Bridging Theory and Practice in Multimodal Deep Learning: A Comprehensive Review in the Large Language Model Era

Anonymous ACL submission

Abstract

In the past few years, the realm of deep learning has captivated widespread interest, with multimodal deep learning (MMDL) rising as an exceptionally promising area. MMDL specializes in processing and amalgamating data from varied communication channels, including text, speech, vision, and spatial indicators. This article delivers an exhaustive exploration of MMDL methodologies and their expansive applications. Furthermore, we delve into a detailed examination of diverse MMDL techniques, encapsulating the progression of model architectures, advancements in data augmentation, refresh methods, and optimization tactics. The main goal of this review is to tackle the pressing challenges and delineate the trajectory for future research in the dynamic field of deep learning, especially focusing on the era of Large Language Models (LLMs). We believe that this comprehensive review will greatly enhance the comprehension of MMDL and act as a crucial tool for researchers aiming to delve into new and promising research paths.

1 Introduction

Multimodal deep learning is a promising technique that utilizes information from multiple modalities, including vision, text, audio, and others, to enhance learning outcomes (Summaira et al., 2022). The exponential growth of knowledge in our world has created a need for more efficient and effective learning approaches. Humans have the ability to leverage cross-modal information to efficiently learn new concepts, which has inspired the development of multimodality (Vasco et al., 2022; Lin et al., 2023).

Multimodality is fundamental to many areas of our society, such as scientific research (Nancekivell et al., 2021; Yan et al., 2022), education (Magnusson and Godhe, 2019), medical diagnosis (Sveric et al., 2022; Chen et al., 2023c) and many more. However, multimodal deep learning is a critical yet underexplored problem in some areas, and its exploration is essential for the development of intelligent agents (Abdulrahman and Richards, 2022; Brinkschulte et al., 2022). The human ability to leverage cross-modal information is essential for effective learning and recognition of visual objects, even with limited examples (Lin et al., 2023). In this regard, verbal language has been shown to facilitate the recognition of visual objects, and the neuroscience literature provides ample evidence that cognitive representations are inherently multimodal (Jackendoff, 1987; Smith and Gasser, 2005). For example, different types of stimuli, such as visual images, textual strings, and audio clips, can evoke the same neurons, indicating the existence of cross-modal or inter-modal representations (Quiroga et al., 2005; Nanay, 2018). These representations are fundamental to the human perceptual-cognitive system and play a crucial role in the acquisition of new concepts and knowledge (Gibson, 1969; Cohn, 2016). Thus, multimodal deep learning techniques hold the potential to significantly enhance learning outcomes in various applications, including speech recognition (Kumar et al., 2022; Kshirsagar et al., 2023), multimedia indexing (Snoek et al., 2006), human behavior analysis (Pantic and Rothkrantz, 2003), video captioning (Song et al., 2018), visual question answering (Antol et al., 2015), among others. The combination of multiple modalities enables the deep learning models to have a more comprehensive understanding of the environment since certain cues are only present in specific modalities. For example, the task of emotion recognition (Koolagudi and Rao, 2012) is not only reliant on facial expressions which are captured through the visual modality, but also on tone and pitch of the voice, captured through the audio modality. The inclusion of both modalities can encode a vast amount of information about the emotional state.

