Zhao, P.; Yang, Y.; Tao, C.; Zhao, D.; Lin, Q. MMDialog: A Large-scale Multi-turn Dialogue Dataset Towards Multi-modal Open-domain Conversation. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 7348–7363.
68. Zhang, H.; Li, X.; Bing, L. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858* **2023**.
69. Ben Abacha, A.; Demner-Fushman, D. A question-entailment approach to question answering. *BMC bioinformatics* **2019**, *20*, 1–23.
70. Kung, P.N.; Yin, F.; Wu, D.; Chang, K.W.; Peng, N. Active Instruction Tuning: Improving Cross-Task Generalization by Training on Prompt Sensitive Tasks. In Proceedings of the The 2023 Conference on Empirical Methods in Natural Language Processing, 2023.
71. Penedo, G.; Malartic, Q.; Hesslow, D.; Cojocaru, R.; Cappelli, A.; Alobeidli, H.; Pannier, B.; Almazrouei, E.; Launay, J. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116* **2023**.
72. Honda, S.; Shi, S.; Ueda, H.R. Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery. *arXiv preprint arXiv:1911.04738* **2019**.
73. Han, X.; Zhang, Z.; Ding, N.; Gu, Y.; Liu, X.; Huo, Y.; Qiu, J.; Yao, Y.; Zhang, A.; Zhang, L.; et al. Pre-trained models: Past, present and future. *AI Open* **2021**, *2*, 225–250.
74. Rasheed, H.; Maaz, M.; Shaji, S.; Shaker, A.; Khan, S.; Cholakkal, H.; Anwer, R.M.; Xing, E.; Yang, M.H.; Khan, F.S. Glamm: Pixel grounding large multimodal model. *arXiv:2311.03356*.
75. Dettmers, T.; Pagnoni, A.; Holtzman, A.; Zettlemoyer, L. Qlora: Efficient finetuning of quantized llms. *arXiv:2305.14314* **2023**.
76. Li, J.; Liu, Y.; Fan, W.; Wei, X.Y.; Liu, H.; Tang, J.; Li, Q. Empowering Molecule Discovery for Molecule-Caption Translation with Large Language Models: A ChatGPT Perspective. *arXiv preprint arXiv:2306.06615* **2023**.
77. Wang, D.; Shang, Y. A new active labeling method for deep learning. In Proceedings of the 2014 International joint conference on neural networks (IJCNN). IEEE, 2014, pp. 112–119.
78. Zhao, Q.; Gao, X.; Li, J.; Luo, L. Optimization algorithm for point cloud quality enhancement based on statistical filtering. *Journal of Sensors* **2021**, *2021*, 1–10.
79. Shen, Y.; Song, K.; Tan, X.; Li, D.; Lu, W.; Zhuang, Y. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv:2303.17580* **2023**.
80. Yuan, Y.; Li, W.; Liu, J.; Tang, D.; Luo, X.; Qin, C.; Zhang, L.; Zhu, J. Osprey: Pixel Understanding with Visual Instruction Tuning. *arXiv:2312.10032*.

