Comprehensive Survey on Multimodal Large Language Models: Advances, Challenges, and Future Directions

SurveyForge

Abstract— The landscape of artificial intelligence is rapidly evolving with the advent of Multimodal Large Language Models (MLLMs), which integrate diverse data types such as text, images, and audio to enhance machine comprehension and interaction capabilities. This comprehensive survey explores the historical progression, design paradigms, and core functions of MLLMs, emphasizing innovative architectural strategies like transformers and mixture of experts for efficient modality synthesis. Despite significant advances, challenges persist regarding the seamless fusion of modalities and ethical concerns over bias and privacy, necessitating continued refinement and ethical guidelines to ensure responsible deployment. Key trends indicate a promising trajectory towards scalable and efficient models that address real-world complexity and dynamic contexts through adaptive training methodologies and robust modality fusion strategies. The implications of these developments are vast, extending across domains such as healthcare, autonomous systems, and assistive technologies, where MLLMs hold transformative potential. Ultimately, the future success of MLLMs hinges on interdisciplinary collaboration and robust ethical frameworks that guide their integration into society, enhancing their utility while safeguarding ethical standards.

Index Terms—Multimodal integration methods, Cross-modal interaction, MLLM applications

1 Introduction

In recent years, the field of artificial intelligence (AI) has undergone significant transformations, largely attributed to the advancements in language models. Among these, Multimodal Large Language Models (MLLMs) stand out as a critical innovation, offering capabilities that extend beyond the confines of unimodal data processing to a more integrated and comprehensive comprehension of the world [1]. This subsection delves into the historical evolution, core significance, and transformative potential of MLLMs across varied applications, reflecting on how these models have revolutionized AI and paved the way for future innovations.

Initially, the concept of multimodality emphasized combining disparate data types such as text, images, and audio, among others, into a coherent system that could better mimic human-like understanding and reasoning [2]. Early efforts in this domain faced challenges related to data alignment and the synergistic fusion of modalities, which hindered effective cross-modal interactions [3]. However, the advent of sophisticated architectures like transformers has unlocked unprecedented potential in this area. The adoption of these architectures facilitates the seamless integration of different modalities, offering enhanced dimensionality and interaction capabilities that were previously unattainable [4].

The historical evolution of MLLMs can be traced through various developmental phases characterized by increasing model complexity and capacity for cross-modal reasoning [5]. Initially, the focus was on creating foundational models capable of handling single modalities. As research progressed, there was a significant shift towards developing models that could analyze and synthesize information from multiple sources simultaneously. This evolution was marked by seminal works that introduced frameworks for modality collaboration and integration [6]. These advancements have enabled MLLMs to excel in tasks that require holistic data interpretation, from visual question answering to complex cognitive tasks such as multimodal sentiment analysis and contextual understanding [7].

The significance of MLLMs in artificial intelligence is multifaceted. At its core, the integration of multiple data modalities within a unified framework allows for a more nuanced understanding of context, leading to better performance in multimodal tasks such as image captioning, speech recognition, and autonomous navigation [8]. For instance, models that leverage textual and visual data in tandem have demonstrated the ability to perform complex reasoning tasks, such as interpreting and generating visual content based on textual prompts [9]. This capability not only enhances the accuracy of AI systems but also broadens the scope of applications to domains that require high precision and contextual awareness, such as healthcare and autonomous systems [10].

However, the transition to multimodal frameworks presents several challenges and trade-offs, particularly concerning the integration complexity and computational demands of these models. Notably, while the incorporation of multi-modal data enhances model capabilities, it also introduces vulnerabilities like a decrease in performance when one of the modalities is missing or incomplete, as explored in recent studies [11]. There is a need for robust design frameworks that can accommodate multimodal inputs seamlessly while maintaining efficiency and scalability [12]. Furthermore, developing standardized metrics and benchmarks for evaluating the performance of MLLMs

across diverse modalities remains an ongoing challenge, necessitating more comprehensive approaches to assessment [13].

Looking forward, the future directions in the realm of MLLMs are poised towards enhancing scalability and interaction mechanisms, ensuring these models can efficiently manage larger datasets and more complex tasks [14]. Emerging trends suggest a promising trajectory for the integration of MLLMs with knowledge graphs and vector databases, potentially alleviating issues like hallucinations and knowledge limitations inherent in current models [15]. Ultimately, ongoing research must focus on refining the modality integration frameworks and addressing the ethical considerations surrounding MLLM deployment, particularly concerning bias and data privacy [10].

In summary, Multimodal Large Language Models represent a pivotal advancement in AI, embodying the convergence of diverse modalities to achieve more comprehensive intelligence. As these models continue to evolve, they hold the potential to redefine interactions across countless applications, driving progress towards truly intelligent systems capable of holistic reasoning and decision-making.

2 Core Architectures and Design Principles

2.1 Architectural Paradigms

Multimodal Large Language Models (MLLMs) are designed to process diverse data types, such as text, images, and audio, within a unified framework. This subsection explores the architectural paradigms that underpin the design of these models, focusing on the mechanisms by which they integrate and process multiple modalities. As the landscape of MLLMs evolves, diverse architectural choices have emerged, each with unique strengths and limitations.

One prominent architectural paradigm in MLLM design is the transformer-based architecture. This approach extends the original transformer model, renowned for its ability to handle sequence data, by embedding multiple modalities within the same framework. The core idea is to encode different modalities into a unified representation space, allowing cross-modal interactions via self-attention layers [4]. Innovations such as Vision-Language Models (VLMs), such as CLIP, exploit this adaptability by employing joint embeddings to align textual and visual information, facilitating coherent integration across modalities [16]. However, the reliance on transformers introduces high computational costs, as self-attention mechanisms require quadratic complexity, posing a challenge for scaling [14].

To address the computational burden while enhancing scalability, the Mixture of Experts (MoE) framework presents a viable alternative. The MoE architecture dynamically allocates a subset of model parameters per input sample, effectively optimizing resource utilization. This sparsification technique allows for efficient handling of multiple modalities by allocating different expert subnetworks to process specific data types [7]. By activating only a portion of the model, MoE minimizes computation, making it feasible to train larger models and enabling real-time multimodal interactions. However, managing the routing mechanisms and expert assignments remains a complex task, potentially leading to challenges in maintaining the model's coherence and ensuring robust performance across tasks [12].

Beyond transformers and MoE, state space models offer another architectural strategy, particularly in applications requiring the handling of long sequences and complex temporal dependencies. State space models define a latent space where the temporal evolution of data is modeled explicitly, allowing for smoother and more interpretable integration across modalities. These models excel in scenarios where maintaining context over longer horizons is crucial, such as in dynamic scene understanding or in applications involving sequential decision-making processes. However, state space models often require more intricate tuning to balance the trade-off between expressiveness and computational efficiency, posing challenges when integrated with high-dimensional multimodal data [5].

Comparatively, transformer-based architectures provide superior flexibility in handling diverse data formats, facilitating seamless integration by embedding different modalities into the same representational space. Nonetheless, this comes at the cost of increased computational demand, necessitating strategies for optimization or simplification such as pruning or distillation [17]. In contrast, the MoE and state space models emphasize efficiency, offering scalability and reduced computational overhead, but often at the expense of architectural and implementation complexity [18].

Emerging trends indicate a shift towards hybrid models that integrate the strengths of different architectural paradigms. For instance, efforts to combine transformers with MoE architectures aim to leverage the flexibility of self-attention mechanisms while capitalizing on the efficiency of sparse activations [19]. Such advancements pave the way for the development of MLLMs that balance computational efficiency and representational power, enabling broader applicability across varied domains and tasks.

In conclusion, the architectural paradigms underlying MLLMs are diverse, each designed to address specific challenges in multimodal integration and processing. The continued evolution of these paradigms hinges on the development of hybrid models that synergistically integrate multiple architectural strategies. Future research should focus on optimizing these architectures for scalability, robustness, and efficiency, ensuring that MLLMs can seamlessly adapt to the growing complexity and diversity of multimodal data landscapes.

2.2 Multimodal Integration Techniques

In the realm of Multimodal Large Language Models (MLLMs), effectively integrating diverse modalities such as text, vision, and audio into a cohesive framework is critical for enhancing the model's capacity to comprehend and generate meaningful outputs from complex inputs. This subsection explores essential multimodal integration techniques, emphasizing strategies that ensure efficient and effective cross-modal interactions.

Integrating multiple modalities within a unified model is a multifaceted task that demands sophisticated architecture. Techniques for multimodal integration typically involve three core components: semantic alignment, joint embedding spaces, and dynamic modality fusion. These compo-

nents function synergistically to enable MLLMs to process information by maximizing each modality's strengths.

Alignment strategies are pivotal in ensuring that distinct modalities are semantically coherent and effectively contribute to a unified representation. A common approach uses cross-attention mechanisms, which dynamically highlight relevant features in one modality based on information from another. The Multimodal Transformer (MulT) employs directional pairwise cross-modal attention to efficiently handle unaligned data streams without explicit pre-alignment [20]. This approach addresses the inherent differences in modality sampling rates and their nonlinear dependencies, although it comes with increased model complexity.

Creating joint embedding spaces is fundamental for unifying multimodal inputs into a common semantic space. Such embeddings facilitate shared understanding and seamless integration of modality-specific information into a coherent model input. Various methods have been proposed to construct these spaces. For instance, techniques involving Latent Space Models align visual and textual data within a shared latent space using novel constraints, enhancing predictions through concurrent descriptions across multiple modalities [21]. Variational Mixture-of-Experts Autoencoders further improve learning across modalities by decomposing inputs into shared and private subspaces, promoting coherent cross-generation [22]. While joint embedding techniques enable effective multimodal modeling, selecting appropriate dimensionality and ensuring scalability remain challenging.

Dynamic modality fusion focuses on adaptively combining inputs from various modalities based on contextual relevance and task requirements. Techniques like the Mixture of Experts (MoE) leverage modular architectures, activating specific components based on input modality, facilitating scalable integration and reduced computational costs [23]. This modular approach allows dynamic allocation of resources, flexibly integrating modalities as required by the task. However, managing expert routing and resolving data sparsity issues remain significant challenges.

Integration approaches provide robust solutions to leveraging the strengths of multiple modalities. Cross-attention mechanisms elegantly handle misaligned and diverse input modalities but require careful design to manage computational overhead. Joint embedding spaces offer a unified representation, promoting effective cross-modal interactions, though scaling efficiently to include novel modalities without loss of information is complex. Finally, dynamic fusion methodologies afford efficiency and flexibility but necessitate sophisticated routing algorithms to prevent redundancy and ensure task-specific performance gains.

As multimodal applications grow increasingly demanding, integration techniques are evolving to accommodate new requirements. Future advancements may focus on adaptive fusion strategies that enable fine-grained modality selection per context, aligning with emergent concepts of Artificial General Intelligence [24]. Enhancing robustness to missing modalities and optimizing joint embeddings for large-scale, diverse data types are key research areas. Moreover, fostering collaboration between modalities through shared auxiliary tasks and developing cross-modal reasoning mechanisms will enhance holistic understanding [25].

In conclusion, while significant strides have been made in developing multimodal integration techniques within MLLMs, research must continue to overcome challenges related to scalability, efficiency, and robustness. The dynamic interplay between theoretical frameworks and practical applications promises to spur further innovations, shaping the trajectory of multimodal AI systems.

2.3 Processing and Interaction Mechanisms

In the realm of Multimodal Large Language Models (MLLMs), the processing and interaction mechanisms are pivotal to achieving effective integration and interpretation of diverse modalities such as text, images, and audio. This subsection delves into the tools and techniques employed to facilitate seamless multimodal input processing and interaction, examining both foundational and emerging approaches.

A critical component of multimodal processing is the establishment of connectors that link disparate modalities, effectively bridging distinct representational spaces. Vision-language connectors, for instance, play a crucial role in linking visual encoders to language models, thereby optimizing the information flow between visual and linguistic representations [26]. Such connectors utilize cross-attention mechanisms to align and integrate visual and textual information without necessitating explicit data alignment, providing flexibility in handling diverse data configurations [20].

Tokenization techniques form another essential processing mechanism within MLLMs, transforming multimodal inputs into sequences of tokens that are universally interpretable by language models. Advanced tokenization strategies ensure that the semantic essence of each modality is preserved during the conversion, facilitating multimodal fusion at subsequent stages [20]. However, a significant challenge remains in ensuring that tokenization does not discard modality-specific nuances, which are critical for nuanced interaction and information retrieval.

Decoder designs in MLLMs have evolved significantly to accommodate the diverse outputs generated across multiple modalities. Recent innovations emphasize the need for flexible decoder architectures that can adapt to various output forms, ensuring coherence and relevance in response generation [25]. These decoders often employ modality-aware mechanisms to customize the output based on context-specific needs, leveraging insights from each modality to enrich the generated responses.

Multimodal attention mechanisms stand out as a sophisticated processing tool that addresses the complexity of interactions between different modalities. By leveraging attention models that prioritize certain modalities or features within a modality, such mechanisms can finely tune the focus of the MLLM, thereby enhancing the interpretative and generative capabilities of the models [27]. This includes the capacity to attend to temporally or spatially specific features, facilitating context-aware multimodal interactions [27].

Despite these advancements, processing and interaction mechanisms in MLLMs still face several challenges. The modulation of interactions to avoid biases stemming

from dominant modalities remains an open issue [28]. Furthermore, the sensitivity of these models to missing or noisy data highlights an ongoing need for robust processing strategies that can accommodate incomplete or imperfect multimodal inputs without significant degradation in performance [11].

Looking forward, integrating self-supervised learning paradigms presents a promising avenue for enhancing multimodal model resilience and adaptability. These techniques propose engaging models with unlabeled data to learn representations that are both modality-agnostic and capable of adaption to unseen scenarios [29]. Additionally, emerging practices in multimodal in-context learning (M-ICL) demonstrate potential in optimizing the interaction capabilities of MLLMs through improved contextual understanding and reasoning across diverse inputs [30].

In summary, processing and interaction mechanisms in Multimodal Large Language Models are characterized by a dynamic interplay between cutting-edge techniques in attention, tokenization, and strategic decoder design. While significant progress has been made, the field is ripe for continued innovation, particularly in developing mechanisms that are both adaptive and robust to the complexities and imperfections of real-world multimodal data. As these models further integrate and expand across new modalities, maintaining clarity in communication pathways and processing strategies will be essential in achieving the transformative potential that MLLMs hold. Future research should focus on refining these mechanisms to optimize both the efficiency and accuracy of multimodal processing, thereby broadening the applicability and effectiveness of these models across diverse disciplines.

2.4 Computational Efficiency and Scalability

Focusing on computational efficiency and scalability is crucial in the development of Multimodal Large Language Models (MLLMs). These attributes are vital to ensuring that models can manage the vast and diverse datasets inherent in multimodal environments while keeping computational costs sustainable and allowing deployment across various platforms, including resource-constrained devices. This subsection comprehensively explores architectural strategies and technical methodologies aimed at boosting the efficiency and scalability of MLLMs. By examining a range of efficient pre-training techniques, applications of low-rank and sparse mechanisms, and deployment strategies, we provide a clear perspective on the current state and potential future directions in this domain.

Efficient pre-training techniques have surfaced as key strategies to reduce the computational expenses associated with MLLMs. One promising approach utilizes data augmentation and curriculum learning strategies, allowing models to learn more effectively by introducing varied and increasingly challenging training data inputs. Data augmentation enriches training datasets without necessitating extensive data acquisition, thereby enabling robust learning processes. Similarly, curriculum learning structures the training process by progressively increasing data complexity, which enhances model learning efficiency [2].

Another avenue towards achieving computational efficiency and scalability is the deployment of low-rank approximations and sparse attention mechanisms. These approaches are instrumental in reducing the number of parameters and computational requirements for model operation. By decomposing weight matrices into low-rank structures, computational demands are decreased without notably affecting model performance. Sparse attention mechanisms, which focus attention on the most pertinent parts of the input data, address the quadratic complexity typically linked to conventional attention models [20], [31]. This not only bolsters computational efficiency but also allows for faster inference times.

The challenge of resource-constrained deployment further necessitates specialized strategies. Techniques like model compression, quantization, and knowledge distillation are employed to adapt large models for deployment on edge devices. Compression reduces model size by eliminating redundancies, while quantization decreases precision to lower computational overhead, maintaining acceptable accuracy levels. Knowledge distillation involves training a smaller model to mirror the outputs of a larger model, enabling the practical deployment of computationally intensive models without significant performance loss [24], [32].

Emerging trends showcase the integration of mixture of experts (MoE) frameworks, which dynamically allocate computations across different model parameter subsets based on input data. This selective activation of model segments facilitates efficient scaling without a proportional rise in computational resources. The design architecture of Vision-Language Models (VLMs), leveraging a mixture of vision encoders, exemplifies this approach, granting enhanced performance at reduced computational costs [33].

However, numerous challenges persist. Balancing model complexity and computational demand remains a pivotal concern, especially as the scale and complexity of multimodal tasks increase. While low-rank and sparse techniques mitigate computational needs, they often require meticulous tuning to achieve an optimal balance of efficiency and performance. Furthermore, ensuring the robustness of deployment strategies in varied computational environments remains a significant challenge, as evidenced by differing performances across frameworks [5].

In conclusion, as MLLMs continue to advance, addressing the dual challenges of computational efficiency and scalability necessitates innovative architectural and algorithmic solutions. Future research directions may focus on refining adaptive techniques like curriculum learning to dynamically match model capabilities, further exploring MoE frameworks for optimal resource allocation, and developing more robust benchmarks for evaluating efficiency and scalability in practical scenarios. This ongoing evolution promises to broaden the accessibility and applicability of MLLMs, ensuring their sustainable growth and impact in various multimodal applications.

2.5 Emerging Design Innovations

The domain of Multimodal Large Language Models (MLLMs) is rapidly advancing with several innovative design approaches that promise to enhance their capabilities and applications. This subsection delves into recent ad-

vancements and emerging trends that significantly transform the architectural and functional landscape of MLLMs.

One of the prevailing trends in emerging design innovations is the increased emphasis on interactive and conversational agents. These systems are increasingly being integrated with real-time interaction capabilities that allow for dynamic dialogue management and multimodal user interfaces. Models like InstructionGPT-4 illustrate this trend by demonstrating how fine-tuning with high-quality instruction-following data can significantly enhance dialogue capabilities [34]. Furthermore, efforts to make models more responsive and versatile in dialogue settings, such as those explored in Composable Diffusion, highlight the potential of these innovations to facilitate simultaneous, multimodal interactions [35].

A key strength of these design innovations is their focus on robustness and adaptability. In real-world applications, MLLMs must be resilient to noisy inputs and environmental variability. Approaches that focus on improving the robustness of MLLMs often leverage Sparse Mixture of Experts (MoE) frameworks, which allow models to efficiently manage complexity by distributing tasks among specialized submodels [36]. This not only enhances performance but also ensures that the models remain adaptable across various scenarios and tasks. Incorporating MoE can help offload computational demands and improve reliability, which is critical in deploying MLLMs in resource-constrained environments like edge devices [37].