In recent years, the field of multimodal learning

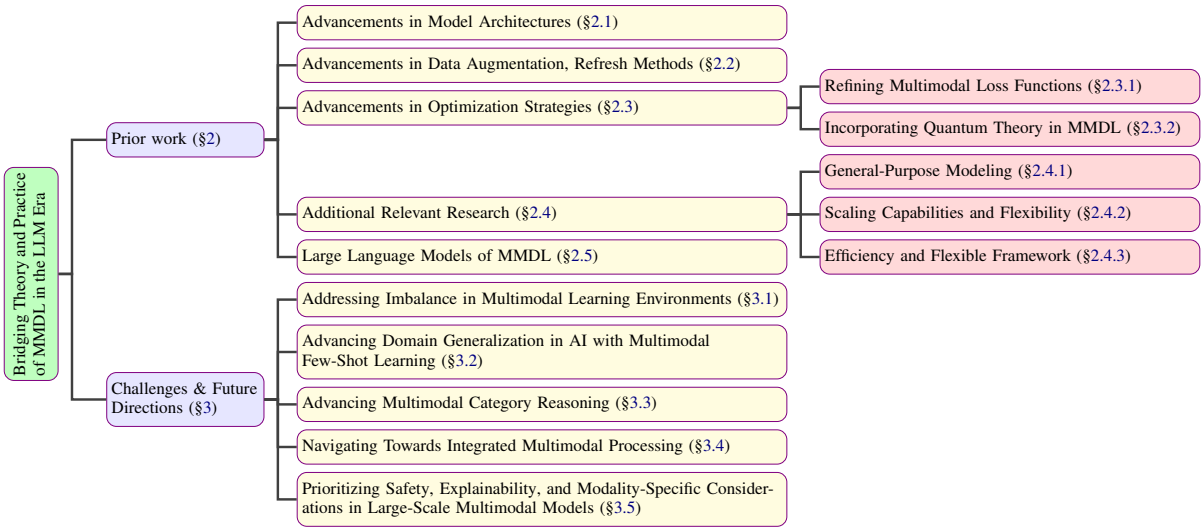

Figure 1: The overall structure of this paper.

has experienced a surge in growth and development, with numerous studies exploring various aspects of this field. However, due to the diverse nature of multimodal data (Lahat et al., 2015) and the interdisciplinary nature of the field (Magnusson and Godhe, 2019; Yan et al., 2022; Chen et al., 2023c), research in this area tends to be fragmented and isolated within different domains. This lack of an integrated overview poses a challenge to researchers seeking a comprehensive understanding of the latest developments in the rapidly evolving field of multimodal learning, especially in the context of the explosive growth of large language models.

To this end, our paper endeavors to offer a comprehensive examination of the existing literature on multimodal learning, as shown in Figure 1. And our contribution can be summarized as following: (1) We present an in-depth overview of various MMDL methodologies, including model architectural evolution, data augmentation and refresh, and optimization strategies. (2) We summarise the application of current large-scale models in MMDL in four ways and the research pain points of multimodal learning in the LLM era, which opens up new avenues for investigating multimodality. (3) We provide a comparative analysis of current large scale models in the field of multimodal learning on various benchmark and evaluation metrics. Comparative analysis helps researchers to get directions in future research. (4) We emphasize the five primary challenges and possible future research areas of MMDL. By focusing on the challenges and opportunities inherent in multimodal learning, we aim to bridge the gap between theoretical understanding and practical implementation.

2 Prior work

2.1 Advancements in Multimodal Learning Architectures

In recent times, the domain of multimodal representation learning, particularly in the context of vision-text tasks, has seen a significant surge in interest. This area of study has become a focal point of attention within the academic and scientific circles. Therefore, in our exposition, we primarily focus on delineating the model architecture, with specific emphasis on vision-text based models.

ViLT is a type of visual language model that was proposed by (Kim et al., 2021). This model does not require convolution or region supervision, and it embeds text into Vision Transformer (Yuan et al., 2021) with minimal design for visual and language pre-training. Specifically, the embedding layers of raw pixels are shallow and computationally light, similar to text tokens, with most of the computation concentrated on modeling modality interactions, as shown in Figure 2 (b). While external visual information typically provides a richer representation of modalities compared to text information, which aligns with human perception, ViLT's suboptimal performance can be attributed to the shallow encoding of the visual modality, despite its fast computational speed. ALBEF (Li et al., 2021a) addresses this performance issue by employing a deeper visual encoder and aligning image and text representations prior to fusing them through cross-modal attention, as depicted in Figure 2(c). Another ap-