 doi:10.20944/preprints202512.1986.v1

81. Chen, C.; Liu, M.; Codella, N.; Li, Y.; Yuan, L.; Gurari, D. Fully authentic visual question answering dataset from online communities. *arXiv preprint arXiv:2311.15562* **2023**.
82. Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; Anderson, P. Nocaps: Novel object captioning at scale. In Proceedings of the ICCV, 2019.
83. Jiang, J.; Shu, Y.; Wang, J.; Long, M. Transferability in deep learning: A survey. *arXiv preprint arXiv:2201.05867* **2022**.
84. Friedman, D.; Dieng, A.B. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research* **2023**.
85. Lyu, C.; Wu, M.; Wang, L.; Huang, X.; Liu, B.; Du, Z.; Shi, S.; Tu, Z. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093* **2023**.
86. He, P.; Liu, X.; Gao, J.; Chen, W. DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. In Proceedings of the International Conference on Learning Representations, 2020.
87. Liu, Y.; Cao, J.; Liu, C.; Ding, K.; Jin, L. Datasets for Large Language Models: A Comprehensive survey, 2024, [arXiv:cs.CL/2402.18041].
88. Schuhmann, C.; Köpf, A.; Vencu, R.; Coombes, T.; Beaumont, R. Laion coco: 600m synthetic captions from laion2b-en. *https://laion.ai/blog/laion-coco/* **2022**.
89. Zhao, Z.; Guo, L.; Yue, T.; Chen, S.; Shao, S.; Zhu, X.; Yuan, Z.; Liu, J. ChatBridge: Bridging Modalities with Large Language Model as a Language Catalyst. *arXiv:2305.16103* **2023**.
90. Su, Y.; Lan, T.; Li, H.; Xu, J.; Wang, Y.; Cai, D. PandaGPT: One Model To Instruction-Follow Them All. *arXiv:2305.16355* **2023**.
91. Xue, L.; Yu, N.; Zhang, S.; Li, J.; Martín-Martín, R.; Wu, J.; Xiong, C.; Xu, R.; Niebles, J.C.; Savarese, S. ULIP-2: Towards Scalable Multimodal Pre-training for 3D Understanding, 2023, [arXiv:cs.CV/2305.08275].
92. Xu, J.; Mei, T.; Yao, T.; Rui, Y. Msr-vtt: A large video description dataset for bridging video and language. In Proceedings of the CVPR, 2016.
93. Hernandez, D.; Brown, T.; Conerly, T.; DasSarma, N.; Drain, D.; El-Showk, S.; Elhage, N.; Hatfield-Dodds, Z.; Henighan, T.; Hume, T.; et al. Scaling laws and interpretability of learning from repeated data. *arXiv preprint arXiv:2205.10487* **2022**.
94. Zhang, R.; Hu, X.; Li, B.; Huang, S.; Deng, H.; Qiao, Y.; Gao, P.; Li, H. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In Proceedings of the CVPR, 2023.
95. Ye, J.; Liu, P.; Sun, T.; Zhou, Y.; Zhan, J.; Qiu, X. Data Mixing Laws: Optimizing Data Mixtures by Predicting Language Modeling Performance. *arXiv preprint arXiv:2403.16952* **2024**.
96. Xu, Z.; Feng, C.; Shao, R.; Ashby, T.; Shen, Y.; Jin, D.; Cheng, Y.; Wang, Q.; Huang, L. Vision-Flan: Scaling Human-Labeled Tasks in Visual Instruction Tuning. *arXiv:2402.11690* **2024**.
97. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International conference on machine learning. PMLR, 2021, pp. 8748–8763.
98. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the ECCV, 2014.
99. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv:2301.12597* **2023**.
100. Goyal, S.; Choudhury, A.R.; Raje, S.; Chakaravarthy, V.; Sabharwal, Y.; Verma, A. Power-bert: Accelerating bert inference via progressive word-vector elimination. In Proceedings of the International Conference on Machine Learning. PMLR, 2020, pp. 3690–3699.
101. Li, B.; Zhang, Y.; Chen, L.; Wang, J.; Yang, J.; Liu, Z. Otter: A multi-modal model with in-context instruction tuning. *arXiv:2305.03726* **2023**.
102. Fan, S.; Pagliardini, M.; Jaggi, M. Doge: Domain reweighting with generalization estimation. *arXiv preprint arXiv:2310.15393* **2023**.
103. Wang, L.; Yang, N.; Huang, X.; Jiao, B.; Yang, L.; Jiang, D.; Majumder, R.; Wei, F. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533* **2022**.
104. Li, L.; Yin, Y.; Li, S.; Chen, L.; Wang, P.; Ren, S.; Li, M.; Yang, Y.; Xu, J.; Sun, X.; et al. M^3IT: A Large-Scale Dataset towards Multi-Modal Multilingual Instruction Tuning. *arXiv preprint arXiv:2306.04387* **2023**.
105. Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114* **2021**.

 doi:10.20944/preprints202512.1986.v1

106. Zhu, D.; Chen, J.; Shen, X.; Li, X.; Elhoseiny, M. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models, 2023, [arXiv:cs.CV/2304.10592].
107. Chen, D.; Dolan, W.B. Collecting highly parallel data for paraphrase evaluation. In Proceedings of the Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, 2011, pp. 190–200.