Moreover, emerging design innovations are capitalizing on cross-disciplinary applications, with particular emphasis on fields such as healthcare, robotics, and autonomous systems. The application of MLLMs in healthcare, as seen in LLaVA-Rad, demonstrates their potential to interpret complex biomedical data, offering tools for enhanced diagnostics and patient care [38]. Additionally, projects focusing on augmented reality and interactive systems are leveraging the advanced capabilities of MLLMs for more immersive user experiences [39]. The integration of MLLMs into robotics and autonomous systems also offers promising prospects, enabling more intuitive human-machine interactions and decision-making processes [40].

Despite these advancements, several challenges persist. Ensuring robustness against adversarial inputs remains an area of active research, with approaches like RESSA focusing on efficient adaptation to cross-modal interactions [41]. Moreover, as models become more computationally intensive, efforts such as LoRA-FA highlight the need for memory-efficient design innovations that maintain performance without overly taxing system resources [42]. Additionally, ethical concerns regarding bias and privacy are becoming more pronounced as MLLMs grow in capability and application scope. Strategies addressing these concerns are crucial to the responsible deployment of such systems [5].

Looking ahead, the field is poised for further breakthroughs in scalability and efficiency, as evidenced by progress in linear transformer variants like HyperAttention, which promise near-linear time attention mechanisms [43]. These advancements suggest a future where MLLMs not only mimic but surpass human cognitive abilities in synthesizing and reasoning across diverse data streams.

In conclusion, the innovations emerging in the design of MLLMs underscore a vibrant research ecosystem that continually seeks to push the boundaries of what is possible. By addressing the challenges of robustness, adaptability, and ethical deployment, the field is laying the groundwork for MLLMs to become integral across various domains. The progress achieved thus far offers a glimpse into a future where these models can operate as sophisticated, autonomous agents in a plethora of applications, driving significant advancements in artificial intelligence.

3 Training Paradigms and Techniques

3.1 Foundational Training Strategies

The development of Multimodal Large Language Models (MLLMs) is fundamentally driven by training strategies that integrate and harmonize diverse data types across multiple modalities, such as text, image, and audio. This subsection explores the foundational strategies employed in training MLLMs, focusing on methods that leverage vast, heterogeneous datasets and iterative processes to enhance model capacity and efficacy.

At the core of MLLM training is the aggregation and preprocessing of diversified datasets. The ability to harmonize data across different modalities is critical for creating robust and adaptable models. Techniques such as modality alignment and joint embedding strategies are crucial in this stage. According to "Multimodal Deep Learning" [4], achieving effective integration across modalities often begins with preprocessing steps that align datasets to a shared semantic space. The approach ensures that the multimodal inputs can be interpreted cohesively by the language model.

Initial training phases typically involve establishing a baseline model that synthesizes foundational understanding across different modalities. This process often leverages large pre-trained language models extended with additional modality-specific encoders as seen in models like "MM-LLMs" [19]. These encoders adaptively map multimodal data into a unified space, which is essential for the models to generate coherent outputs across varied inputs. During the initial stages, models employ strategies such as masked token prediction or autoencoding to learn generalized features before fine-tuning on specific tasks.

A comparative analysis of foundational strategies reveals diverse approaches in handling modality fusion and integration. While some models utilize early fusion techniques to merge multimodal data in initial layers, others adopt late fusion methods where separate modality-specific pipelines converge at higher network layers. The trade-offs between these methods lie in their computational requirements and flexibility. Early fusion can enhance learning by facilitating cross-modal interactions but may increase computational overhead [44]. Late fusion, conversely, allows tuning of each modality independently, offering a modular approach that simplifies incorporation of additional modes without re-training the entire model.

Emerging trends highlight the implementation of iterative training loops that refine model parameters progressively. Techniques such as co-learning, where the learning process is iteratively adjusted based on multi-task objectives, optimize the training efficacy across diverse modali-

ties. This paradigm supports the continuous adaptation and improvement of MLLMs as they are exposed to broader and more varied data, thereby improving generalization capabilities [5].

Critically, foundational training strategies are challenged by issues of data quality and modality imbalance. Ensuring balanced learning from each modality is paramount, given that more abundant modalities can inadvertently dominate the model's learning focus. Studies such as "PMR: Prototypical Modal Rebalance for Multimodal Learning" [45] emphasize the implementation of measures like data augmentation and prototype-based adjustments to bolster underrepresented modalities, promoting uniform learning.

In conclusion, foundational training strategies for MLLMs are a dynamic blend of data integration, modality fusion, and iterative refinement processes that contribute to the advancement of comprehensive, cohesive models capable of handling diverse tasks. Future directions may explore more sophisticated pre-training methods that can better accommodate new conditions and data types, enhancing the robustness and adaptability of MLLMs. Furthermore, innovations in cross-modal co-learning and iterative feedback loops will likely play a critical role in driving the next wave of advancements. Such developments offer promising prospects for broader applications, with the overarching aim to effectively transform MLLMs into versatile tools that closely mimic human-like understanding and interaction across multiple data modalities.

3.2 Advanced Pre-training and Adaptation Techniques

In the realm of Multimodal Large Language Models (MLLMs), achieving effective pre-training and fine-tuning is vital for harnessing the synergy between various data modalities. This subsection delves into advanced strategies that bolster MLLMs' capabilities to integrate and process diverse inputs, adapting to specific task requirements with heightened efficiency.

Building on foundational strategies, advanced pre-training techniques like cross-modal pre-training play a crucial role by forging unified representations across disparate modalities. A widely adopted approach is the use of contrastive learning, which facilitates coherent integration by minimizing the distance between semantically similar data from different sources [46]. Contrastive methods leverage large collections of paired data, enhancing the model's ability to generate meaningful embeddings that capture the shared semantic spaces of different modalities [25]. Despite their effectiveness, these approaches face challenges in ensuring reliable alignment when modalities display inherent disparities in sampling rates or temporal characteristics [20].

Equally crucial for adapting pretrained MLLMs to downstream applications are task-specific fine-tuning methodologies. Techniques such as parameter-efficient fine-tuning have gained prominence, especially in resource-constrained scenarios. Methods like Low-Rank Adaptation (LoRA) adjust only a fraction of the model parameters, achieving notable task performance gains without incurring substantial computational costs [47]. The agile adaptation across a variety of contexts, while maintaining robustness and accuracy, is further facilitated by sparse mixture-of-experts architectures that activate different sub-networks based on task requirements, effectively scaling model capacity [48].

Beyond these techniques, emerging trends emphasize the synergy between pre-training frameworks and domain-specific knowledge incorporation. Modularized networks combining visual, linguistic, and other sensory knowledge modules address the challenge of modality collaboration. For instance, leveraging visual encoders alongside language models enhances the interpretation of complex visual-textual relationships, which is critical in tasks such as visual question answering [49]. Achieving seamless integration of nonlinear modality inputs without substantial pre-alignment, however, poses an ongoing challenge.

Advanced pre-training techniques also encompass dynamic modality fusion strategies, which adjust modality prioritization in real-time to optimize performance across tasks with varying demands [22]. This dynamic nature, while enhancing adaptability and robustness, requires sophisticated routing and gating mechanisms.

Moreover, the exploration of state space models to reduce episodic overheads associated with attention mechanisms represents another promising pathway towards more efficient pre-training techniques. Such models offer viable alternatives for scenarios demanding consistent handling of long sequences, thus streamlining computational processes [50]. Balancing transfer learning advantages with domain-specific fine-tuning helps mitigate overfitting while preserving generalization capabilities.

In conclusion, advanced pre-training and adaptation techniques hold considerable promise in fortifying MLLMs' multimodal capabilities. Nonetheless, challenges persist, particularly in harmonizing integration architectures to balance efficiency and accuracy with minimal redundancy. Future research is poised to explore innovations in optimizing sparse expert models, adaptive fusion strategies, and leveraging state-space efficiencies, which could lead to models with broad multimodal prowess and fine-grained task adaptability. Such advancements pave the path toward the development of truly general-purpose intelligent systems, seamlessly integrated across multiple data modalities.

3.3 Knowledge Transfer and Sharing Mechanisms

The process of knowledge transfer and sharing within Multimodal Large Language Models (MLLMs) is pivotal for enhancing their ability to leverage different modalities effectively and deliver comprehensive across-domain solutions. This section delves into the theoretical frameworks and practical mechanisms that facilitate cross-modal knowledge transfer, a concept crucial for most contemporary multimodal applications. By integrating insights from various approaches, this analysis will clarify how MLLMs transform and share knowledge, with an emphasis on addressing context-specific challenges and opportunities.

To begin, the notion of cross-modal transfer learning is central to this discussion. Cross-modal transfer learning involves transferring knowledge from a model trained on one modality to enhance the learning experience on another modality. This concept is explored through techniques that

leverage shared latent spaces for representation learning [51]. For instance, multimodal transformer architectures, such as those discussed in the reference [20], implement a shared attention mechanism that guides inter-modality interaction without explicit data alignment, offering a robust solution to misaligned inputs.

One effective knowledge-sharing paradigm is the use of high-dimensional interaction networks, as detailed in [52]. This framework showcases how complex networks can facilitate deeper interactions between modalities, essentially serving as conduits for shared knowledge and improved integration across modalities. The use of adversarial networks to align different modal distributions by translating them into a unified representation space ensures that models are adaptable and resilient to input variations [53].

Various models operate on the principle of a shared embedding space to achieve effective cross-modal knowledge transfer. A pertinent example is ImageBind, which articulates a joint embedding space across multiple modalities by aligning them to images [54]. Such frameworks use images as anchor points to unify diverse data types into a coherent semantic space, thus enriching cross-modal understanding through learned associations. The versatility seen in models like ImageBind showcases the power of using core modalities as integrative hubs for other data types, fostering richly connected networks capable of semantic fusion and cross-modal retrieval.

Moreover, the interplay between modality-independent and modality-dependent factors is critical in the transfer of knowledge. Exploring the dynamics between these factors helps identify bottlenecks and potential enhancements, facilitating better informed cross-modal sharing [55]. Models that bifurcate representations into modality-specific and invariant components demonstrate improved robustness to noise and missing data, underlining the strength of task-specific discrimination principles in fostering effective cross-modal fusion.

A key challenge remains the scalability and efficiency of these mechanisms, particularly as models grow in complexity and size. The shift towards parameter-efficient fine-tuning approaches, such as those proposed in the literature [56], highlights ongoing advancements in reducing computational overhead while capitalizing on cross-modal synergies. These strategies allow MLLMs to dynamically adapt to new tasks and domains by efficiently re-calibrating the existing knowledge architecture, without extensive resource expenditure.

Further exploration involves enhancing alignment techniques for stronger cross-modal connections. Adaptive alignment techniques, such as those in [57], propose tuning alignment capacities to meet specific task requirements, potentially circumventing issues of modality misalignment and ensuring more precise task-oriented knowledge transfer.

In conclusion, the frameworks guiding cross-modal knowledge transfer and sharing within MLLMs are advancing rapidly, driven by innovations in shared embedding paradigms and scalable, adaptive learning techniques. Future research directions could explore the integration of even more diverse modalities into these networks, refining fusion strategies, and improving context-sensitive adaptability. Continued refinement of these models holds the promise of enhancing their applicability in varied and complex real-world tasks, achieving levels of integration that mimic human-like multisensory learning and decision-making.

3.4 Optimization and Resource-Efficient Training

In the rapidly evolving landscape of Multimodal Large Language Models (MLLMs), optimizing resource-efficient training paradigms has become increasingly important. This subsection delves into various strategies aimed at reducing computational demands while maintaining or enhancing model performance. We examine these methodologies through parameter-efficient techniques, low-rank and quantization approaches, and the development of resource-efficient frameworks, aligning them with the broader context of cross-modal knowledge transfer and training challenges discussed previously.

Parameter-efficient fine-tuning (PEFT) emerges as a vital optimization strategy for MLLMs. By emphasizing efficiency in parameter usage, PEFT techniques enable effective model adaptation with minimal resource expenditure. Methods such as Low-Rank Adaptation (LoRA) have proven instrumental, significantly reducing the amount of trainable parameters by focusing on only the low-rank portions of the model's weight matrices [58]. This approach preserves the essential model weights, allowing rapid and efficient adaptation to new tasks or datasets while maintaining foundational characteristics.

Low-rank and quantization techniques further facilitate resource-efficient training by approximating model weights through lower-dimensional matrices or discrete representations. Quantization, which reduces the precision of weights while maintaining robustness, allows for substantial reductions in memory footprint and computational requirements [59]. Similarly, low-rank decomposition approximates full-rank weight matrices with lower-rank counterparts, achieving computational efficiency by reducing the model's dimensional complexity.

Novel architectures also contribute to efficient training. The Mamba language model, extended into multimodal systems owing to its linear computational complexities, exemplifies improved computational efficiency [60]. These models adapt quickly to various input modalities, reducing the extensive computational time typically required for processing large multimodal datasets.

Moreover, token reduction strategies play a pivotal role. Techniques such as Transfusion, which blends language modeling with diffusion processes, allow a single model to work with multimodal sequences, thereby minimizing token counts per input [61]. In tandem, PruMerge employs adaptive token reduction mechanisms, clustering visual tokens by importance and merging redundant ones to retain informational fidelity while reducing computational load [62].

Emerging trends indicate a shift towards self-supervised and weakly-supervised approaches to resource optimization during training. Leveraging vast unlabeled datasets for pre-training enables models to generalize with fewer labeled examples. Multimodal In-Context Instruction Tuning [63]

exemplifies such strategies, improving learning efficiency by aligning large volumes of unlabeled data with task-specific objectives through in-context learning paradigms.

Despite considerable advances, challenges remain, particularly in balancing efficiency and accuracy with highly unstructured or noisy multimodal inputs. Additionally, integrating cutting-edge solutions such as shared multimodal embeddings presents challenges in ensuring consistency across different tasks and data scales.

In conclusion, the future of MLLM training hinges on further refinement of resource-efficient methodologies, sophisticated low-rank approximations, and advanced quantization schemes, building upon the cross-modal knowledge transfer and training strategies previously outlined. As the field continues to grow, ongoing development and empirical studies will enhance the scalability of MLLMs, enabling them to operate efficiently across increasingly complex multimodal domains. Collaborative efforts across sectors will be crucial in meeting the nuanced optimization demands of multimodal language models, aligning them cohesively with the challenges and innovations discussed throughout this survey.

3.5 Challenges and Solutions in Training

The training of Multimodal Large Language Models (MLLMs) introduces a unique set of challenges, stemming primarily from the inherent complexities of integrating and balancing multiple data modalities, as well as computational and data resource constraints. Addressing these challenges requires tailored strategies that ensure effective and efficient model training while maintaining generalization capabilities across diverse multimodal tasks. This subsection explores these challenges and proposes innovative solutions, drawing on recent research advancements to provide a comprehensive framework for overcoming these obstacles.

One primary challenge in training MLLMs is addressing data imbalance and bias, which often emerge due to disparate data quality, quantity, and distribution across modalities. For instance, textual data tends to be more abundant and well-curated compared to visual or audio data, potentially skewing the model's learning process. Addressing this issue necessitates sophisticated data augmentation and preprocessing techniques to ensure balanced representation across modalities. Techniques such as cross-modal data augmentation, which generates synthetic samples that combine elements of various modalities, can mitigate these imbalances by enriching underrepresented samples and ensuring more uniform coverage [64]. Furthermore, alignment strategies that focus on equating the semantic importance of modalities using balanced weighting during training have shown promise [25].

Another significant challenge in MLLM training is preventing overfitting and enhancing generalization, especially given the high-dimensionality and complexity of multimodal data spaces. The adoption of regularization techniques, such as dropout and weight decay, has been an effective strategy to combat overfitting by preventing the model from excessively capturing noise and random fluctuations in the data [65]. Moreover, employing techniques such as sparsity awareness and mixed precision training, as explored in Activation-aware Weight Quantization for LLM Compression and Acceleration [66], can reduce overfitting risks while enhancing computational efficiency.

In addition, computational demands pose a critical challenge in the training process of MLLMs, given the extensive resources required for processing multimodal datasets, pre-training large networks, and fine-tuning models on specific tasks. Resource-efficient training approaches, such as low-rank adaptation (LoRA) [42], sparse fine-tuning methods [67], and techniques like mixture-of-experts (MoE) modeling [37], have been instrumental in reducing the computational footprint without sacrificing performance. These methods leverage advancements in model pruning, parameter sharing, and conditional computation to optimize resource allocation, ultimately enabling scalable MLLM development and deployment.

Emerging trends in self-supervised and semi-supervised learning paradigms offer new avenues for overcoming data constraints and improving model adaptability. Leveraging vast amounts of unlabeled data, these paradigms utilize contrastive and generative pre-training to uncover latent data patterns, allowing MLLMs to learn robust feature representations without reliance on exhaustive annotations [4]. The introduction of retrieval-augmented learning techniques, which augment decisions by integrating external knowledge during training, further enhances the models' domain adaptability and inference accuracy [5].

As MLLMs continue to advance, it is crucial to address the ethical implications associated with their development. Implementing fairness constraints and bias detection mechanisms during training ensures that models make unbiased and equitable decisions. Techniques such as differential privacy and secure multi-party computation serve to uphold data security and privacy throughout the training lifecycle [68]. By aligning technical innovation with ethical guidelines, the research community can foster the responsible deployment of MLLMs, maximizing their positive societal impact.

In conclusion, while the training of MLLMs presents formidable challenges, the development of multifaceted strategies encompassing data optimization, computational efficiency, and ethical considerations, provides a robust foundation for progress. Future research should focus on refining these strategies and exploring new methodologies to further improve the scalability and generalization of MLLMs across diverse applications. Such advancements will be pivotal in unlocking the full potential of multimodal artificial intelligence systems.

3.6 Evaluation and Benchmarking in the Training Process

In the domain of Multimodal Large Language Models (MLLMs), effective evaluation and benchmarking of training processes are crucial for advancing the understanding and development of these sophisticated systems. This subsection delves into the methodologies and benchmarks employed during the training phase of MLLMs, emphasizing their roles in assessing model performance and identifying enhancement opportunities.

Central to MLLM evaluation is the use of benchmark datasets and tasks designed to test a model's capacity to understand and integrate information across multiple modalities such as text, images, and audio. Leading datasets, like COCO and VQA, provide standardized environments challenging models across vision-language tasks, ensuring comprehensive assessment in diverse scenarios [17]. However, these benchmarks have limitations, such as potential data bias and a lack of real-world complexity, which might lead to performance bottlenecks in diverse applications [12].

Performance metrics during training are critical for assessing how well an MLLM performs across different modalities. Conventional metrics, including accuracy, precision, recall, and F1 score, are adapted for evaluating multimodal inputs and outputs. Additionally, metrics tailored for multimodal contexts, such as cross-modal retrieval scores and integrated harmony indexes, offer a comprehensive view of multimodal interactions [25]. These metrics can reveal nuances in modal alignments and the model's ability to leverage synergies between diverse data types.