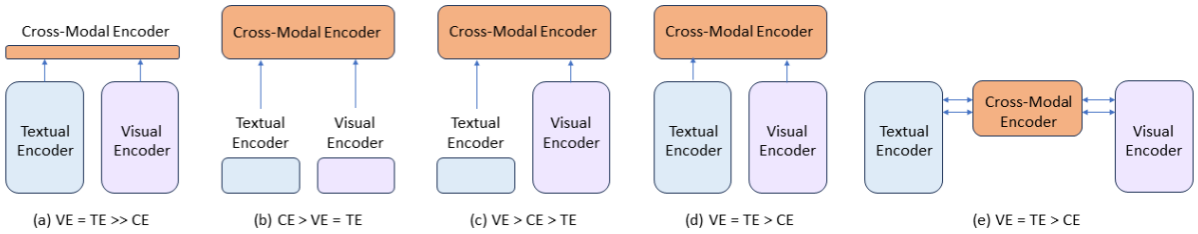

Figure 2: The current vision-language models architecture can be classified into five categories, as represented by (a) to (e). where VE, TE, and CE denote the Visual Encoder, Textual Encoder, and Cross-modal Encoder, respectively. The height of each rectangle in the illustration corresponds to its relative computational cost, and VE = TE denotes that the visual encoder and the textual encoder have comparable parameters or computational costs.

proach, CLIP (Radford et al., 2021), illustrated in Figure 2(a), is renowned for its efficient computation and high-quality feature extraction. For tasks like Visual Question Answering, CLIP relies on a dot product to determine similarity. However, CLIP's encoder suffers from the limitation of equal sizing for visual and textual inputs, resulting in suboptimal performance in various model fusion tasks due to shallow interaction. To better adapt to various multimodal tasks, VLMo (Bao et al., 2022) presents a unified approach to multimodal pre-training and fine-tuning by utilizing modality experts and freezing the parameters of the shared multimodal layer. This approach strikes a balance between computational efficiency and performance. VLMo trains different experts for different tasks, resulting in enhanced multimodal representation learning while reducing computational costs. Table 1 shows the performance of some of the models discussed above.

The above models use the encoder side of the Transformer structure for multimodal tasks, and there are also some recent works studying how to use the decoder side of the Transformer for generation tasks, including BLIP (Li et al., 2022a), CoCa (Yu et al., 2022) and the BEiT series (Bao et al., 2021; Peng et al., 2022; Wang et al., 2022).

BLIP is a recent model that utilizes two deep text-visual encoders for effective information extraction. It combines the advantages of both ALBEF and VLMO, and unifies vision-language understanding and generation in a single framework. The model employs parameter sharing to reduce computational complexity while achieving significant progress. On the other hand, CoCa also adopts a similar model architecture to ALBEF as depicted in Figure 2(c), but uses a text decoder for encoding on the textual side. It demonstrates excellent performance in multimodal generation tasks. BEiT is a type of multimodal model that uses self-supervised learning to pre-training vision Transformers using a masked image modeling task. BEiT-v2 and BEiT-v3 further enhance this approach by employing new pre-training tasks and architecture. BEiT-v3 also draws inspiration from VLMo, utilizing multiple experts and treating image information as a foreign language to unify vision-language tasks, resulting in exceptional performance. Table 2 includes a performance comparison between the different models.

Model	# Pretrain Images	VQA		NLVR2	
		test-dev	test-std	dev	test-P
ViLT (Kim et al., 2021)	4M	71.26	-	75.70	76.13
ALBEF (Li et al., 2021a)	4M	74.54	74.70	80.24	80.50
VLMo (Bao et al., 2022)	4M	76.64	76.89	82.77	83.34

Table 1: Model performance on VQA and NLVR2. We report vqa-score on VQA test-dev and test-standard split, and report accuracy for NLVR2 development and public test set (test-P). The reported results are from published literature (Bao et al., 2022).

However, previous works in multimodal learning have mainly focused on fusing input modalities after significant independent processing, which can be time-consuming and computationally expensive. In contrast, the human brain performs multimodal processing almost immediately (Angelaki and Cullen, 2008). Therefore, one crucial design decision in multimodal learning is how to best combine, or fuse, the different input modalities. To address this issue, recent research (Xu et al., 2022) proposes a cross-model encoder that simultaneously uses multimodal information fusion during model encoding, as shown in Figure 2(e), achieving excellent interaction under two-modal fusion. This approach aims to better simulate the way the human brain processes multiple modali-

ties, which may lead to improved performance and efficiency in multimodal learning tasks.

Model	VQAv2		NLVR2	
	test-dev	test-std	dev	test-P
ALBEF (Li et al., 2021a)	75.84	76.04	82.55	83.14
BLIP (Li et al., 2022a)	78.25	78.32	82.15	82.24
CoCa (Yu et al., 2022)	82.30	82.30	86.10	87.00
BEiT-3 (Wang et al., 2022)	84.19	84.03	91.51	92.58

Table 2: Model performance on visual question answering and visual reasoning. We report vqa-score on VQAv2 test-dev and test-standard splits, accuracy for NLVR2 development set and public test set (test-P). The reported results are from published literature (Wang et al., 2022).

2.2 Enhancing Multimodal Data: Augmentation and Refresh Strategies

In multimodal learning, training models often rely on web-sourced data, which is prone to containing noise that may degrade model performance. To mitigate this, researchers have devised various methods for augmenting and refining data.

A notable technique is DataMix (Liu et al., 2020), which implements a blending-based strategy to create new image-text pairs. This method involves altering existing pairs through random weighted averaging, thus generating unique data instances. Conversely, DataEcho (Cioffi and Bingham, 1994) applies an echo-based technique to modify image-text pairs, producing fresh data pairs and thereby contributing to the data augmentation process.

Differently, DataReMix (Mao et al., 2021) employs a strategy of pair replacement or swapping, aiming to diversify the dataset and enhance model resilience. Beyond simple augmentation and refinement, the advent of multimodal self-supervised learning marks a significant stride forward. This method uses one modality as a supervisory signal for another, such as image-to-speech or image-to-text. For instance, MixGen (Hao et al., 2023) innovates by generating new training samples through a process of image interpolation and text sequence concatenation from existing pairs, enriching the diversity and quality of multimodal data.

Additionally, the data filter and caption technique (Li et al., 2022a) involves fine-tuning models with high-quality, manually labeled data and subsequently filtering and enhancing a vast amount of web data. This enhances the correlation between images and text.

These advancements are proving to be invaluable in augmenting the quality and volume of multimodal data, thereby elevating model performance. They are especially critical in contexts where access to high-quality training data is scarce.

2.3 Optimization Strategies in Multimodal Learning

Optimization strategies play a pivotal role in augmenting the performance and broadening the generalization capabilities of multimodal models, which process diverse data types such as text, images, and audio. These models typically necessitate either alignment or fusion of modalities at feature or decision levels for optimal functioning.

2.3.1 Refining Multimodal Loss Functions

The loss function in a multimodal model quantifies the deviation between the model's output and the actual label, guiding the optimization process. Recent advancements have introduced innovative methods to enhance the alignment or fusion in multimodal models. For example, Xu et al. (2023b) developed a balanced multimodal learning approach using the multimodal cosine loss function. This method adapts feature and weight normalization to multimodal contexts, thus refining the model's discriminative capabilities.

In a similar vein, Yang et al. (2021) introduced TACo, a method for multimodal alignment using three distinct loss functions. This approach leverages unimodal self-supervised information, cross-modal comparison data, and cross-task shared insights to construct these loss functions. They work by improving the representation within each modality, enhancing similarity across different modalities, and utilizing correlations between various tasks.

Other techniques focus on balancing the influence of different modalities or tasks by modulating the loss function's weights. Approaches like dynamic weighting (Abels et al., 2019) or adaptive weighting (Walia et al., 2019) allocate weights based on the modality or task's difficulty, significance, or relevance, thereby optimizing the model more effectively.