A notable trend in evaluating MLLMs is the shift towards more nuanced frameworks encompassing qualitative dimensions, such as robustness and interpretability. Traditional numerical benchmarks are increasingly complemented by qualitative assessments that evaluate interpretative capabilities, vital for human-model interactions in high-stakes environments [69]. This signifies a paradigm shift from focusing solely on quantitative outcomes to understanding the intrinsic decision-making processes within MLLMs.

Emerging benchmarking techniques aim to evaluate complex reasoning, adaptability, and ethical implications of MLLMs. Multi-criteria evaluation frameworks assess not only performance accuracy but also stability, robustness, and ethical considerations, facilitating a holistic view of model capabilities [18]. Such approaches highlight strengths and uncover overlooked weaknesses in model design and function.

Despite progress, challenges remain in evaluating MLLMs. One key issue is the saturation of traditional benchmarks; as models excel at specified tasks, differentiating between truly innovative models and those finely tuned for benchmark performance becomes challenging [3]. Furthermore, rapid advancements in MLLMs necessitate updated benchmarks that reflect real-world complexities and ethical responsibilities.

Looking forward, future directions for evaluating MLLMs should include ethically grounded benchmarks addressing fairness, bias, and societal impacts. New benchmarks should consider safety, privacy, and potential biases affecting predictions and interactions [70]. Incorporating dynamic, customizable benchmarks simulating real-world applications can help align MLLMs closer to practical use cases, improving applicability and trustworthiness in diverse settings.

Ultimately, evaluating MLLMs during training is not merely about assessing immediate performance but understanding broader implications and potential. As research continues pushing MLLMs' boundaries, robust and thoughtful evaluation methods will guide development, ensuring these models fulfill potential in equitable and beneficial ways for society.

4 Multimodal Data Processing and Representation Learning

4.1 Representation Alignment

To achieve cohesive and efficient information processing across multiple modalities, it is imperative to align the underlying representations of different data types. Representation alignment in multimodal large language models is a critical step that ensures a harmonious integration of modalities such as text, image, and audio. This subsection delves into various methodologies that establish alignment between disparate modal representations, highlighting their relative strengths, limitations, and impacts on cross-modal learning.

The primary objective of representation alignment is to create a unified semantic space where representations from different modalities converge, thus facilitating seamless interaction and integration. One widely adopted strategy is supervised embedding alignment, which leverages pre-trained models like CLIP to align visual and textual tokens. These models utilize contrastive learning techniques to bind semantically related elements from different modalities within a shared latent space [2]. Contrastive learning leverages paired data – such as images and corresponding descriptions – to maximize similarity for aligned pairs while minimizing it for unaligned ones. This approach has been shown to drastically improve the coherence and accuracy of fused representations, enabling models to excel in tasks such as visual question answering and cross-modal retrieval.

Implicit multimodal alignment, on the other hand, focuses on achieving alignment without explicit multimodal fine-tuning. This method capitalizes on model architecture designs that intrinsically enable the alignment of perceptual and textual representations. For instance, models that employ attention-based mechanisms like transformers can exploit self-attention layers to progressively merge information from different modalities through shared attention patterns [16]. Despite not necessitating extensive supervised data for every possible modality combination, these methods still achieve substantial fusion quality, albeit facing challenges in scalability and interpretability in complex scenarios.

Comparatively, supervised methods tend to offer more control over the alignment process and typically result in higher accuracy in specific tasks due to the explicit pairing of instances during training. However, they require substantial labeled data, which may not always be available or feasible to procure. Implicit approaches, while demanding less curated data, may struggle with ambiguity and divergences in semantic interpretation without explicit guidance.

Emerging trends in representation alignment involve the exploration of hybrid methods that blend explicit and implicit strategies, as well as the utilization of large pre-trained models with adaptation layers to transfer alignment capacity to new tasks with minimal labeled data. Such solutions aim to harness the benefits of both strategies – leveraging the robustness of supervised alignment and the flexibility of implicit models [4].

Technical nuances such as the design of joint embedding spaces are also critical in aligning different modalities. Joint embedding spaces are crafted to ensure that embeddings from different modalities are not only semantically aligned but are also capable of further fusion or combination during downstream tasks [10]. This involves carefully defining the dimensions and properties of the embedding space, potentially using sophisticated loss functions that can quantify and enforce the similarity across modalities.

Ensuring that aligned representations are robust and free of biases is another critical challenge. Techniques such as data augmentation and adversarial training are often deployed to create more robust embeddings, which can withstand variations and inconsistencies inherent across datasets of different modalities [5].

In synthesizing the directions for future exploration, representation alignment could benefit from a greater focus on domain-specific adaptations that enhance alignment in specialized fields such as medical imaging or autonomous systems. Moreover, the integration of real-world feedback mechanisms could dynamically adjust the alignment process, enabling models to adapt to evolving data landscapes and user interaction contexts [8].

In conclusion, representation alignment remains a cornerstone of effective multimodal processing, offering a pathway to cohesive and interpretable integration of diverse data types. As methodologies evolve, the balance between supervision, scalability, and adaptability will dictate the progress in aligning complex multimodal systems, promising more nuanced and accurate cross-modal understanding.

4.2 Multimodal Fusion Strategies

In the field of multimodal data processing, the fusion of various modalities—such as text, image, audio, and video inputs—into a cohesive representation is critical. This aspect of multimodal large language models capitalizes on the complementary nature of different modalities, thereby enhancing the predictive and reasoning capabilities of AI systems. This subsection delves into the methodologies employed for integrating multimodal inputs, examining the technical nuances, strengths, limitations, and emerging trends.

Starting with early fusion techniques, also known as feature-level fusion, the focus here is on combining raw data from different modalities at the input or shallow layer stages. Such an approach allows for the simultaneous processing of modalities, fostering an intrinsic understanding of cross-modal dependencies at foundational levels. However, early fusion can be vulnerable to robustness issues, especially when handling incomplete or noisy data, as it requires consistent sampling rates and feature dimensions across modalities—an often impractical requirement in real-world scenarios [20].

Conversely, late fusion, or decision-level fusion, involves aggregating the outputs or high-level features from individual modality-specific models. This strategy enables each modality to be processed using architectures tailored to its unique characteristics before a collective decision-making process is engaged. While late fusion preserves the distinct qualities of each modality and effectively manages heterogeneous data, it may lack the synergy that earlier integration could provide, as interactions are primarily considered at the culmination of the processing pipeline [2].

Middle fusion strategies, which reside between early and late paradigms, facilitate integration at intermediate layers through mechanisms like attention-based fusion or cross-modal transformers. These approaches capitalize on shared latent spaces to enable synergistic learning while maintaining sufficient separation to preserve modality-specific information. Notably, Multimodal Transformers utilize directional pairwise cross-modal attention to ensure robust integration across input sequences, effectively addressing challenges such as non-alignment and long-range dependencies inherent in multimodal data [20].

An innovative technique in fusion is the use of mixture-of-experts (MoE) architectures, which dynamically activate a subset of specialized sub-models (experts) to process inputs depending on context. This sparse approach allows for scaling model capacity with reduced computational costs and is particularly effective in integrating modalities like vision and language due to its dynamic routing capabilities [71]. Furthermore, recent literature suggests utilizing entropy-based regularization schemes to enhance the stability and balanced application of these mixtures [46].

Emerging methodologies in fusion incorporate adaptive and dynamic fusion strategies that consider task relevance and data context when prioritizing modalities. For instance, retrieval-augmented mechanisms improve decision-making by integrating external knowledge into the fusion process, thereby utilizing information beyond the immediate inputs [24]. Moreover, alignment techniques that minimize the need for extensive fine-tuning across different modalities hold promise for enhanced flexibility and scalability in large model architectures [24].

Despite these advancements, significant challenges persist. Designing fusion strategies that uniformly manage the diverse and often asynchronous characteristics of multimodal data while maintaining computational efficiency remains a pressing issue. Ensuring robustness against missing or noisy modalities is crucial for real-world applications. Addressing cross-modal inconsistencies and the influence of biased representations from one modality on the overall prediction poses further hurdles [11].

Looking forward, the development of flexible, adaptive fusion frameworks that consider the variable significance of modalities in context will likely lead future research. Strategies that dynamically incorporate user feedback and enable real-time context adaptation, using hybrid approaches that combine knowledge-driven and data-driven methods, will likely define the evolution of multimodal fusion. This trajectory promises not only improvements in performance but also the incorporation of accountability and transparency, which are key to ethical AI deployment [72].

In conclusion, while substantial progress has been made, the quest for optimal multimodal fusion is ongoing. Continued research explores innovative paradigms that converge the fields of computer vision, natural language processing, and other domains to realize holistic AI solutions.

4.3 Robust Representation Learning

In the domain of multimodal large language models, robust representation learning is paramount for ensuring that

learned models generalize effectively across a variety of input scenarios. This subsection delves into techniques to enhance the robustness and generalization capabilities of multimodal representations, addressing challenges such as modality-specific biases and hallucinations. By evaluating and synthesizing different approaches, we aim to identify trends, offer insights, and propose future directions in this crucial area of research.

A key challenge within multimodal learning is the presence of modality-specific biases, which can lead to hallucinations—outputs that are not grounded in any input modality. Such biases often stem from discrepancies in the contributions of different modalities during the learning process. Techniques such as Noise Perturbation and Regularization, as introduced in frameworks like NoiseBoost, have shown efficacy in balancing attention distribution across modalities [28]. This approach involves perturbing inputs with noise to enforce more equitable attention across modalities, thereby mitigating issues where one modality disproportionately influences the neural decision process.

Another promising strategy is data-driven augmentation, which leverages techniques such as attribute-based diversification to enrich datasets and fortify model robustness. By dynamically transforming feature spaces using multimodal data, these methods ensure broad data coverage and enhance inferential accuracy [44]. This approach effectively combats data sparsity and improves the robustness of learned representations, supporting more resilient model behavior across diverse scenarios.

The use of self-supervised and semi-supervised learning paradigms has also proven effective in this field. These paradigms facilitate the learning of joint embeddings by capitalizing on inherent data patterns without extensive reliance on labeled datasets [73]. Contrastive pre-training, for instance, leverages latent relationships across modalities to build representations that are invariant to certain perturbations, fostering robustness in zero-shot and low-resource settings.

Moreover, representing multimodal data in a unified, modality-agnostic space has garnered considerable attention. Approaches like Modality-Invariant and -Specific Representations framework (MISA) project each modality into a shared space, fostering robust learning even in cases of missing modalities [74]. The framework separates modality-specific characteristics from shared components, thus reducing the impact of modality gaps and fortifying the robustness of embeddings.

Despite these advancements, several challenges persist. This includes the need for models to effectively adapt to new modalities and scaling issues inherent in fusing diverse data types. The emerging techniques, such as cross-modal translations to ensure joint representation learning in the presence of missing data, show potential in addressing these issues [53]. Learning methods that incorporate adaptive mechanisms to dynamically adjust importance across modalities based on context are also gaining traction [75].

In synthesizing these approaches, it becomes evident that the future trajectory in robust representation learning will likely emphasize hybrid strategies that integrate the strengths of individual techniques. For instance, combining data augmentation with self-supervised learning stakes a claim to address both data sparsity and bias simultaneously [76].

In conclusion, robust multimodal representation learning is a multifaceted challenge that is being addressed through a diverse array of strategies. While significant progress has been made in mitigating modality-specific biases and hallucinations, future research endeavors should focus on hybrid approaches that amalgamate various advancements to offer comprehensive solutions. Integrating methods like dynamic adaptation and cross-modal learning will bolster the generalization capabilities of these models, paving the way for more reliable and versatile multimodal applications. By continuing to innovate in these areas, the field will substantiate its drive towards realizing truly intelligent and adaptable models.

4.4 Self-supervised and Semi-supervised Learning

In the field of Multimodal Large Language Models (MLLMs), self-supervised and semi-supervised learning paradigms are pivotal for exploiting the vast, yet often unlabeled, multimodal datasets prevalent today. These techniques have the potential to reshape how models learn and interpret multimodal signals, enhancing their applicability and performance without incurring the high costs associated with acquiring labeled data.

Self-supervised learning (SSL) leverages intrinsic data structures by formulating pretext tasks that allow for automatic label generation. Contrastive learning stands out in this domain, training models to identify similarities and differences among data point pairs [20]. Within a multimodal context, this often entails developing cross-modal representations, where, for instance, a text snippet must be accurately paired with its corresponding image amidst a selection of distractors. Lin et al. propose models focused on directional pairwise cross-modal attention, adeptly capturing interactions between non-aligned sequences, thus bolstering representational robustness across modalities without the need for manually curated labels [20].

Generative-based methods, another significant SSL avenue, prioritize the reconstruction of incomplete or obscured data, employing architectures such as autoencoders and transformers. These techniques predict missing image patches from surrounding context, fostering the learning of robust feature representations. The Perceiver model exemplifies this approach, employing iterative attention to condense inputs into compact latent representations [77], facilitating the handling and generation of extensive inputs with greater efficiency. These generative mechanisms inherently leverage consistent data patterns across multiple modalities, offering scalable solutions for large-scale learning challenges.

In contrast, semi-supervised learning (SSL) enhances limited labeled data with a more extensive reservoir of unlabeled examples. Techniques like retrieval-augmented learning are crucial here, improving model performance by integrating external knowledge bases, thus resolving ambiguities seen in sparse labeled datasets [78]. Such methods enable MLLMs to excel in zero-shot tasks, utilizing retrieved data to inject semantic context into the inferential process.

Innovative retrieval-based strategies, as seen in MIMIC-IT [63], develop rich instruction-response pairs, empowering models to not only think but act by using large volumes of unlabeled data for performance gains. These scenarios enable multimodal models to fine-tune with synthetically generated labels, reinforcing robust learning paradigms with minimal supervised input.

Despite these advancements, challenges persist, notably the need to reconcile modality disparities in data resolution and sampling rates [79]. Furthermore, designing architectures that balance learning efficacy with computational efficiency remains a crucial task. Emerging architectures must creatively integrate modalities to preserve the structural integrity of learning pipelines without incurring excessive computational drawbacks.

Looking ahead, self-supervised and semi-supervised learning in multimodal contexts shows great promise. Emerging trends suggest hybrid models combining generative and contrastive techniques, hypothesis-driven self-supervision guiding unsupervised learning, and more sophisticated resource-efficient models scaling multimodal applications [80]. Additionally, the fusion with reinforcement learning frameworks heralds new opportunities for models to refine continuously through dynamic interactions with varied environments.

As research continues, synthesizing SSL and semi-supervised approaches with multimodal architectures could unlock unparalleled flexibility and cognitive resilience, paving the way for models that can truly generalize and immerse in learning across diverse modalities. These efforts aim to cultivate a new generation of intelligent systems, capable of intuitive learning from their multifaceted and information-rich environments, complementing the overarching need for robust representation learning and adaptive methodologies essential for dynamic representation adaptation in MLLMs.

4.5 Dynamic Representation Adaptation

Dynamic representation adaptation is a crucial facet of multimodal data processing and representation learning, particularly within the context of multimodal large language models (MLLMs). This subsection delves into the methodologies that enable the dynamic and context-sensitive adjustment of multimodal representations, fostering their adaptability across a range of data contexts and model tasks. Such adaptability is instrumental in realizing the potential of MLLMs across various domains, enhancing their applicability and efficiency in real-world scenarios.

The core of dynamic representation adaptation lies in its ability to modulate and align multimodal inputs based on changing task requirements and data contexts. One prominent approach involves modality switching and adaptation, where models can dynamically adjust or switch between modalities depending on the task at hand. This capability ensures that the most relevant data is prioritized at runtime. For example, the work on conditional multimodal generation emphasizes the conditional adjustments needed for generating one modality from another [81].

Scaling with Mixture of Experts (MoE) presents another innovative framework that supports dynamic adaptation through the selective activation of expert sub-models dedicated to specific modalities. This approach not only enhances the model's ability to manage diverse inputs efficiently but also optimizes resource allocation by activating only the necessary model components. MoE strategies have been effectively utilized to manage the complexities and computational demands of large-scale models [36], [37].

Dynamic adaptation also incorporates techniques like feature re-calibration and selective attention mechanisms, which permit the model to focus selectively on task-critical features while disregarding less relevant information. Such mechanisms are facilitated by advanced neural architectural components including attention mechanisms and gating networks, which allow the dynamic fine-tuning of the interpretative focus in relation to the context [43].

However, utilizing dynamic representation adaptation is not without its challenges. Key among these is the computational overhead associated with continuously adjusting the model representations in response to new data inputs and task requirements. While MoE approaches address this to an extent by enabling efficient computation only in relevant parts of the model, they introduce additional complexity in training and model architecture management. Furthermore, ensuring the robustness of these adaptive processes in the face of noisy or incomplete data remains a significant challenge. Models must be able to adapt dynamically without succumbing to performance degradation or increased error rates [65].

An emerging area of interest within this domain is the development of task-aware dynamic adaptation frameworks, which leverage task-specific contextual information to refine and guide the adaptation process. Such techniques involve training models on task-conditioned datasets that align similar contextual cues with specific adaptation strategies, thereby enhancing both accuracy and robustness. Incorporating task knowledge can also improve the interpretability of the adaptation process by providing explicit guidance to the model on modality importance [82].

Looking ahead, the trajectory of dynamic representation adaptation is poised to focus increasingly on enhancing scalability and computational efficiency. Furthermore, integrating dynamic representation adaptation with other AI subfields such as transfer learning and domain adaptation can further broaden its applicability. Future research may explore the synergistic merging of dynamic adaptation with continuous learning paradigms to better accommodate evolving data streams and shifting context within environments featuring high variability and uncertainty [5].

In conclusion, dynamic representation adaptation represents a frontier of innovation in multimodal learning, empowering MLLMs to adjust responsively to a myriad of contexts and tasks. Continued advancements in this field promise to unlock new potentials for AI applications, particularly in environments that demand high adaptability and precision. As we refine these adaptive models, the potential for creating versatile, efficient, and intelligent systems expands, paving the way for next-generation AI technologies. This underscores the need for ongoing research and development, drawing on interdisciplinary insights to overcome current limitations and chart new pathways for innovation.

5 Evaluation and Benchmarking

5.1 Performance Metrics and Standard Benchmarks

The evaluation of Multimodal Large Language Models (MLLMs) is a complex endeavor requiring a comprehensive framework to assess diverse model capabilities across different modalities. This subsection delves into the performance metrics and standard benchmarks pertinent to MLLMs, outlining their significance, application, and areas of challenge.

A robust set of performance metrics is essential for evaluating MLLMs, encompassing both traditional metrics and those adapted for multimodal contexts. Standard metrics include accuracy, precision, recall, and F1 score, commonly used in classification tasks to measure the rate of correct predictions [24]. However, multimodal tasks necessitate the adaptation of these metrics to account for the integration and interaction of multiple data types. For example, in tasks such as Visual Question Answering (VQA), metrics like BLEU and CIDEr, originally developed for natural language processing tasks, are employed to evaluate the fluency and comprehensibility of generated responses against reference answers [5].