New loss functions have also been designed to boost the alignment or fusion of modalities. These include methods based on contrast learning (Li et al., 2021a, 2022a), self-supervised learning (Alayrac et al., 2020), and cross-task learning (Chen et al., 2017; Hu et al., 2020). These techniques utilize various levels of information, such as

unimodal intra-modal data, cross-modal similarity, and cross-task shared information, to forge more effective loss functions that enhance the model's generalization capabilities.

2.3.2 Incorporating Quantum Theory in Multimodal Learning

Handling multimodal information and sentiment analysis involves understanding human cognition, a task where classical probabilistic methods often struggle. These traditional methods typically fall short in effectively capturing the dynamic interplay between modalities and contexts from a cognitive perspective. Quantum theory, however, has demonstrated its prowess in overcoming the limitations of classical probability theory in modeling human cognition. It not only achieves superior performance but also offers enhanced interpretability in this context (Zhang et al., 2020; Li et al., 2021b).

Several groundbreaking studies have explored quantum-inspired models for sentiment analysis and multimodal information processing. Gkoumas et al. (2021b) developed a quantum cognition-based fusion strategy. In this model, utterances are conceptualized as quantum states within a complex-valued emotional Hilbert space, with single-modal decisions represented as incompatible observables. This approach allows for an innovative handling of diverse emotional judgment scenarios.

Gkoumas et al. (2021a) introduced a quantum probability neural model specifically for video emotion analysis. The model uses the concept of entanglement, a form of inseparability in quantum mechanics, for the fusion of two modalities. It effectively captures both classical and non-classical correlations between these modalities by quantifying non-classical correlations accurately.

Li et al. (2021b) proposed a quantum-inspired network for dialogue emotion recognition. This network adeptly fuses multimodal data and integrates dialogue context to accurately identify emotions in each utterance. Additionally, Zhang et al. (2020) introduced a quantum-inspired multi-modal network (QMN) framework. This framework incorporates a density matrix-based CNN (Kalchbrenner et al., 2014), a quantum measurement-inspired influence model, and a quantum interference-inspired decision fusion method. It is designed to model both intra- and inter-utterance interactions, significantly enhancing emotion recognition in speakers.

2.4 Exploring Additional Relevant Research

2.4.1 Advancements in General-Purpose Modeling

Foundation models have garnered significant interest for their versatility across various downstream applications. Despite architectural similarities, most pre-trained models are generally optimized for specific tasks or modalities. Baevski et al. (2022) introduced a universal learning framework applicable to different modalities, yet it still relies on modality-specific encoders. Tsimpoukelli et al. (2021) demonstrated the transferability of in-context learning capabilities of frozen language models to vision-language settings. Alayrac et al. (2022) implemented a broad-spectrum understanding of images, videos, and text through a large-scale frozen language model. Reed et al. (2022) developed a multifaceted agent functioning as a multi-modal, multi-task, and multi-embodiment generalist policy. Furthermore, Hao et al. (2022) proposed the METALM model, leveraging a semi-causal language model as a universal interface to various foundation models. This model integrates a suite of pre-trained encoders to process diverse modalities and interact with the language model, thereby facilitating the resolution of a range of tasks without necessitating individual task retraining.

2.4.2 Scaling Capabilities and Flexibility

Pre-trained models have proven efficacious in both vision and language tasks, as highlighted by (Dosovitskiy et al., 2021; Zhai et al., 2022) in vision and (Raffel et al., 2020; Kaplan et al., 2020) in language. To scale effectively, a flexible task interface is essential for large language models to excel in diverse tasks. Chen et al. (2022) introduced PaLI, a model that concurrently processes language and vision. PaLI generates text from visual and textual inputs, handling an array of vision, language, and multimodal tasks across different languages. The model utilizes a scaled-up 4B parameter Vision Transformer as its vision backbone, optimizing compute resources through the use of pre-trained models. Lu et al. (2022a) proposed Unified-IO, a Seq2Seq model capable of executing various tasks using a single architecture without necessitating task or modality-specific components. This unification is achieved by converting every task's output into a sequence of discrete tokens. Unified-IO demonstrates robust performance

across diverse benchmarks, including GRIT benchmark, NYUv2-Depth, ImageNet, VQA2.0, OK-VQA, Swig, VizWizGround, BoolQ, and SciTail, without the need for task-specific fine-tuning.