Beyond these, novel performance metrics specific to MLLMs have been proposed, such as the Empirical Multimodally-Additive Function Projection (EMAP), which isolates the contribution of cross-modal interactions by attributing changes in performance exclusively to multimodal synergies [3]. This metric addresses a critical challenge in multimodal evaluation — ensuring that models are genuinely leveraging multimodal data rather than relying on unimodal signals.

To provide a consistent framework for evaluation, several standard benchmarks have been developed. The COCO dataset is widely used for image captioning tasks, allowing MLLMs to be evaluated on their ability to generate descriptive text from visual content [2]. Similarly, the VQA dataset assesses a model's comprehension by posing intricate questions that require understanding and reasoning over paired text-image inputs. Other datasets, such as CLEVR, focus on evaluating a model's capability to perform logical reasoning within complex visual scenarios, thereby pushing the boundaries of current MLLM capabilities in terms of reasoning and deduction [44].

Emerging benchmarks such as MM-Vet extend the evaluation to more intricate multimodal tasks, testing models on scenarios that necessitate the integration of core vision-language capabilities [13]. These benchmarks incorporate a range of tasks and metrics to holistically evaluate advanced MLLM functionalities, such as event recognition and reasoning about dynamic visual scenes.

Despite the advances, current practices in multimodal evaluation face significant challenges. One key issue is the saturation of benchmarks where models consistently achieve high scores without demonstrating genuine multimodal comprehension, as identified through diagnostic tools like EMAP [3]. This calls for the development of more discriminative benchmarks that can tease apart performance attributable to true multimodal integration versus unimodal baseline strategies.

Additionally, issues such as data bias and lack of context capture in existing benchmarks pose further challenges. These problems can undermine the fairness and applicability of MLLMs in real-world scenarios [5]. Addressing these issues requires not only the design of new datasets that are more representative of diverse populations and settings but also the incorporation of contextual reasoning abilities into evaluation frameworks.

Looking ahead, the evolution of MLLM evaluation practices may include the development of benchmarks that simulate dynamic, real-world environments more accurately. This could involve using synthetic datasets to test models on unforeseen and adaptive scenarios, thereby encouraging robust, adaptable learning strategies [18]. Furthermore, as the field progresses, there is a growing call for qualitative assessments that incorporate user feedback and human judgment, thus providing additional dimensions to traditional quantitative evaluations.

Overall, advancing the field of MLLM evaluation necessitates a concerted effort to develop more comprehensive metrics and benchmarks that can accurately capture the nuanced capabilities of these models while addressing inherent challenges of bias and context-dependence. Through such enhancements, we can ensure that MLLMs are not only state-of-the-art in performance metrics but also robust, fair, and applicable in diverse real-world contexts.

5.2 Challenges in Multimodal Evaluation Practices

The evaluation and benchmarking of Multimodal Large Language Models (MLLMs) present a multifaceted landscape, encompassing numerous challenges and limitations that demand substantial research attention to enhance current practices. This subsection delves into these challenges, considering biases, contextual understanding, benchmark saturation, and evaluation methodologies that critically impact the assessment of these advanced AI systems.

Multimodal datasets inherently carry biases due to the varied origins and methods of data collection, affecting the fairness and accuracy of evaluations. For instance, socioeconomic or cultural biases present in visual datasets can significantly skew model outputs, impacting downstream applications requiring fairness and equity. Addressing these biases is crucial yet challenging, necessitating comprehensive strategies in data curation and selection processes. The "Quantifying & Modeling Multimodal Interactions: An Information Decomposition Framework" underscores the importance of evaluating datasets' multimodal interactions to accurately quantify redundancies and synergies essential for robust task performance. This approach aids in identifying biases and guiding corrective measures, although its complexity often limits practical application in large-scale evaluations [44].

A notable challenge is the limited capacity of current benchmarks to encapsulate the intricate context dependencies inherent in real-world multimodal data. Traditional benchmarks like VQA or COCO often concentrate on isolated tasks and scenarios, failing to account for the cross-modal interdependencies and sequential nature of real-world tasks. These limitations hinder MLLMs' ability to accurately reflect practical applicability, prompting the need for more comprehensive and context-rich evaluation benchmarks. Benchmarks such as MMMU: "A Massive

Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI" aim to address this by furnishing a diverse set of tasks requiring integrated multimodal reasoning, though balancing task diversity with the necessary depth to capture true contextual understanding remains a significant challenge [72].

Benchmark saturation poses another daunting challenge, where models achieve high scores across standard tasks without yielding genuine advances in interpretative or generative capabilities. This saturation mirrors the phenomenon in NLP models that optimize towards dataset-specific cues rather than authentic comprehension. To avoid stagnation, it is imperative to continuously evolve benchmarks to incorporate novel tasks and criteria that emphasize generalization and adaptability.

A further challenge is the frequent reliance on isolated modality performance evaluation, often overlooking the interconnectivity that characterizes real-world data. Evaluative frameworks such as "Multimodal Intelligence: Representation Learning, Information Fusion, and Applications" highlight the necessity for criteria that acknowledge the dynamics between modalities, an area underexplored in typical benchmark designs [25]. Additionally, the development of complex inter-modality tasks stresses the need for improved evaluation metrics that can adequately measure system coherence and multimodal interaction dynamics [29].

Technical limitations in evaluation design also present challenges. The variance in modality-specific token representations complicates direct comparative analysis, while current scoring mechanisms lack the resolution to discern nuanced differences in model performance [11]. This underscores the need for novel evaluation frameworks that utilize advanced scoring systems capable of managing high-dimensional multimodal interactions and representations.

Innovative directions suggest integrating qualitative evaluation paradigms alongside conventional quantitative metrics. The implementation of human-in-the-loop assessments could offer deeper insights into model shortcomings regarding ethical use, context understanding, and real-world applicability. Furthermore, fostering benchmark designs based on real-time data could alleviate the drawbacks of static benchmark information, thus enhancing adaptability and continuous evolution in evaluation practices.

In conclusion, advancing towards robust multimodal evaluation practices necessitates transitioning from isolated, modality-specific evaluations to comprehensive frameworks incorporating bias assessment, contextual depth, real-world applicability, and innovative evaluation metrics. Emerging benchmarks and methodologies indicate progress but require further refinement to adequately address the multifaceted nature of multimodal data. By adopting a multidimensional evaluation perspective grounded in both technological advancements and contextual understanding, the field can aspire to more effective integration and advancement of MLLMs, fostering progress with broader and more equitable societal implications.

5.3 Advanced Benchmarking Techniques

In advancing Multimodal Large Language Models (MLLMs), developing pertinent benchmarking techniques is paramount to thoroughly evaluating their capacity for complex reasoning, scalability, and flexibility. Traditional evaluations often encapsulate limited dimensions of performance, typically focusing on base metrics such as accuracy and F1 score. However, the intricate nature of MLLMs necessitates more sophisticated benchmarking frameworks capable of capturing their multifaceted capabilities and applications.

Emerging benchmarking approaches are increasingly incorporating task customization and dynamism. Tailored benchmarks like Task-Me-Anything exemplify this shift by allowing users to design and modify evaluation tasks that more accurately reflect real-world applications and the specific needs of end-users [73]. These frameworks empower practitioners to create scenarios that meticulously test models' adaptability and performance across varying contexts, enhancing the precision of comparative analyses between models.

Continuous evaluation systems represent another evolutionary leap in benchmarking MLLMs. Frameworks such as LMMS-EVAL and Multimodal LIVEBENCH dynamically update to include the most current datasets and tasks, reflecting the burgeoning advancements and challenges in multimodal AI. This adaptability ensures that assessments remain relevant and challenging, pushing models to adapt gradually to diverse and evolving expectations which, in turn, contributes to a continuous improvement cycle within the field of MLLM development.

Moreover, multi-criteria evaluation frameworks have emerged, aiming to provide a holistic understanding of model performance across various non-traditional metrics, such as ethical implications and energy efficiency. These frameworks demand that MLLMs are scrutinized not only based on how accurately they perform tasks but also concerning their robustness and potential biases, scalability issues, and adherence to ethical AI guidelines. The introduction of such diverse evaluation criteria is shaping a more nuanced understanding of what it means for an MLLM to be truly versatile and scalable [5].

In terms of scalability, techniques in benchmarking models often explore their ability to efficiently process increased data volume and complexity—key traits necessary to pave the path toward artificial general intelligence (AGI). Consequently, novel benchmarks are being designed to evaluate not just performance up to a task capacity but the models' efficiency in extrapolating solutions to problems involving previously unseen data or tasks.

Trade-offs in benchmarking are inevitable and require careful balancing. There remains a constant challenge in accounting for the inherent trade-offs between scalability and performance against comprehensive but resource-intense evaluation methods. Implementing complex, customizable benchmarks implicates more intricate and costly processing or training environments. Yet, they provide a depth of insight that basic benchmarks lack and may uncover latent capabilities or weaknesses only evident under more dynamic conditions [79].

A significant obstacle in the current landscape of advanced benchmarking techniques is achieving a consensus on standard metrics and datasets that reflect true generalization abilities without succumbing to saturation—a point where models perform optimally across a set bench-

mark without showing genuine field-scale understanding or adaptability [1]. Enabling real-world scenario simulation for MLLMs thus necessitates the introduction of hybrid evaluation approaches integrating both quantitative metrics and qualitative observations, potentially supplemented by human-in-the-loop assessments to maximize fidelity in practical applications [2].

In summary, while the foundational work in advancing benchmarking frameworks for MLLMs has made significant strides toward evaluating them in broader, more comprehensive ways, there remains ample room for development. Future directions should prioritize enhancing benchmark adaptability, ensuring evaluations are reflective of real-world application diversity while fostering improvements in scalability and ethical AI deployment [69]. These benchmarking advancements are not only crucial in assessing present capabilities but also in shaping the future trajectory of MLLM development towards more inclusive and intelligent multimodal systems.

5.4 Proposing Improvements and Innovations

In the realm of Multimodal Large Language Models (MLLMs), evaluating the intricate interplay across varied modalities remains a multifaceted challenge. Current evaluation paradigms predominantly emphasize traditional metrics such as accuracy and F1 scores; however, these often fail to capture the nuanced capabilities of these advanced models. Within this context, exploring innovative evaluation strategies becomes imperative to enhance and delineate the comprehensive capabilities of MLLMs, setting the stage for their future evolution and deployment.

One promising approach to transcend conventional metrics is the hybrid evaluation framework, which integrates both quantitative measures and qualitative assessments, such as user feedback, offering a more comprehensive view of MLLM performance. This holistic approach addresses the limitations of quantitative metrics alone, which frequently overlook the subjective user experience and qualitative nuances inherent in human-like interactions, as discussed in [79]. By incorporating qualitative user feedback, we can gain insights into the model's usability, adaptability, and real-world applicability.

Furthermore, the proposal of benchmarks specifically targeting ethical dimensions, such as safety and privacy, is crucial for the responsible development of MLLMs. As articulated in discussions surrounding ethical guidelines, integrating benchmarks that evaluate these dimensions ensures that MLLMs not only meet performance metrics but also align with ethical standards safeguarding user interests. Such frameworks entail evaluating models on their ability to preserve user data privacy, prevent discrimination, and maintain fairness across different demographic groups.

Regarding advancing cross-domain and transfer task evaluation, developing benchmarks that emphasize a model's ability to generalize knowledge across different tasks and domains stands out as a pivotal direction. Currently, benchmarks tend to confine models within narrowly defined tasks, as noted in [83]. The capacity to seamlessly adapt knowledge across diverse scenarios underscores an MLLM's versatility and potential for broader real-world deployment.

Innovative frameworks, such as multi-criteria evaluation systems, mark significant strides toward addressing these challenges. These systems appraise models not only based on standard accuracy or precision metrics but also on stability, scalability, and ethical benchmarks, as illustrated in [84]. Such comprehensive assessment mechanisms enable researchers to appreciate multiple facets of MLLM performance, advancing the frontier of multimodal AI capabilities.

Emerging trends also underscore the importance of continuous evaluation systems. These systems, like LMMS-EVAL mentioned in [31], allow for a dynamic feedback loop where models are evaluated in real-time, and adjustments can be made iteratively based on evolving data and application environments. These systems are pivotal in maintaining relevance given the ever-growing complexity and demands of multimodal integrations.

Moreover, synthesizing current understandings in multimodal evaluation reveals gaps and opportunities for refining evaluation practices. This synthesis should focus on aligning model evaluation with real-world conditions, ensuring that tasks and datasets reflect the complexities and unpredictabilities encountered outside lab environments. As explored in [85], a data-driven approach to evaluation can offer vital clues into model behavior under diverse input conditions, enhancing robustness and adaptability.

In conclusion, propelling the evaluation paradigms of MLLMs forward involves integrating multi-faceted evaluation approaches that authentically capture model capabilities and limitations. This requires blending traditional metrics with qualitative insights, ethical considerations, and real-world applicability. Future directions should focus on developing evaluation frameworks that prioritize versatility, fairness, and ethical compliance, paving the way for MLLMs to make compelling advances across a spectrum of applications while maintaining societal values and norms. Through a conscientious and innovative review of evaluation practices, the future of MLLMs can be shaped not only to reflect technical sophistication but to align with a wider vision of responsible and beneficial AI advancement.

5.5 Comparative Analysis of Assessment Frameworks

This subsection delves into the comparative analysis of assessment frameworks used to evaluate Multimodal Large Language Models (MLLMs). The evaluation and benchmarking of these models are crucial for understanding their capabilities and guiding further advancements in the field. Various frameworks have emerged to assess MLLMs, each with distinct approaches, strengths, and limitations. This analysis aims to provide a nuanced understanding of these frameworks, highlighting their ability to capture the full potential of MLLMs while identifying areas for improvement.

Assessment frameworks for MLLMs are designed to evaluate a range of competencies, from basic task performance to more complex capabilities such as reasoning, scalability, and ethical considerations. Frameworks like MM-Vet [13] emphasize the evaluation of multimodal tasks, examining the integration of core vision-language capabilities. These frameworks are structured around evaluating models on a variety of challenging tasks, which provide a comprehensive view of a model's capabilities. Another

example is MMT-Bench, which assesses MLLMs across a wide array of multimodal tasks requiring expert knowledge and deliberate visual reasoning [86].

One of the primary strengths of current assessment frameworks is their comprehensive coverage of task diversity. Frameworks like MMT-Bench encompass a wide range of scenarios, thus allowing researchers to evaluate models in various contexts, which is crucial for understanding the generalizability and applicability of MLLMs across domains [86]. Furthermore, frameworks such as MM-Vet are adept at evaluating the integration of vision and language capabilities, offering insights into how well these models can synthesize information from multiple modalities [13].

Despite their strengths, many current frameworks face limitations regarding scalability and real-time evaluation. Most frameworks are designed to evaluate models with fixed datasets and predetermined tasks, which may not fully capture the dynamic nature of real-world applications. For instance, the static nature of these frameworks often fails to account for evolving data distributions or emerging use cases, thereby limiting their long-term applicability [87]. Additionally, while some frameworks effectively measure task performance, they may not adequately capture ethical dimensions such as safety and privacy, which are becoming increasingly important [88].

Recent trends in assessment frameworks are starting to address some of these limitations by incorporating real-time evaluation techniques and broader criteria. For example, Multimodal LIVEBENCH utilizes continuously updating news and online forums to assess models' generalization abilities in dynamic environments, thus aiming for a low-cost and zero-contamination evaluation approach [87]. Another promising direction is the integration of ethical benchmarks that evaluate MLLMs on parameters like safety and privacy, ensuring that these models align with societal values [88].

A critical technical aspect of these frameworks is the need for robust evaluation metrics that can adapt to different modalities and task requirements. Some frameworks have begun utilizing LLM-based evaluators for open-ended outputs, which can provide a unified scoring metric across different question types and answer styles [13]. These metrics are crucial for fair and comprehensive evaluations, as they allow for consistent comparisons across models. Furthermore, employing advanced evaluation techniques like multi-criteria evaluation frameworks can provide deeper insights into models' stability, robustness, and scalability [13].

In conclusion, current assessment frameworks for MLLMs have made significant strides in evaluating model capabilities across a spectrum of tasks and modalities. However, there remains a need for further development to address limitations related to scalability, ethical considerations, and real-time adaptability. Future research should focus on creating more dynamic and inclusive benchmarks that incorporate continuous updates and ethical evaluations. By doing so, the academic community can ensure that these frameworks not only measure existing capabilities but also guide the development of more advanced and responsible MLLMs. Moreover, expanding the integration of real-world data into evaluation processes will enhance the real-world applicability of these models, ensuring that MLLMs continue to evolve in alignment with societal and technological advancements.

6 Applications and Use Cases

6.1 Healthcare and Medical Applications

The advent of Multimodal Large Language Models (MLLMs) marks a significant leap forward in healthcare applications, garnering notable advancements in diagnostic precision, medical imaging interpretation, and patient-centered interaction systems. These advancements stem from MLLMs' intrinsic ability to process and synthesize complex, multimodal datasets, ranging from detailed textual clinical notes to high-dimensional imaging data. In this subsection, we explore the potentials and challenges associated with the integration of MLLMs into healthcare systems, critically examining the current methodologies deployed and their realized impact.

A primary application of MLLMs in healthcare is the enhancement of medical imaging analysis. Recent studies have illustrated how MLLMs, by integrating visual and linguistic data, have improved diagnostic accuracy and efficiency in radiology and pathology [89]. These models can effectively interpret imaging data alongside clinical reports, facilitating early detection of ailments such as tumors or microfractures, which may not be as easily discerned through unimodal analysis. Technologies such as the Radiology-Llama2 use domain-specific data and instruction tuning to specialize MLLMs for radiological tasks, achieving state-of-the-art performance benchmarks and underscoring the importance of contextual domain expertise in medical applications [89].

The capability of MLLMs to deliver informed medical question-answering (QA) systems has revolutionized the decision-support landscape. These systems, leveraging extensive medical literature and datasets, can offer rapid, evidence-based responses to complex clinical queries [10]. Besides assisting practitioners in clinical decision-making, these models enhance comprehension by translating complex medical jargon into layperson-friendly language, bridging communication gaps between healthcare providers and patients. This integration is pivotal in improving patient engagement and adherence to treatment protocols.

Moreover, MLLMs are advancing personalized patient interaction capabilities by generating adaptive communication strategies tailored to individual patient profiles and input data dynamics. Tools combining patient-generated data with MLLM capabilities show promise in personalizing healthcare delivery, enhancing patient experience, and supporting continued patient care outside conventional clinical settings [90]. For instance, automated systems can monitor and analyze patient health indicators, suggesting lifestyle modifications or alerting healthcare providers about potential health risks.

Despite these advancements, the integration of MLLMs in healthcare is accompanied by significant challenges. The diversity in healthcare datasets—characterized by imbalance, privacy concerns, and modality heterogeneity—affects model training and performance [10]. Additionally, maintaining patient privacy and ensuring data security remain

paramount concerns, given the sensitive nature of healthcare data. Advocated solutions include differential privacy techniques and robust encryption protocols to protect user data, accommodating the ethical and legal stipulations necessary in healthcare [91].

Emerging trends indicate a shift towards more extensive deployment of self-supervised learning paradigms, which can capitalize on the unlabeled nature of large-scale medical datasets, increasing model robustness while mitigating data annotation burdens [14]. Moreover, developments in cross-modal transfer learning promise to enhance the adaptability and effectiveness of MLLMs across varying medical contexts, thereby broadening their usability and application spectrum [5].

In conclusion, while MLLMs possess remarkable potential to transform healthcare systems, the field must surmount intrinsic challenges related to data diversity, privacy, and integration to harness these models fully. Future directions may involve developing more efficient self-supervised learning frameworks, addressing modality-specific challenges, and fostering collaborative efforts between technologists and healthcare professionals to ensure that AI systems meet real-world clinical needs holistically. By leveraging these advanced technologies with caution and foresight, MLLMs can significantly enhance healthcare delivery quality, leading toward more proactive and personalized medicine.

6.2 Autonomous Systems and Robotics

Autonomous systems and robotics are at the cutting edge of technological advancements, with Multimodal Large Language Models (MLLMs) significantly advancing decision-making and interaction capabilities within these domains. These models have a unique ability to process and integrate diverse data modalities such as visual, linguistic, and sensory inputs, forming the bridge between comprehensive environmental understanding and robust autonomous operation.

MLLMs' impact on autonomous systems is primarily rooted in their proficiency in interpreting multimodal data, which enables robotics to achieve a nuanced understanding of environments. Models like Perceiver and similar architectures enhance an autonomous system's capacity to process high-dimensional inputs from varied modalities, which are essential in tasks such as autonomous driving and human-robot interaction [77]. Through its asymmetric attention mechanism, the Perceiver efficiently scales to handle large inputs by distilling information into a compact latent space, aligning perfectly with the real-time processing needs of autonomous systems.

The integration of MLLMs into robotic systems has extended their applicability. In autonomous driving, MLLMs play a crucial role in enabling advanced natural language processing abilities for navigation and decision-making. These systems can interpret and respond to verbal instructions from passengers, allowing them to adapt dynamically to complex driving situations [92]. The use of models like RoboMamba, which utilize state-space architectures for computational efficiency and effective reasoning, has significantly bolstered real-time processing and adaptability, key in the ever-changing environments faced by autonomous vehicles [93].

Additionally, MLLMs enhance human-robot interaction (HRI) by facilitating more intuitive interfaces that transcend basic command-response interactions. With novel architectures such as MoE-LLaVA, utilizing sparse mixture-of-experts frameworks, MLLMs manage multimodal inputs effectively and provide contextually aware, natural dialogues between humans and robots [47]. This enhances robots' ability to execute complex, multi-step instructions, thus broadening their roles from home assistance robots to sophisticated industrial automation. By leveraging extensive language and vision datasets, these models refine interaction capabilities, fostering deeper and more effective HRI [25].

Nevertheless, challenges remain. The need for real-time application demands continuous optimization of MLLM computational efficiency. Innovations such as FlashConv, which boost the execution speed of state-space models, are crucial for overcoming computational constraints during model inference and deployment [94]. Additionally, ensuring robustness against environmental variability and noise is critical, particularly for outdoor settings or unpredictable scenarios in driverless vehicle operations. MLLMs need to evolve with more resilient mechanisms that maintain functionality across diverse conditions.

Looking ahead, the integration of MLLMs in autonomous systems is poised for groundbreaking advancements across cross-disciplinary domains. Emerging models like MoAI, combining specialized vision and language capabilities, could redefine the scope of autonomous applications [95]. Future research should focus on enhancing these models' interoperability with specialized sensory systems, promoting greater autonomy and precision in task execution.

The trajectory of MLLMs in autonomous systems is set for significant growth as innovations in architectures and training methodologies unfold. This evolution promises not only to optimize current robotic functionalities but also to unlock new possibilities in various fields, including logistics and exploration. Ultimately, the synergy between MLLMs and autonomous systems heralds a transformative era in intelligent automation, connecting seamlessly with the preceding advancements in healthcare and paving the way for further innovations in assistive technologies.

6.3 Assistive Technologies

Multimodal Large Language Models (MLLMs) are poised to make significant strides in assistive technologies, offering enhanced communication interfaces and heightened environmental interaction for individuals with disabilities. By amalgamating data from multiple modalities such as text, audio, image, and sensors, these models extend the boundaries of traditional assistive devices, allowing for deeper understanding and interaction with user environments. This section delves into various facets of MLLM-driven assistive devices, providing a comparative analysis of their approaches, strengths, limitations, and emergent trends.

An essential aspect of MLLMs in assistive technology is their ability to facilitate more nuanced and effective

communication for individuals facing speech or mobility challenges. Speech and language interfaces augmented by MLLMs can provide real-time translation between spoken language and text, offering seamless integration into existing communication systems. This is particularly beneficial for users with limited verbal communication abilities, allowing them to interact with digital interfaces using either enhanced voice control or textual input. The capability for real-time responsiveness also ensures that these systems can adapt dynamically to the variability inherent in human communication [2], [55].

The integration of vision-language models can significantly amplify the capabilities of assistive devices tailored to users with visual impairments. For example, models like ImageBind, which extend modality interactions to novel sensory inputs, allow users to receive audio descriptions of their surroundings, thereby enhancing spatial awareness. This can facilitate greater independence for visually impaired users in executing daily tasks such as navigation [54].

Another burgeoning application of MLLMs in assistive technologies includes intelligent prosthetics. By combining data from multiple sensors, these prosthetics can adjust to varying environmental conditions, providing a more intuitive and responsive interaction that mimics natural limb movement. The integration of multimodal inputs from visual, auditory, and tactile sensors ensures that these prosthetics can dynamically respond to both the user's commands and the surroundings, improving autonomy and user satisfaction [76].

The inherent challenge of incorporating MLLMs into assistive technology lies in creating models that can navigate the noise and variation typical of real-world environments. The robustness and fine-grained adaptability of these models are crucial for maintaining performance in unstructured settings. Research into context-based multimodal fusion illustrates promising directions, employing contrastive alignment strategies to maintain performance even when input modalities are partially missing or misaligned. This aligns well with practical requirements for devices like hearing aids, which must operate efficiently in various acoustic environments [28], [91].

However, the adoption of MLLMs in assistive technologies is not without limitations. Computational demands and resource intensity present challenges in deploying these models on lightweight or portable devices such as smartphones or tablets. Emerging trends address these concerns by optimizing pre-training and fine-tuning processes, exemplified by methods such as Parameters-Efficient and Scalable Multimodal Fusion via Mixture of Prompt Experts, which achieve significant reductions in computational resource requirements while maintaining high performance [56].

Looking forward, the future of assistive technologies driven by MLLMs could pivot towards integrating more diverse data modalities, like physiological signals, to provide even richer insights and enhance user interactions. Developments could further explore adaptive learning frameworks that continuously improve personalization by learning from user feedback. This avenue, combined with privacy-preserving techniques, should focus on ensuring safe data handling, crucial for sensitive personal data common in assistive technology applications [29], [44].

In summary, MLLMs possess the potential to fundamentally transform assistive technologies, fostering greater independence and improved quality of life for individuals with disabilities. Despite obstacles in computational efficiency and model robustness, ongoing advancements promise to address these challenges. By enhancing interaction capabilities and maintaining user-centric designs, MLLMs stand as the cornerstone for future innovations in assistive technology, paving the way for more accessible and inclusive solutions.

6.4 Content Creation and Multimedia Generation

Multimodal Large Language Models (MLLMs) have ushered in a new era in content creation and multimedia generation, revolutionizing how creative industries and digital media synthesize diverse outputs from intricate multimodal inputs. The seamless integration of text, audio, visual, and other modalities underpins the crafting of compelling and dynamic content, standing as a testament to the advances discussed in assistive technology and education. This subsection critically analyzes the current state of MLLMs in content creation, evaluates various approaches, and explores future prospects and challenges, connecting these innovations with prior developments.

Key advancements in content creation have been driven by models such as Perceiver, which utilize asymmetric attention mechanisms to transform large multimodal inputs into coherent outputs while preserving semantic integrity [77]. Meanwhile, Unified-IO 2 demonstrates how diverse data types can be transposed into a unified semantic space, allowing a single model framework to process and generate across modalities efficiently [96]. These technological breakthroughs align with the trend of refining communication interfaces in assistive technology, showcasing the extensive capabilities of MLLMs.

The strengths of models like PaLM-E lie in their dynamic responsiveness to multimodal inputs, which is particularly valuable in interactive entertainment applications requiring real-time adaptability [97]. Despite these strengths, challenges such as scalability and resource intensity persist. High computational demands are prohibitive, although efforts like the Cobra model address efficiency through linear computational complexity approaches [60]. These challenges echo the computational considerations highlighted in the educational domain.

In the realm of multimedia generation, autoregressive models like VisionLLM enhance MLLM capabilities by aligning vision-centric tasks with language instructions, allowing for task customization tailored to specific creative needs [98]. However, reliance on manually curated datasets for training MLLMs can introduce biases and limit content diversity, a concern also prevalent in education. This issue is explored in Quantifying & Modeling Multimodal Interactions, which delves into redundancies and synergies in multimodal interactions [44].

Recent trends focus on enhancing interactive experiences, as illustrated by the MM-REACT system, which integrates general-purpose models like ChatGPT with spe-

cialized vision experts to achieve complex multimodal reasoning [99]. This approach parallels the educational developments of adaptive learning tools and virtual tutors. Despite these advances, the alignment of diverse modality training and inference remains a critical research focus. Robust benchmark frameworks, such as VALSE, which evaluates visio-linguistic grounding capabilities, will be key in assessing advancements and guiding future improvements [100].

In conclusion, while multimodal LLMs underscore transformative potential in content creation and multimedia generation, ongoing research is required to address challenges in scalability, modality alignment, and bias mitigation. Future directions might involve developing more sustainable and efficient training techniques, improving ethical standards, and exploring interdisciplinary applications in creative domains. As these systems evolve, collaboration across fields will be essential to harness their full creative potential while addressing the ethical and societal implications of their deployment, setting the stage for their impactful integration into educational settings and beyond.

6.5 Education and Skill Development

In recent years, Multimodal Large Language Models (MLLMs) have revolutionized educational paradigms by offering highly personalized and interactive learning environments that seamlessly integrate a wide array of data modalities. These models hold the promise of tailoring educational experiences to meet diverse learning needs, enhancing comprehension, skill acquisition, and retention. This subsection delves into the applications of MLLMs in education and skill development, exploring their methodological innovations, evaluating their efficacy, and identifying future research directions.

At the core of MLLMs' impact on education is their ability to create virtual tutors and assistants that adapt to students' individual learning styles and pace. Leveraging multimodal datasets that combine text, audio, visual, and tactile information, these systems provide personalized feedback and context-rich explanations, enhancing students' understanding across complex subjects [68]. By analyzing each student's interaction with the model, MLLMs can dynamically adjust instructional content, offering a scaffolded learning approach that supports incremental skill acquisition [13].

Comparative analyses of various approaches highlight the unique strengths of MLLMs in facilitating skill training and simulation. By integrating realistic visual, auditory, and kinesthetic cues within simulations, MLLMs offer immersive environments where learners can practice tasks that mimic real-world scenarios. This capability is especially valuable in disciplines requiring hands-on practice, such as medicine, where virtual dissections or diagnostic practices can be simulated. Unlike traditional didactic methods, these simulations utilize interactive 3D models and haptic feedback for skill refinement, enabling learners to experiment with complex systems without the resource constraints or safety concerns of physical experimentation [4].

Technical innovations such as dynamic modality fusion further optimize the educational applications of MLLMs by selectively prioritizing the most relevant inputs based on learning contexts [81]. This adaptability is crucial for environments where diverse input types must be processed simultaneously, such as integrating video tutorials with real-time problem-solving in mathematics or physics. Despite these advantages, challenges like ensuring equitable access to technology and mitigating cognitive overload remain. It's pivotal that educational tools balance the cognitive load by presenting information in digestible segments, which has been shown to facilitate better retention and understanding [25].

Emerging trends indicate a growing focus on seamlessly integrating MLLMs with existing educational technologies, such as learning management systems (LMS), to streamline user experiences and broaden reach. By embedding MLLMs within LMS platforms, institutions can harness their potential to drive engagement through features like automated assessment and real-time feedback, which can tailor learning at scale and provide insights into student progress across diverse cohorts [101].

However, the adoption of MLLMs in education comes with its trade-offs, particularly concerning data privacy and model bias. Given that these models rely on vast datasets, safeguarding students' privacy through robust data anonymization and encryption protocols is imperative. Furthermore, detecting and mitigating biases embedded in training datasets is essential to providing fair educational experiences [4].

As we look to the future, multidisciplinary research initiatives should focus on refining the accuracy and inclusivity of MLLMs in educational settings. Advancements in multimodal interaction mechanisms, such as haptic technology and augmented reality (AR), hold the potential to further deepen engagement by bridging the gap between virtual and physical learning experiences [17]. Additionally, fostering cross-disciplinary collaborations can spearhead the development of tailored educational tools that meet specific learning objectives across various fields, from language acquisition to STEM education [36].

In conclusion, the continued evolution of MLLMs in education represents a transformative trend with far-reaching implications for learners across the globe. Through careful consideration of technical, ethical, and pedagogical factors, MLLMs can provide inclusive, engaging, and effective educational experiences that empower learners with the skills required for a rapidly changing world.

7 Ethical Considerations and Societal Impacts

7.1 Bias and Fairness in Multimodal Models

Multimodal Large Language Models (MLLMs) are increasingly central to AI applications, providing sophisticated interactions across diverse data types such as text, images, and sounds. As these models permeate critical areas like healthcare, legal judgments, and content recommendations, addressing inherent biases and ensuring fairness become crucial to their ethical deployment. This subsection delves into the challenges of bias within MLLMs, detailing strategies for detection and mitigation while highlighting the profound societal implications of biased outcomes.

Multimodal models are particularly susceptible to biases due to the heterogeneity and complexity of the data they process. Bias often arises from imbalanced datasets, where over-represented demographic groups in training data skew model predictions. Such biases can lead to unfair treatment, amplifying stereotypes or reinforcing societal inequities through discriminatory outcomes in hiring processes, law enforcement predictive models, and beyond. The potential for biased results within these domains underscores the urgent need for comprehensive detection and mitigation strategies to ensure equitable model behavior [17].

One effective approach to identifying bias within MLLMs is through the use of diagnostic tools and evaluation frameworks that uncover discrepancies in model outputs across different demographic groups. Techniques such as sensitivity analysis and counterfactual evaluations can provide insights into how models might shift predictions based on changes in input attributes indicative of social identity markers like race or gender. However, these methodologies often require controlled and extensive preprocessing of data to be effective, which can limit their applicability in real-time deployments [3].

To address these biases, several mitigation strategies have been proposed, ranging from data-centric methods to sophisticated algorithmic approaches. Data augmentation techniques, which involve synthetically balancing datasets by generating under-represented examples, have shown promise in addressing bias at the data level. These methods work well within the scope of multi-modal learning, enhancing fairness by ensuring that all demographic groups are equitably represented in the training set [64]. Algorithmic fairness techniques like adversarial debiasing and fairness constraints attempt to directly incorporate fairness objectives into the learning process. By penalizing unfair predictions during training, these approaches strive towards models that not only perform well but also adhere to fairness guidelines across their outputs [5].

However, these strategies have their limitations. Data augmentation, while effective, may not fully capture the complexities and nuances of real-world biases. Additionally, adversarial debiasing can lead to trade-offs between model accuracy and fairness, necessitating careful balancing to avoid compromising model effectiveness. These trade-offs represent one of the principal challenges in deploying unbiased MLLMs. Furthermore, the absence of universally accepted metrics for bias and fairness complicates the comparison and assessment of different debiasing techniques, making it difficult to standardize practices across applications and domains [69].

Emerging trends in addressing bias involve leveraging advances in interpretability and explainability within MLLMs. By demystifying the decision-making processes of AI models, researchers aim to identify sources of bias more effectively and foster trust among users. Initiatives like Explainable AI (XAI) offer frameworks for dissecting model decisions, thereby providing insights into how models weigh different modalities and features, which can reveal potential biases embedded in the model architecture or data [18].

Moreover, fostering fairness in MLLMs requires a proactive and continuous approach beyond technical solutions. Ethical guidelines and policies should be established, emphasizing accountability at each stage of the model's lifecycle—from data collection through to deployment. Involving stakeholders from diverse backgrounds in the development process can also provide necessary perspectives to preemptively address potential biases [102].

In conclusion, while considerable advances have been made in understanding and mitigating biases in MLLMs, substantial work remains. Future research should focus on developing robust, standardized benchmarks for bias detection and fairness assessment. The integration of interpretability techniques with ethical frameworks presents a promising avenue to enhance the fairness and societal acceptance of MLLMs. As these models continue to evolve and increasingly integrate into society, ensuring fairness not only reinforces the technical efficacy of MLLMs but also strengthens their societal impact, aligning technological progress with human-centric values.

7.2 Privacy and Security in Multimodal Systems

Multimodal Large Language Models (MLLMs) are at the forefront of AI development, as they process and integrate diverse modalities including text, images, audio, and potentially other forms of data. This multiplicity of inputs brings about pressing privacy and security concerns, stemming from the complexity and sensitivity of collected multimodal data. This subsection delves into the multifaceted privacy and security implications posed by MLLMs, evaluates current protection strategies, and highlights future research directions to safeguard user information cohesively within the broader context of bias mitigation and ethical deployment.

The handling of vast volumes of heterogeneous data lies at the core of privacy risks in multimodal systems. These data often encompass personal, sensitive, or proprietary elements, and the risk of inadvertent data leakage or memorization is heightened within MLLMs, given how they integrate multiple modalities. This integration can uncover intricate behavioral and contextual insights that might not be apparent when data is considered in isolation. Recent investigations have illustrated these vulnerabilities, showing that sensitive training data can inadvertently be recalled and exposed. To mitigate such risks, advancements in differential privacy, encryption techniques, and stringent data governance frameworks have been suggested, aligning with efforts to achieve fairness and transparency in model deployment [25].

From a security standpoint, the cross-modal functionalities of MLLMs present unique challenges. Integrating visual, auditory, and textual data opens up potential novel attack vectors, such as modality-specific hacks exemplified by adversarial attacks on image inputs or audio perturbations targeting model weaknesses [20]. Research has identified emerging threats where the conversion of information from one modality to another, like generating text from images, could risk unauthorized disclosure of sensitive data, underlining the requirement for comprehensive security strategies.

Addressing these concerns involves incorporating robust security architectures during both the training and deployment phases of MLLMs. Techniques such as adversarial

training and robust optimization have been employed to fortify models against known security threats, although these approaches must continuously evolve to respond to the emergent complexities presented by multimodal data [50]. Additionally, employing privacy-preserving mechanisms like homomorphic encryption and federated learning can significantly diminish the risks related to data centralization and unauthorized access [71].

Emerging trends point to a shift toward more sophisticated privacy- and security-aware MLLM frameworks. Future research is expected to focus on contextual privacy, ensuring user data is protected not only at rest or in transit but also within the operational realm of the MLLM itself, safeguarding data confidentiality in outputs. Techniques such as secure multiparty computation and zero-knowledge proofs might find greater adoption, offering cryptographic assurances with minimal overhead.

It is imperative to synchronize the development of multimodal systems with comprehensive legal and ethical standards that enhance transparency and accountability, complementing ongoing discussions on ethical governance. Policy frameworks should mandate the integration of privacy-by-design principles and necessitate routine audits to confirm compliance with data protection regulations. Beyond legal imperatives, industry-driven standards might emerge as influential mechanisms to standardize secure practices across multimodal AI implementations, establishing baselines for model audits and data requirement assessments [44].

In conclusion, while MLLMs promise transformative advancements, their associated privacy and security challenges demand serious attention within the ethical and biased context of their deployment. Addressing these concerns requires a comprehensive approach, incorporating technological innovations, regulatory measures, and ethical guidelines to ensure trust in these powerful models. Future research should prioritize refining existing techniques, exploring novel cryptographic applications, and establishing integrated frameworks for privacy and security in MLLMs, enhancing the safe and responsible deployment of multimodal technologies.

7.3 Ethical Frameworks and Policy Implications

In the burgeoning field of Multimodal Large Language Models (MLLMs), the integration of different modalities such as text, images, audio, and more has rapidly expanded the functional capabilities of artificial intelligence systems. However, this technological advance brings to the forefront profound ethical considerations and necessitates robust policy frameworks to govern its implementation. This subsection aims to dissect these ethical challenges and propose pathways for policy development that can ensure responsible deployment of MLLMs.

A primary concern within the ethical domain is the alignment of MLLMs with existing societal norms and ethical principles. Ethical frameworks are essential to guide the development and deployment of MLLMs in a manner that conforms to universal human rights and public interest. These frameworks should encompass guidelines for fairness, transparency, accountability, and privacy. Key to this is the construction of ethical guidelines that balance innovation with ethical integrity, ensuring that MLLMs function within societal norms without stifling technological progress.

Ethical frameworks must address the issue of bias, which inherently poses a risk when models are trained on large datasets with historical biases. These biases can manifest in the output of MLLMs, leading to skewed decisions that reinforce stereotypes or discriminatory practices. For instance, models might generate biased responses when tasked with applications across socioeconomically disadvantaged demographics or diverse cultural backgrounds. Strategies such as data auditing, bias detection algorithms, and inclusive data representation are crucial in mitigating these concerns [2]. Moreover, emerging models propose adversarial training techniques and diversity-aware design frameworks that aim to cultivate equitable model outcomes [55].

A critical element of ethical governance is data privacy. MLLMs often require vast amounts of multimodal data, raising concerns around the safeguarding of personal information. The privacy risks associated with data collection and use include unauthorized data sharing and breaches, as well as the potential misuse of personal data insights retrieved from multimodal inputs. To counteract these risks, ethical frameworks must incorporate strong data privacy protocols, such as encrypted data storage, consent management systems, and differential privacy approaches to protect user data. These technical measures must be reinforced by policy frameworks that mandate compliance with data protection regulations like GDPR and CCPA, ensuring that MLLMs operate within legal boundaries [5].

Emerging trends suggest a growing need for cross-disciplinary policies that integrate technical remedies with legal structures. This integration is critical in areas prone to ethical dilemmas, such as surveillance and autonomous decision-making, where policy vacuums can result in unintended consequences. Multimodal systems, when deployed in law enforcement or healthcare, require stringent ethical review processes to delineate the boundaries of acceptable use. This necessitates a collaborative approach involving technologists, ethicists, policymakers, and end-users in the design and implementation of MLLMs to ensure they align with ethical standards and public welfare [103].

Furthermore, the concept of ethical governance extends to international cooperation in defining and enforcing global standards for MLLM deployment. This involves harmonizing disparate regulations and standards across jurisdictions, fostering transparency, and adopting accountability measures. Establishing international bodies to oversee and coordinate these efforts can mitigate discrepancies and promote a unified approach to ethical compliance [1].

Future directions in ethical frameworks and policy implications for MLLMs involve advancing the understanding of ethical AI through continuous research and development of dynamic ethical standards that can adapt to emerging challenges. This includes developing new evaluation metrics that explicitly assess ethical compliance and societal impact, extending beyond traditional performance metrics focused on accuracy and efficiency [5].

In conclusion, the ethical deployment of MLLMs hinges on the development of comprehensive ethical frameworks

and policy adaptations that prioritize societal well-being without compromising on innovation. By fostering a multi-stakeholder dialogue and incorporating ethical standards into the core of MLLM technologies, it is possible to safeguard against potential risks while amplifying the societal benefits of multimodal intelligence. This balanced approach ensures that the evolution of MLLMs will be beneficial, equitable, and aligned with human-centric values.

7.4 Societal Impacts and Responsible AI Use

The integration of Multimodal Large Language Models (MLLMs) into society marks a significant shift in the landscape of artificial intelligence, offering unprecedented capabilities while simultaneously raising critical ethical and societal concerns. Building upon the discussion on ethical considerations, this subsection delves into the broader societal impacts of MLLMs and underscores the importance of responsible AI practices to both mitigate potential negative repercussions and harness societal benefits.

MLLMs, with their capacity to analyze and synthesize diverse modalities such as text, images, and audio, herald advancements across multiple sectors, including healthcare, education, and entertainment [24], [104]. For instance, in healthcare, these models boost diagnostic precision through the amalgamation of medical imaging and textual analysis, potentially revolutionizing patient care. However, their deployment raises intricate ethical issues, as discussed earlier, particularly around bias, privacy, and environmental sustainability.

A primary societal concern revolves around the impact of MLLMs on labor markets and economic structures. The automation potential of these models could transform various job sectors, especially those reliant on data processing and basic analysis, thus accelerating job displacement. The challenge lies in balancing technological progress with socio-economic stability, ensuring that the workforce evolves through reskilling and education initiatives to meet the new demands created by these technologies [105].

Moreover, the environmental impact of MLLMs is significant and cannot be overlooked. The energy-intensive nature of training extensive models contributes notably to carbon emissions, highlighting sustainability issues in AI research [31]. Innovative approaches are essential to enhance model efficiency and reduce resource consumption, such as developing more energy-efficient architectures and exploring alternative training paradigms [31].

Promoting responsible AI practices involves designing MLLMs that are not only technologically advanced but also ethically sound. This entails adopting frameworks that prioritize transparency, accountability, and inclusivity in AI development. Ensuring diverse data representation during model training can aid in mitigating biases, thereby promoting fairness in AI outcomes. Additionally, establishing safety and ethical benchmarks is crucial for evaluating the societal impact of these models, ensuring they function within acceptable ethical boundaries [106].

In aiming for positive societal outcomes, MLLMs could be leveraged to tackle global challenges, such as enhancing accessibility for individuals with disabilities and refining disaster response mechanisms through improved data analysis and interpretation. Furthermore, integrating MLLMs into public safety and emergency response protocols can heighten situational awareness and decision-making accuracy, potentially saving lives and conserving resources [105].

Despite these advantages, the path to comprehensive and ethical deployment is fraught with hurdles. Emerging trends underscore the necessity for interdisciplinary approaches that merge technical expertise with ethical scholarship to address these complex issues. Cross-disciplinary collaborations, involving AI ethicists, computer scientists, and sociologists, could yield novel insights for designing AI systems resilient against ethical pitfalls, such as bias and discrimination [98].

In conclusion, the societal impacts of MLLMs are profound and multifaceted, requiring a balanced strategy that maximizes benefits while minimizing risks. As ethical governance underscores, responsible AI practices are pivotal in ensuring these technologies contribute positively to society, fostering a future where AI not only complements human efforts but also adheres to the highest ethical standards. Future research should focus on enhancing the robustness of MLLMs against ethical challenges and pursuing sustainable practices in AI development, paving the way for more socially responsible and impactful technologies.

8 Future Directions and Emerging Trends

8.1 Scalability Enhancements

The burgeoning field of Multimodal Large Language Models (MLLMs) presents both an opportunity and a challenge in scaling these models to efficiently manage increasing data volumes and complexity. This subsection delves into the potential scalability enhancements for MLLMs, emphasizing computational efficiency and model capacity expansion.

The primary challenge in scaling MLLMs lies in their inherent complexity due to the integration of multiple modalities, each requiring distinct processing pathways and ample data representation capacity. As MLLMs grow to accommodate diverse input types — from textual and visual data to auditory and beyond — the computational requirements exponentially increase. This necessitates novel approaches to make these models both scalable and efficient.

Existing literature on scalability improvements for MLLMs offers several compelling approaches. One such approach is the adoption of low-rank and sparse techniques, which, by reducing the dimensionality of matrices involved in model operations, can lead to significant computational savings while maintaining the overall performance of the model. These techniques can be particularly effective when integrated into the model's core architecture, facilitating scalability without sacrificing accuracy or utility [14].

Additionally, mixture of experts (MoE) architectures provides a promising avenue for scalability by dynamically selecting specific subnetworks to process sections of the input, thereby effectively utilizing model parameters only where necessary. This paradigm aids in managing computational loads and allows the model to scale efficiently with increasing data inputs. The MoE strategy further benefits from its capability to train sub-models on distinct modalities separately, allowing for tuned specialization and thereby

enhancing cross-modal learning capabilities, although challenges remain in optimizing the gating mechanisms that decide the allocation of computational resources [4].

Another crucial area for enhancement is within the training regimes of MLLMs. Utilizing techniques such as curriculum learning, where models are gradually exposed to increasingly complex multimodal data, can mitigate training hurdles, reducing computational strain and facilitating smoother scaling. Further, employing contrastive learning techniques allows for more robust representations by learning from the natural congruence present in multimodal datasets, intrinsically enhancing scalability by leveraging shared information across modalities [5].

Beyond architectural and training innovations, an important direction for scalability lies in the optimization of data pipeline processes. Efficient data processing at the input level, utilizing advanced pre-processing techniques and data normalization methods, can substantially decrease the processing burden of MLLMs. Concurrently, exploring distributed training environments, which partition data across multiple processors effectively, presents a viable strategy to handle large-scale data input without overwhelming system resources [107].

Lastly, a promising and emerging area involves the exploration of novel hardware and software co-designs that are optimized for multimodal operations. These systems can potentially address the power-guzzling tendencies of traditional hardware installations by incorporating task-specific accelerators, such as Graphical Processing Units (GPUs) and Tensor Processing Units (TPUs), which are engineered to handle the intricate task of multimodal data integration more efficiently.

Despite strides in these aforementioned areas, the scalability of MLLMs continues to be encumbered by the challenges of maintaining model accuracy and integrity under reduced computational constraints. Forward-looking research must continue to balance these scalability concerns with the pivotal demands of maintaining model robustness and preserving cross-modal interaction insights. Practical innovations in architectural design, computational optimization, and even data-centric methodologies will play a seminal role in overcoming these hurdles. Efforts to establish a coherent framework for scalability that aligns with the expansive capabilities of MLLM will be essential in ensuring these models can reach their potential in real-world applications without impractical resource consumption [108].

In conclusion, addressing scalability in MLLMs requires a multi-faceted approach that encompasses not only architectural and algorithmic developments but also innovative strategies for data management and hardware utilization. The journey towards scalable, efficient MLLMs will undoubtedly be pivotal in unlocking the potential of these models in diverse, real-world applications, underscoring their transformative capability in the realm of intelligence augmentation across industries.

8.2 Cross-Modal Interaction and Integration

The realm of Multimodal Large Language Models (MLLMs) stands at the forefront of innovation, thanks to the significant advancements in cross-modal interaction and integration techniques, which are pivotal to optimizing the synergy among diverse data modalities such as text, vision, and audio. The ability to seamlessly integrate and interact across these modalities greatly enhances model performance and efficacy across tasks, setting the stage for more complex real-world applications. This subsection delves into the cutting-edge methods that facilitate cross-modal interaction and integration, examines their applications, and spotlights emerging trends that are reshaping this field.

Recent technological strides have pushed beyond the confines of traditional early and late fusion techniques, forging paths toward more dynamic and adaptive mechanisms for integrating multimodal data. The emergence of models like LIMoE and MoVA highlights a transformative leap toward employing sophisticated gating functions and task-specific expert routing strategies. These methods are designed to handle modality-specific features adeptly, enabling the model to adaptively draw on relevant information based on the contextual demands [46], [109]. By moving past the static constraints of conventional fusion methodologies, these approaches craft more nuanced cross-modal representations.

Central to effective multimodal integration is cross-modal alignment, a critical factor in ensuring consistent representation across varied input forms. The Multimodal Transformer (MulT) is an exemplary model, utilizing a directional pairwise cross-modal attention mechanism to facilitate simultaneous interaction between non-aligned sequences. This approach fosters latency adaptation between modalities, thereby refining data interpretation even when they are not temporally synchronized [20]. Such mechanisms ensure that despite temporal discrepancies, the model adeptly extracts and correlates key features across different modalities.

Moreover, strides have been made in conceiving methods that bolster the robustness and efficacy of cross-modal interactions. The development of joint embedding spaces, such as those explored in the Perceiver model, represents a significant milestone. These spaces allow for the mapping of modalities into a unified latent space, minimizing dependency on modality-specific assumptions. The use of an asymmetric attention paradigm enables transformers to effectively manage diverse modalities, achieving considerable scalability [77].

Despite notable progress, challenges persist in cross-modal interaction methodologies. Variability in performance across different datasets, as seen in approaches like FuseMoE, underscores the need for models with greater generalizability that maintain high performance irrespective of data patterns and quality [110]. Additionally, the amplification of modality-specific biases in multimodal contexts necessitates the development of techniques to ensure fairness and equity in model predictions.

Emerging trends continue to expand the horizons of cross-modal integration. The fusion of multimodal models with reinforcement learning frameworks, illustrated by platforms like LAMM, is paving the way for enhanced agency and decision-making capabilities, potentially fostering more autonomous and contextually aware interactions [111]. Furthermore, the exploration of meta-learning strategies and insights from neuromorphic computing is inspiring new architectures that seek to emulate human-level cross-modal

data processing and fusion.

The advancements in cross-modal interaction and integration within MLLMs signal a promising trajectory for the future. Ongoing research is set to tackle existing technical limitations, propelling models that are not only more efficient and capable but also equitable in their outputs. Future directions may see the refinement of task-specific expert systems that dynamically calibrate interaction strategies in response to environmental and input complexity. Moreover, redefining multimodal integration protocols with privacy-preserving measures becomes imperative as these models increasingly permeate real-world applications where data privacy and security are critical concerns. These continuous efforts are anticipated to revolutionize the way MLLMs address and solve complex challenges, thereby amplifying their impact across various industries.

8.3 Efficient Training and Inference Solutions

The advent of Multimodal Large Language Models (MLLMs) has revolutionized how we process and integrate diverse data types from text, audio, and visual inputs to generate more insightful outputs. Yet, as these models become increasingly intricate, the quest for training and inference efficiency escalates, requiring innovative techniques to manage the extensive computational demands without compromising on performance. Thus, this subsection delves into the strategies aimed at achieving efficient training and inference solutions for MLLMs.

Efficient training methodologies often begin with optimizing the computational overheads without losing fidelity in model performance. One prominent approach is Parameter-Efficient Fine-Tuning (PEFT), where the emphasis lies in selectively updating only a subset of all model parameters across tasks, thus significantly reducing the computational burden. Techniques like Tuning LayerNorm stand out in this domain, allowing modifications that enhance training efficiency by optimizing essential components, achieving the dual goals of economical computation and maintaining model precision [112].

Furthermore, the introduction of novel bundle-training frameworks, like Mixture of Prompt Experts (MoPE), offers another layer of flexibility. MoPE disentangles traditional prompting mechanisms and adaptively routes data to the most promising experts within a network, thereby enhancing both adaptivity and expressiveness [56]. Unlike traditional models that require extensive fine-tuning of all layers, MoPE utilizes fewer parameters and achieves substantial improvements in the fusion of multimodal data, which is particularly useful in resource-constrained setups.

On the inference side, techniques such as Visual Tokens Withdrawal (VTW) represent a paradigm shift towards rapid processing capabilities. By intelligently identifying and skipping redundant tokens during inference phases, models can preserve computational resources while ensuring sustained high performance levels. This strategy not only decreases latency but also significantly slashes the energy consumption typically associated with running sophisticated models in real-time applications [113].

Moreover, the fusion of multimodal inputs even during inference can be greatly enhanced by efficient architecture designs like X-VILA, which aligns multiple modality-specific encoders effectively with large language model inputs. This facilitation of cross-modality unification enhances the robustness of models against changes in data type while retaining a simplistic yet powerful interface [114]. Such architectures demonstrate the potential for reducing computational load by leveraging pre-trained models and focusing on essential alignments without necessitating redesigns from scratch.

While these methodologies provide promising pathways, they come with their own set of challenges. For example, PEFT techniques sometimes risk underfitting in highly complex models due to overly aggressive pruning of parameters. Meanwhile, fusion methods like MoPE may encounter difficulties in correctly routing data when facing unprecedented combinations of modalities [2]. Inference strategies such as VTW must balance between token withdrawal and performance preservation, requiring intricate calibration depending on application demands [113].

The future of efficient MLLM training and inference is likely to witness transformative developments in hybrid models that combine the computational efficiency of sparse networks and the generalization capability of dense architectures. As researchers refine these approaches, the scope for scalable and adaptable model architectures will expand, prompting new applications in domains that were previously considered computationally prohibitive. Indeed, by continuing to explore and refine efficient modeling techniques, the field stands poised not only to alleviate the computational burdens but also to enhance the versatility and accessibility of MLLMs across various multidisciplinary applications.

In summarizing, the pathway forward for MLLMs will necessitate a balanced approach incorporating both cutting-edge technical solutions and innovative architectural designs. The aim is not merely to conserve computational resources but to enable these models to operate effectively across diverse scenarios, underscoring the growing imperative for cross-modal efficiency as a critical determinant of future advances in artificial intelligence.

8.4 Ethical and Societal Implications

As the field of Multimodal Large Language Models (MLLMs) progresses, addressing their ethical and societal effects becomes increasingly pertinent. Given their proficiency in integrating diverse data types—text, audio, and images—MLLMs present significant ethical challenges, including bias, fairness, privacy, and security. This section delves into these issues, examines current approaches to tackle them, and anticipates future directions to ensure ethical MLLM deployment.

Bias is a critical concern, as MLLMs can inherit and amplify biases prevalent in their training data, affecting gender, race, ethnicity, and socioeconomic status. Such biases can skew outputs across applications like recruitment, legal systems, and healthcare diagnostics [115]. Addressing this requires implementing effective bias detection and mitigation strategies, such as data diversification, counterfactual data generation, and algorithmic techniques like constrained optimization and fairness constraints. However,

these methods often struggle with the fairness-performance trade-off, necessitating more advanced solutions [17], [85].

Privacy concerns stem from MLLMs handling sensitive multimodal data, which could inadvertently compromise user information. The models' ability to process and infer meaning across modalities elevates the risk of privacy breaches, especially when models memorize sensitive data from training. Techniques like differential privacy, federated learning, and data anonymization offer viable yet performance-impacting solutions [11]. As MLLMs gain traction in various sectors, ensuring strong data encryption and secure model frameworks is imperative.

From a broader perspective, MLLM deployment influences labor markets and environmental sustainability. Automating tasks typically performed by humans could exacerbate job displacement, particularly in language and vision processing roles. Conversely, the energy demands for training and deploying large MLLMs pose environmental sustainability challenges [31]. Addressing these issues requires balancing ethical AI principles with sustainable practices, emphasizing energy-efficient architectures and training strategies [60].

Future research on MLLMs' societal and ethical implications should focus on fostering cross-disciplinary collaboration for comprehensive ethical guidelines and frameworks. Enhanced public policy and regulatory measures could align MLLM development with societal values, ensuring accountability and transparency through stakeholder engagement in shaping policies on data privacy, model fairness, and AI transparency [106].

Emerging trends suggest incorporating ethical considerations into MLLM design and deployment from the outset. Key components are developing techniques and tools that automatically detect and mitigate biases during model training and application, fostering trust and reliability in MLLMs [17]. Additionally, promoting open, responsibly-sourced data and enforcing ethical evaluation standards can guide responsible advancements in this domain [96].

In conclusion, while MLLMs possess transformative potential, their ethical and societal implications demand careful attention and proactive measures. By embedding fairness, robust security, and sustainability goals into MLLM development, researchers can ensure these models contribute positively to society while minimizing negative impacts. As this field evolves, interdisciplinary efforts and pioneering approaches to ethical AI deployment will be crucial in navigating the intricate challenges posed by MLLMs.

8.5 Emerging Applications and Use Cases

As the field of Multimodal Large Language Models (MLLMs) progresses, we are witnessing the emergence of novel applications and use cases that harness the distinctive capabilities of these models to address real-world challenges effectively. This section delves into burgeoning areas where MLLMs are making significant impacts, facilitated by their ability to integrate and process diverse modalities—text, visual data, audio, and beyond—to provide transformative solutions across various domains.

One salient application area is the enhancement of AI-driven agents, particularly in robotics and human-computer interaction. Projects like LaMI demonstrate how MLLMs can enable robots to understand and respond to complex instructions in dynamic environments [17]. By processing multimodal inputs, robots can navigate and manipulate environments with greater dexterity, improving their utility in settings such as healthcare, where precision and adaptability are critical. These developments in AI agents not only exemplify advanced robotic autonomy but also highlight the trade-offs between model complexity and real-time processing capabilities [38].

Furthermore, augmented reality (AR) and interactive systems represent another frontier for MLLMs. Innovations such as mixed reality platforms leverage the seamless integration capabilities of MLLMs to create immersive environments where digital and physical worlds converge [5]. These systems facilitate natural user interactions via multimodal signals, which can significantly enhance applications in education and entertainment. As the sophistication of AR technology grows, MLLMs will be crucial in processing the vast amount of data necessary to create seamless and interactive experiences.

In specialized domains, tailored applications of MLLMs are also emerging. In healthcare, MLLMs are showing promise in personalizing patient care by integrating medical records, imaging, and genetic information to provide comprehensive diagnostic and therapeutic insights [38]. The integration of multimodal data allows for a more holistic patient assessment and individualized treatment plans, thus improving outcomes and efficiency in healthcare provision.

Another transformative use case is in personalized recommendation systems. By combining user interactions across multiple platforms—such as text messages, videos, or social media interactions—MLLMs provide highly tailored content recommendations, engaging users in unique and personalized experiences [116]. The richness of multimodal data allows recommendations to consider not only explicit user inputs but also implicit preferences inferred from various forms of media consumption.

The deployment of MLLMs in these emerging applications is, however, not without challenges. A primary concern is the computational and data demands of these models, which can limit scalability and accessibility, particularly in resource-constrained environments [117]. Strategies like activation-aware weight quantization and novel architectural designs aim to address these issues by improving model efficiency without compromising performance, as illustrated by methods such as AWQ and Cobra [60], [66].

Moreover, the ethical and societal implications of MLLM applications warrant critical consideration. Systems that rely on large-scale data aggregation raise privacy concerns, demanding robust frameworks for data management and consent [5]. As MLLMs become more pervasive, ensuring that these innovations do not reinforce biases or exacerbate inequalities is paramount [5].

Looking forward, the future direction for MLLMs in emerging applications will likely involve overcoming these challenges through innovation in model efficiency, ethical data use, and increasing reliance on domain-specific knowledge to craft solutions that are not only technologically advanced but also socially responsible. Technologies such as quantum computing and edge AI may also play a role in ad-

dressing the computational constraints currently associated with MLLMs, paving the way for even more sophisticated and widely-used applications. It is crucial for the research community to continue exploring these avenues to maximize the potential impact of MLLMs while conscientiously managing their societal implications.

9 Conclusion

In synthesizing the comprehensive insights gathered from this extensive survey, the significance and potential of Multimodal Large Language Models (MLLMs) in advancing artificial intelligence become increasingly clear. MLLMs leverage the integration of diverse modalities—such as text, vision, and audio—to enhance model understanding and enable AI systems to perform with a nuanced level of comprehension akin to human intelligence. This subsection focuses on analyzing key findings, evaluating current approaches, identifying emerging challenges, and proposing future directions.

The landscape of MLLMs is defined by architectural innovations that aim to bridge the gap between different modalities. Transformer models have become the standard-bearer in this domain, offering a flexible architecture that allows for effective cross-modal learning and interaction [7]. In particular, models such as the Mixture of Experts (MoE) have optimized the scalability and efficiency of handling various modalities, enabling the dynamic allocation of resources based on input requirements [2]. While these approaches have advanced the field significantly, they also highlight the inherent trade-offs between computational complexity and performance optimization.

Evaluating the strengths and limitations of current MLLM architectures reveals a persistent challenge: the effective fusion of modalities to maintain coherence and context. Methods such as joint embedding spaces and adaptive modality fusion demonstrate progress in achieving coherent representation [69]. Yet, the empirical findings suggest that true synergy across modalities remains elusive, as evidenced by scenarios where high-performing models may inadvertently rely predominantly on unimodal signals despite being equipped for cross-modal interactions [3]. This indicates a misalignment in intended multimodal functionality and the operational reality—a gap that poses opportunities for further algorithmic refinement.

Emerging trends in the field center around enhancing interaction mechanisms and improving robustness against the variability of input data. Recent advancements focus on training paradigms that emphasize cross-modal transfer learning, enabling models to apply learned information from one modality to enhance understanding in another [7]. Techniques like empirical multimodally-additive function projection (EMAP) provide diagnostic tools to isolate unimodal influences, ensuring that cross-modal enhancements are both effective and perceptible [3]. These developments suggest an optimistic trajectory for addressing the limitations of current architectures.

As MLLMs evolve, ethical considerations and societal impacts emerge as critical areas of focus. Bias and fairness issues, privacy preservation, and the societal implications of deploying these models are paramount, particularly as multimodal systems are transitioned from controlled environments to real-world applications [118]. Addressing these concerns requires not only technical solutions but also a collaborative effort among researchers, policymakers, and the broader community to guide the responsible evolution and deployment of MLLMs [5].

Looking ahead, the future of MLLMs lies in developing scalable, efficient, and ethically sound models that can handle complex real-world data. The exploration of parameter-efficient fine-tuning and rapid inference mechanisms represents a promising direction for enhancing the deployment and scalability of these models [14]. Moreover, as applications like healthcare and autonomous systems highlight the transformative capabilities of MLLMs [10], it becomes essential to refine these models further to ensure they meet practical and ethical standards in sensitive domains.

In conclusion, MLLMs exhibit immense potential to advance artificial intelligence through enhanced multimodal integration. However, unlocking their full potential entails addressing fundamental challenges related to cross-modal interactions, ethical considerations, and practical scalability. The synthesis of current findings underscores the importance of continued interdisciplinary research and collaboration to push the boundaries of what is possible with multimodal AI systems. As the field advances, it will be imperative for researchers to maintain a focus on both the technical excellence and responsible innovation of MLLMs to ensure they contribute positively to society. This survey serves as a foundational step toward realizing that vision, providing a comprehensive analysis and stimulating further inquiry into the future of MLLMs in artificial intelligence.

References

[1] J. Huang and J. Zhang, "A survey on evaluation of multimodal large language models," *ArXiv*, vol. abs/2408.15769, 2024. 1, 15, 21

[2] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, pp. 423–443, 2017. 1, 4, 9, 10, 13, 15, 18, 21, 24, 26

[3] J. Hessel and L. Lee, "Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think!" *ArXiv*, vol. abs/2010.06572, 2020. 1, 9, 13, 20, 26

[4] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," in *International Conference on Machine Learning*, 2011, pp. 689–696. 1, 2, 5, 8, 9, 19, 23

[5] P. Liang, A. Zadeh, and L.-P. Morency, "Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions," *ArXiv*, vol. abs/2209.03430, 2022. 1, 2, 4, 5, 6, 8, 10, 12, 13, 14, 17, 20, 21, 23, 25, 26

[6] Q. Ye, H. Xu, J. Ye, M. Yan, A. Hu, H. Liu, Q. Qian, J. Zhang, F. Huang, and J. Zhou, "mplug-owi2: Revolutionizing multi-modal large language model with modality collaboration," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13 040–13 051, 2023. 1

[7] C. Li, Z. Gan, Z. Yang, J. Yang, L. Li, L. Wang, and J. Gao, "Multimodal foundation models: From specialists to general-purpose assistants," *Found. Trends Comput. Graph. Vis.*, vol. 16, pp. 1–214, 2023. 1, 2, 26

[8] J. Xie, Z. Chen, R. Zhang, X. Wan, and G. Li, "Large multimodal agents: A survey," *ArXiv*, vol. abs/2402.15116, 2024. 1, 10

[9] Y. Huang, W. Zhang, L. Feng, X. Wu, and K. C. Tan, "How multimodal integration boost the performance of llm for optimization: Case study on capacitated vehicle routing problems," *ArXiv*, vol. abs/2403.01757, 2024. 1

[10] H. Xiao, F. Zhou, X. Liu, T. Liu, Z. Li, X. Liu, and X. Huang, "A comprehensive survey of large language models and multimodal large language models in medicine," *ArXiv*, vol. abs/2405.08603, 2024. 1, 2, 10, 16, 26

[11] M. Ma, J. Ren, L. Zhao, D. Testuggine, and X. Peng, "Are multimodal transformers robust to missing modality?" *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18 156–18 165, 2022. 1, 4, 10, 14, 25

[12] J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. Raileanu, and R. McHardy, "Challenges and applications of large language models," *ArXiv*, vol. abs/2307.10169, 2023. 1, 2, 9

[13] W. Yu, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, X. Wang, and L. Wang, "Mm-vet: Evaluating large multimodal models for integrated capabilities," *ArXiv*, vol. abs/2308.02490, 2023. 2, 13, 15, 16, 19

[14] M. Xu, W. Yin, D. Cai, R. Yi, D. Xu, Q. Wang, B. Wu, Y. Zhao, C. Yang, S. Wang, Q. Zhang, Z. Lu, L. Zhang, S. Wang, Y. Li, Y. Liu, X. Jin, and X. Liu, "A survey of resource-efficient llm and multimodal foundation models," *ArXiv*, vol. abs/2401.08092, 2024. 2, 17, 22, 26

[15] Z. Jing, Y. Su, Y. Han, B. Yuan, H. Xu, C. Liu, K. Chen, and M. Zhang, "When large language models meet vector databases: A survey," *ArXiv*, vol. abs/2402.01763, 2024. 2

[16] D. Caffagni, F. Cocchi, L. Barsellotti, N. Moratelli, S. Sarto, L. Baraldi, L. Baraldi, M. Cornia, and R. Cucchiara, "The (r)evolution of multimodal large language models: A survey," *ArXiv*, vol. abs/2402.12451, 2024. 2, 9

[17] J. Wu, W. Gan, Z. Chen, S. Wan, and P. S. Yu, "Multimodal large language models: A survey," *2023 IEEE International Conference on Big Data (BigData)*, pp. 2247–2256, 2023. 2, 9, 19, 20, 25

[18] Y. Wang, W. Chen, X. Han, X. Lin, H. Zhao, Y. Liu, B. Zhai, J. Yuan, Q. You, and H. Yang, "Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning," *ArXiv*, vol. abs/2401.06805, 2024. 2, 9, 13, 20

[19] D. Zhang, Y. Yu, C. Li, J. Dong, D. Su, C. Chu, and D. Yu, "Mm-llms: Recent advances in multimodal large language models," in *Annual Meeting of the Association for Computational Linguistics*, 2024, pp. 12 401–12 430. 2, 5

[20] Y.-H. H. Tsai, S. Bai, P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2019, pp. 6558–6569, 2019. 3, 4, 6, 7, 10, 11, 20, 23

[21] Y.-S. Cho, G. V. Steeg, E. Ferrara, and A. Galstyan, "Latent space model for multi-modal social data," *Proceedings of the 25th International Conference on World Wide Web*, 2015. 3

[22] Y. Shi, S. Narayanaswamy, B. Paige, and P. H. S. Torr, "Variational mixture-of-experts autoencoders for multi-modal deep generative models," in *Neural Information Processing Systems*, 2019, pp. 15 692–15 703. 3, 6

[23] N. Du, Y. Huang, A. M. Dai, S. Tong, D. Lepikhin, Y. Xu, M. Krikun, Y. Zhou, A. W. Yu, O. Firat, B. Zoph, L. Fedus, M. Bosma, Z. Zhou, T. Wang, Y. E. Wang, K. Webster, M. Pellat, K. Robinson, K. Meier-Hellstern, T. Duke, L. Dixon, K. Zhang, Q. V. Le, Y. Wu, Z. Chen, and C. Cui, "Glam: Efficient scaling of language models with mixture-of-experts," in *International Conference on Machine Learning*, 2021, pp. 5547–5569. 3

[24] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," *ArXiv*, vol. abs/2306.13549, 2023. 3, 4, 10, 13, 22

[25] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal intelligence: Representation learning, information fusion, and applications," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, pp. 478–493, 2019. 3, 6, 8, 9, 14, 17, 19, 20

[26] R. Kiros, R. Salakhutdinov, and R. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *ArXiv*, vol. abs/1411.2539, 2014. 3

[27] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. Hershey, T. K. Marks, and K. Sumi, "Attention-based multimodal fusion for video description," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 4203–4212, 2017. 3

[28] X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu, "Balanced multimodal learning via on-the-fly gradient modulation," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8237, 2022. 4, 11, 18

[29] P. Xu, X. Zhu, and D. Clifton, "Multimodal learning with transformers: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 12 113–12 132, 2022. 4, 14, 18

[30] F. B. Baldassini, M. Shukor, M. Cord, L. Soulier, and B. Piwowarski, "What makes multimodal in-context learning work?" *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1539–1550, 2024. 4

[31] Y. Jin, J. Li, Y. Liu, T. Gu, K. Wu, Z. Jiang, M. He, B. Zhao, X. Tan, Z. Gan, Y. Wang, C. Wang, and L. Ma, "Efficient multimodal large language models: A survey," *ArXiv*, vol. abs/2405.10739, 2024. 4, 15, 22, 25

[32] F. Bordes, R. Y. Pang, A. Ajay, A. C. Li, A. Bardes, S. Petryk, O. Mañas, Z. Lin, A. Mahmoud, B. Jayaraman, M. Ibrahim, M. Hall, Y. Xiong, J. Lebensold, C. Ross, S. Jayakumar, C. Guo, D. Bouchacourt, H. Al-Tahan, K. Padthe, V. Sharma, H. Xu, X. E. Tan, M. Richards, S. Lavoie, P. Astolfi, R. A. Hemmat, J. Chen, K. Tirumala, R. Assouel, M. Moayeri, A. Talattof, K. Chaudhuri, Z. Liu, X. Chen, Q. Garrido, K. Ullrich, A. Agrawal, K. Saenko, A. Celikyilmaz, and V. Chandra, "An introduction to vision-language modeling," *ArXiv*, vol. abs/2405.17247, 2024. 4

[33] M. Shi, F. Liu, S. Wang, S. Liao, S. Radhakrishnan, D.-A. Huang, H. Yin, K. Sapra, Y. Yacoob, H. Shi, B. Catanzaro, A. Tao, J. Kautz, Z. Yu, and G. Liu, "Eagle: Exploring the design space for multimodal llms with mixture of encoders," *ArXiv*, vol. abs/2408.15998, 2024. 4

[34] L. Wei, Z. Jiang, W. Huang, and L. Sun, "Instructiongpt-4: A 200-instruction paradigm for fine-tuning minigpt-4," *ArXiv*, vol. abs/2308.12067, 2023. 5

[35] Z. Tang, Z. Yang, C. Zhu, M. Zeng, and M. Bansal, "Any-to-any generation via composable diffusion," *ArXiv*, vol. abs/2305.11846, 2023. 5

[36] S. Shen, Z. Yao, C. Li, T. Darrell, K. Keutzer, and Y. He, "Scaling vision-language models with sparse mixture of experts," *ArXiv*, vol. abs/2303.07226, 2023. 5, 12, 19

[37] J. Li, X. Wang, S. Zhu, C.-W. Kuo, L. Xu, F. Chen, J. Jain, H. Shi, and L. Wen, "Cumo: Scaling multimodal llm with co-upcycled mixture-of-experts," *ArXiv*, vol. abs/2405.05949, 2024. 5, 8, 12

[38] J. M. Z. Chaves, S.-C. Huang, Y. Xu, H. Xu, N. Usuyama, S. Zhang, F. Wang, Y. Xie, M. Khademi, Z. Yang, H. Awadalla, J. Gong, H. Hu, J. Yang, C. yue Li, J. Gao, Y. Gu, C. Wong, M.-H. Wei, T. Naumann, M. Chen, M. Lungren, S. Yeung-Levy, C. P. Langlotz, S. Wang, and H. Poon, "Training small multimodal models to bridge biomedical competency gap: A case study in radiology imaging," *ArXiv*, vol. abs/2403.08002, 2024. 5, 25

[39] Z. Shao, Z. Yu, J. Yu, X. Ouyang, L. Zheng, Z. Gai, M. Wang, and J. Ding, "Imp: Highly capable large multimodal models for mobile devices," *ArXiv*, vol. abs/2405.12107, 2024. 5

[40] S. Xu, C. K. Thomas, O. Hashash, N. Muralidhar, W. Saad, and N. Ramakrishnan, "Large multi-modal models (lmms) as universal foundation models for ai-native wireless systems," *IEEE Network*, vol. 38, pp. 10–20, 2024. 5

[41] T. Wang, W. Zhou, Y. Zeng, and X. Zhang, "Efficientvlm: Fast and accurate vision-language models via knowledge distillation and modal-adaptive pruning," *ArXiv*, vol. abs/2210.07795, 2022. 5

[42] L. Zhang, L. Zhang, S. Shi, X. Chu, and B. Li, "Lora-fa: Memory-efficient low-rank adaptation for large language models fine-tuning," *ArXiv*, vol. abs/2308.03303, 2023. 5, 8

[43] I. Han, R. Jayaram, A. Karbasi, V. Mirrokni, D. P. Woodruff, and A. Zandieh, "Hyperattention: Long-context attention in near-linear time," *ArXiv*, vol. abs/2310.05869, 2023. 5, 12

[44] P. P. Liang, Y. Cheng, X. Fan, C. K. Ling, S. Nie, R. Chen, Z. Deng, N. Allen, R. Auerbach, F. Mahmood *et al.*, "Quantifying & modeling multimodal interactions: An information decomposition framework," *Advances in Neural Information Processing Systems*, vol. 36, 2024. 5, 11, 13, 18, 21

[45] Y. Fan, W. Xu, H. Wang, J. Wang, and S. Guo, "Pmr: Prototypical modal rebalance for multimodal learning," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20 029–20 038, 2022. 6

[46] B. Mustafa, C. Riquelme, J. Puigcerver, R. Jenatton, and N. Houlsby, "Multimodal contrastive learning with limoe: the language-image mixture of experts," *ArXiv*, vol. abs/2206.02770, 2022. 6, 10, 23

[47] S. Chen, Z. Jie, and L. Ma, "Llava-mole: Sparse mixture of lora experts for mitigating data conflicts in instruction finetuning mllms," *ArXiv*, vol. abs/2401.16160, 2024. 6, 17

[48] F. Xue, Z. Zheng, Y. Fu, J. Ni, Z. Zheng, W. Zhou, and Y. You, "Openmoe: An early effort on open mixture-of-experts language models," *ArXiv*, vol. abs/2402.01739, 2024. 6

[49] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang, "Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," in *International Conference on Machine Learning*, 2022, pp. 23 318–23 340. 6

[50] Y. Qiao, Z. Yu, L. Guo, S. Chen, Z. Zhao, M. Sun, Q. Wu, and J. Liu, "Vl-mamba: Exploring state space models for multimodal learning," *ArXiv*, vol. abs/2403.13600, 2024. 6, 21

[51] M. Suzuki and Y. Matsuo, "A survey of multimodal deep generative models," *Advanced Robotics*, vol. 36, pp. 261 – 278, 2022. 7

[52] S. Mai, H. Hu, and S. Xing, "Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion," *ArXiv*, vol. abs/1911.07848, 2019. 7

[53] H. Pham, P. Liang, T. Manzini, L.-P. Morency, and B. Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *AAAI Conference on Artificial Intelligence*, 2018, pp. 6892–6899. 7, 11

[54] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind one embedding space to bind them all," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15 180–15 190, 2023. 7, 18

[55] Y.-H. H. Tsai, P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," *ArXiv*, vol. abs/1806.06176, 2018. 7, 18, 21

[56] R. Jiang, L. Liu, and C. Chen, "Mope: Parameter-efficient and scalable multimodal fusion via mixture of prompt experts," *ArXiv*, vol. abs/2403.10568, 2024. 7, 18, 24

[57] F. Zhao, T. Pang, C. Li, Z. Wu, J. Guo, S. Xing, and X. Dai, "AlignGPT: Multi-modal large language models with adaptive alignment capability," *ArXiv*, vol. abs/2405.14129, 2024. 7

[58] T. Vallaeys, M. Shukor, M. Cord, and J. Verbeek, "Improved baselines for data-efficient perceptual augmentation of llms," *ArXiv*, vol. abs/2403.13499, 2024. 7

[59] W. Li, Y. Yuan, J. Liu, D. Tang, S. Wang, J. Zhu, and L. Zhang, "Tokenpacker: Efficient visual projector for multimodal llm," *ArXiv*, vol. abs/2407.02392, 2024. 7

[60] H. Zhao, M. Zhang, W. Zhao, P. Ding, S. Huang, and D. Wang, "Cobra: Extending mamba to multi-modal large language model for efficient inference," *ArXiv*, vol. abs/2403.14520, 2024. 7, 18, 25

[61] C. Zhou, L. Yu, A. Babu, K. Tirumala, M. Yasunaga, L. Shamis, J. Kahn, X. Ma, L. Zettlemoyer, and O. Levy, "Transfusion: Predict the next token and diffuse images with one multi-modal model," *ArXiv*, vol. abs/2408.11039, 2024. 7

[62] Y. Shang, M. Cai, B. Xu, Y. J. Lee, and Y. Yan, "Llava-prumerge: Adaptive token reduction for efficient large multimodal models," *ArXiv*, vol. abs/2403.15388, 2024. 7

[63] B. Li, Y. Zhang, L. Chen, J. Wang, F. Pu, J. Yang, C. Li, and Z. Liu, "Mimic-it: Multi-modal in-context instruction tuning," *ArXiv*, vol. abs/2306.05425, 2023. 7, 12

[64] Y. Zhou, C. Guo, X. Wang, Y. Chang, and Y. Wu, "A survey on data augmentation in large model era," *ArXiv*, vol. abs/2401.15422, 2024. 8, 20

[65] M. He, Y. Liu, B. Wu, J. Yuan, Y. Wang, T. Huang, and B. Zhao, "Efficient multimodal learning from data-centric perspective," *ArXiv*, vol. abs/2402.11530, 2024. 8, 12

[66] J. Lin, J. Tang, H. Tang, S. Yang, X. Dang, and S. Han, "Awq: Activation-aware weight quantization for llm compression and acceleration," *ArXiv*, vol. abs/2306.00978, 2023. 8, 25

[67] A. Ansell, I. Vuli'c, H. Sterz, A. Korhonen, and E. Ponti, "Scaling sparse fine-tuning to large language models," *ArXiv*, vol. abs/2401.16405, 2024. 8

[68] A. Rahate, R. Walambe, S. Ramanna, and K. Kotecha, "Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions," *Inf. Fusion*, vol. 81, pp. 203–239, 2021. 8, 19

[69] S. Uppal, S. Bhagat, D. Hazarika, N. Majumder, S. Poria, R. Zimmermann, and A. Zadeh, "Multimodal research in vision and language: A review of current and emerging trends," *Inf. Fusion*, vol. 77, pp. 149–171, 2022. 9, 15, 20, 26

[70] W. Ye, G. Zheng, Y. Ma, X. Cao, B. Lai, J. Rehg, and A. Zhang, "Mm-spubench: Towards better understanding of spurious biases in multimodal llms," *ArXiv*, vol. abs/2406.17126, 2024. 9

[71] M. Artetxe, S. Bhosale, N. Goyal, T. Mihaylov, M. Ott, S. Shleifer, X. V. Lin, J. Du, S. Iyer, R. Pasunuru, G. Anantharaman, X. Li, S. Chen, H. Akın, M. Baines, L. Martin, X. Zhou, P. S. Koura, B. O'Horo, J. Wang, L. Zettlemoyer, M. T. Diab, Z. Kozareva, and V. Stoyanov, "Efficient large scale language modeling with mixtures of experts," in *Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 11 699–11 732. 10, 21

[72] X. Yue, Y. Ni, K. Zhang, T. Zheng, R. Liu, G. Zhang, S. Stevens, D. Jiang, W. Ren, Y. Sun, C. Wei, B. Yu, R. Yuan, R. Sun, M. Yin, B. Zheng, Z. Yang, Y. Liu, W. Huang, H. Sun, Y. Su, and W. Chen, "Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, 2023. 10, 14

[73] A. S. Sundar and L. Heck, "Multimodal conversational ai: A survey of datasets and approaches," *ArXiv*, vol. abs/2205.06907, 2022. 11, 14

[74] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and -specific representations for multimodal sentiment analysis," *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. 11

[75] Z. Wang, Z. Wan, and X. Wan, "Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis," *Proceedings of The Web Conference 2020*, 2020. 11

[76] Y. Yin, F. Meng, J. Su, C. Zhou, Z. Yang, J. Zhou, and J. Luo, "A novel graph-based multi-modal fusion encoder for neural machine translation," in *Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3025–3035. 11, 18

[77] A. Jaegle, F. Gimeno, A. Brock, A. Zisserman, O. Vinyals, and J. Carreira, "Perceiver: General perception with iterative attention," *ArXiv*, vol. abs/2103.03206, 2021. 11, 17, 18, 23

[78] F. Raue, T. Breuel, A. Dengel, and M. Liwicki, "Symbol grounding association in multimodal sequences with missing elements," *J. Artif. Intell. Res.*, vol. 61, pp. 787–806, 2015. 11

[79] P. Liang, L. Ziyin, A. Zadeh, and L.-P. Morency, "Multimodal language analysis with recurrent multistage fusion," in *Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 150–161. 12, 14, 15

[80] H. Pham, T. Manzini, P. Liang, and B. Póczos, "Seq2seq2sentiment: Multimodal sequence to sequence models for sentiment analysis," *ArXiv*, vol. abs/1807.03915, 2018. 12

[81] G. Pandey and A. Dukkipati, "Variational methods for conditional multimodal deep learning," *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 308–315, 2016. 12, 19

[82] B. Zhou, Y. Hu, X. Weng, J. Jia, J. Luo, X. Liu, J. Wu, and L. Huang, "Tinyllava: A framework of small-scale large multimodal models," *ArXiv*, vol. abs/2402.14289, 2024. 12

[83] J. Wang, Y. Ming, Z. Shi, V. Vineet, X. Wang, and N. Joshi, "Is a picture worth a thousand words? delving into spatial reasoning for vision language models," *ArXiv*, vol. abs/2406.14852, 2024. 15

[84] L. Chen, Y. Zhang, S. Ren, H. Zhao, Z. Cai, Y. Wang, P. Wang, X. Meng, T. Liu, and B. Chang, "Pca-bench: Evaluating multimodal large language models in perception-cognition-action chain," *ArXiv*, vol. abs/2402.15527, 2024. 15

[85] T. Bai, H. Liang, B. Wan, L. Yang, B. Li, Y. Wang, B. Cui, C. He, B. Yuan, and W. Zhang, "A survey of multimodal large language model from a data-centric perspective," *ArXiv*, vol. abs/2405.16640, 2024. 15, 25

[86] K. Ying, F. Meng, J. Wang, Z. Li, H. Lin, Y. Yang, H. Zhang, W. Zhang, Y. Lin, S. Liu, J. Lei, Q. Lu, R. Chen, P. Xu, R. Zhang, H. Zhang, P. Gao, Y. Wang, Y. Qiao, P. Luo, K. Zhang, and W. Shao, "Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi," *ArXiv*, vol. abs/2404.16006, 2024. 16

[87] K. Zhang, B. Li, P. Zhang, F. Pu, J. A. Cahyono, K. Hu, S. Liu, Y. Zhang, J. Yang, C. Li, and Z. Liu, "Lmms-eval: Reality check on the evaluation of large multimodal models," *ArXiv*, vol. abs/2407.12772, 2024. 16

[88] H. Duan, J. Yang, Y. Qiao, X. Fang, L. Chen, Y. Liu, X. wen Dong, Y. Zang, P. Zhang, J. Wang, D. Lin, and K. Chen, "Vlmevalkit: An open-source toolkit for evaluating large multi-modality models," *ArXiv*, vol. abs/2407.11691, 2024. 16

[89] Z. Liu, Y. Li, P. Shu, A. Zhong, L. Yang, C. Ju, Z. Wu, C.-Y. Ma, J. Luo, C. Chen, S. Kim, J. Hu, H. Dai, L. Zhao, D. Zhu, J. Liu, W. Liu, D. Shen, T. Liu, Q. Li, and X. Li, "Radiology-llama2:

Best-in-class large language model for radiology," *ArXiv*, vol. abs/2309.06419, 2023. 16
[90] Y. Zheng, W. Gan, Z. Chen, Z. Qi, Q. Liang, and P. S. Yu, "Large language models for medicine: A survey," *ArXiv*, vol. abs/2405.13055, 2024. 16
[91] N. Pfleger, "Context based multimodal fusion," in *International Conference on Multimodal Interaction*, 2004, pp. 265–272. 17, 18
[92] C. Cui, Y. Ma, X. Cao, W. Ye, Y. Zhou, K. Liang, J. Chen, J. Lu, Z. Yang, K.-D. Liao, T. Gao, E. Li, K. Tang, Z. Cao, T. Zhou, A. Liu, X. Yan, S. Mei, J. Cao, Z. Wang, and C. Zheng, "A survey on multimodal large language models for autonomous driving," *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pp. 958–979, 2023. 17
[93] J. Liu, M. Liu, Z. Wang, L. Lee, K. Zhou, P. An, S. Yang, R. Zhang, Y. Guo, and S. Zhang, "Robomamba: Multimodal state space model for efficient robot reasoning and manipulation," *ArXiv*, vol. abs/2406.04339, 2024. 17
[94] T. Dao, D. Y. Fu, K. K. Saab, A. Thomas, A. Rudra, and C. Ré, "Hungry hungry hippos: Towards language modeling with state space models," *ArXiv*, vol. abs/2212.14052, 2022. 17
[95] B.-K. Lee, B. Park, C. W. Kim, and Y. Ro, "Moai: Mixture of all intelligence for large language and vision models," *ArXiv*, vol. abs/2403.07508, 2024. 17
[96] J. Lu, C. Clark, S. Lee, Z. Zhang, S. Khosla, R. Marten, D. Hoiem, and A. Kembhavi, "Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26 429–26 445, 2023. 18, 25
[97] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, and P. R. Florence, "Palm-e: An embodied multimodal language model," in *International Conference on Machine Learning*, 2023, pp. 8469–8488. 18
[98] W. Wang, Z. Chen, X. Chen, J. Wu, X. Zhu, G. Zeng, P. Luo, T. Lu, J. Zhou, Y. Qiao, and J. Dai, "Visionllm: Large language model is also an open-ended decoder for vision-centric tasks," *ArXiv*, vol. abs/2305.11175, 2023. 18, 22
[99] Z. Yang, L. Li, J. Wang, K. Lin, E. Azarnasab, F. Ahmed, Z. Liu, C. Liu, M. Zeng, and L. Wang, "Mm-react: Prompting chatgpt for multimodal reasoning and action," *ArXiv*, vol. abs/2303.11381, 2023. 19
[100] L. Parcalabescu, M. Cafagna, L. Muradjan, A. Frank, I. Calixto, and A. Gatt, "Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena," *ArXiv*, vol. abs/2112.07566, 2021. 19
[101] F. Bao, S. Nie, K. Xue, C. Li, S. Pu, Y. Wang, G. Yue, Y. Cao, H. Su, and J. Zhu, "One transformer fits all distributions in multi-modal diffusion at scale," in *International Conference on Machine Learning*, 2023, pp. 1692–1717. 19
[102] S. Woo, S. Lee, Y. Park, M. A. Nugroho, and C. Kim, "Towards good practices for missing modality robust action recognition," in *AAAI Conference on Artificial Intelligence*, 2022, pp. 2776–2784. 20
[103] Z. Sun, S. Shen, S. Cao, H. Liu, C. Li, Y. Shen, C. Gan, L. Gui, Y.-X. Wang, Y. Yang, K. Keutzer, and T. Darrell, "Aligning large multimodal models with factually augmented rlhf," *ArXiv*, vol. abs/2309.14525, 2023. 21
[104] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, pp. 5625–5644, 2023. 22
[105] M. Beck, K. Poppel, M. Spanring, A. Auer, O. Prudnikova, M. K. Kopp, G. Klambauer, J. Brandstetter, and S. Hochreiter, "xlstm: Extended long short-term memory," *ArXiv*, vol. abs/2405.04517, 2024. 22
[106] J. Li and W. Lu, "A survey on benchmarks of multimodal large language models," *ArXiv*, vol. abs/2408.08632, 2024. 22, 25
[107] J. Z. Pan, S. Razniewski, J.-C. Kalo, S. Singhania, J. Chen, S. Dietze, H. Jabeen, J. Omeliyanenko, W. Zhang, M. Lissandrini, R. Biswas, G. de Melo, A. Bonifati, E. Vakaj, M. Dragoni, and D. Graux, "Large language models and knowledge graphs: Opportunities and challenges," *ArXiv*, vol. abs/2308.06374, 2023. 23
[108] Z. Qin, D. Chen, W. Zhang, L. Yao, Y. Huang, B. Ding, Y. Li, and S. Deng, "The synergy between data and multi-modal large language models: A survey from co-development perspective," *ArXiv*, vol. abs/2407.08583, 2024. 23
[109] Z. Zong, B. Ma, D. Shen, G. Song, H. Shao, D. Jiang, H. Li, and Y. Liu, "Mova: Adapting mixture of vision experts to multimodal context," *ArXiv*, vol. abs/2404.13046, 2024. 23
[110] X. Han, H. Nguyen, C. Harris, N. Ho, and S. Saria, "Fusemoe: Mixture-of-experts transformers for fleximodal fusion," *ArXiv*, vol. abs/2402.03226, 2024. 23
[111] Z. fei Yin, J. Wang, J. Cao, Z. Shi, D. Liu, M. Li, L. Sheng, L. Bai, X. Huang, Z. Wang, W. Ouyang, and J. Shao, "Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark," *ArXiv*, vol. abs/2306.06687, 2023. 23
[112] S. Moon, A. Madotto, Z. Lin, T. Nagarajan, M. Smith, S. Jain, C.-F. Yeh, P. Murugesan, P. Heidari, Y. Liu, K. Srinet, B. Damavandi, and A. Kumar, "Anymal: An efficient and scalable any-modality augmented language model," *ArXiv*, vol. abs/2309.16058, 2023. 24
[113] Y. Li, R. Quan, L. Zhu, and Y. Yang, "Efficient multimodal fusion via interactive prompting," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2604–2613, 2023. 24
[114] H. Ye, D.-A. Huang, Y. Lu, Z. Yu, W. Ping, A. Tao, J. Kautz, S. Han, D. Xu, P. Molchanov, and H. Yin, "X-vila: Cross-modality alignment for large language model," *ArXiv*, vol. abs/2405.19335, 2024. 24
[115] M. Chen, Y. Cao, Y. Zhang, and C. Lu, "Quantifying and mitigating unimodal biases in multimodal large language models: A causal perspective," *ArXiv*, vol. abs/2403.18346, 2024. 24
[116] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, and X. Peng, "Smil: Multimodal learning with severely missing modality," *ArXiv*, vol. abs/2103.05677, 2021. 25
[117] Z. Lin, G. Qu, Q. Chen, X. Chen, Z. Chen, and K. Huang, "Pushing large language models to the 6g edge: Vision, challenges, and opportunities," *ArXiv*, vol. abs/2309.16739, 2023. 25
[118] A. Tamkin, M. Brundage, J. Clark, and D. Ganguli, "Understanding the capabilities, limitations, and societal impact of large language models," *ArXiv*, vol. abs/2102.02503, 2021. 26