2.4.3 Advancing Efficiency and Flexibility in Multimodal Frameworks

For the practical deployment of multimodal frameworks, developing an efficient and adaptable framework is critical. In this context, Li et al. (2022b) introduced FLZP, a novel and efficient language-image pre-training method. This approach enhances the learning capabilities and efficiency of CLIP by incorporating MAE. The FLZP model undertakes an exploration into the scaling of model size, dataset size, and training epochs, yielding impressive outcomes across a variety of vision-language benchmarks.

Additionally, Zhu et al. (2022c) proposed Uni-Perceiver, a unified architecture for generic perception tailored for zero-shot and few-shot tasks. This model harmonizes vision and language modalities into a singular framework, demonstrating robust performance across a spectrum of diverse tasks. The utilization of a large-scale dataset encompassing images and text in its training phase enables it to learn rich representations, which can be further fine-tuned for specific downstream applications.

Moreover, Zhu et al. (2022a) introduced Uni-Perceiver-MoE, a sparse generalist model featuring conditional mixture of expertse. This model integrates vision and language modalities into a single unified system. Uni-Perceiver-MoE has garnered considerable attention for its ability to efficiently handle a wide range of tasks while maintaining a unified approach. The flexibility and efficiency of these frameworks mark significant strides in the field of multimodal learning, paving the way for more practical and versatile applications.

2.5 Exploring the Landscape of Large Language Models in Multimodal Learning

The arena of multimodal learning has been revolutionized by the advent of large-scale language model pretraining, which has demonstrated exceptional performance in a variety of downstream tasks, sparking widespread research interest. A key distinguishing factor among these models is their pretraining objectives and architectural designs. Notably, GPT series (Radford et al., 2018, 2019; Brown et al., 2020) have pioneered in pretraining causal language models using decoder-only Transformers, revealing remarkable capabilities in few-shot and in-context learning.

GPT-4 (Tasar and Tasar, 2023), as one of the most prominent models in this field, has emerged as a titan with its 1.5 trillion parameters. It uniquely processes both image and text inputs, producing text outputs. Trained on a comprehensive multimodal dataset, including web texts, images, videos, and audio, GPT-4 has demonstrated proficiency in tasks such as natural language understanding and generation, image captioning, visual question answering, and more. Its performance on various benchmarks has been compared to human-level proficiency, as detailed in Table 3.

Our analysis primarily focuses on the application of these large-scale models in multimodal learning, categorized into four main approaches. (1) Freezing LLMs and training additional structures like visual encoders to adapt them for specific tasks, exemplified by mPLUG-Owl (Ye et al., 2023), LLaVA (Liu et al., 2023a), Mini-GPT4 (Zhu et al., 2023), and PaLM-E (Driess et al., 2023). (2) Converting visual information into textual input for LLMs, as seen in PICA (Yang et al., 2022), PromptCap (Hu et al., 2022), and ScienceQA (Lu et al., 2022b). (3) Utilizing visual modalities to influence LLM decoding, such as in ZeroCap (Tewel et al., 2022). (4) Employing LLMs as a central hub for integrating and leveraging multimodal models, like VisualChatGPT (Wu et al., 2023a) and MM-REACT (Yang et al., 2023b).

Due to the rapid development of large language models (LLMs), there is an increasing trend toward using LLMs as backbones for constructing large-scale multimodal models. These models primarily focus on the fusion of vision and text modalities, aiming to create versatile and widely applicable multimodal deep learning models. Table 4 in Appendix A presents an overview of the key technologies and applications pertinent to Multimodal Large Language Models. This includes various innovative approaches like Multimodal Instruction Tuning (M-IT), Multimodal In-Context Learning (M-ICL), Multimodal Chain-of-Thought (M-CoT), LLM-Aided Visual Reasoning (LAVR), Multimodal Hallucination (MMH), and Multimodal RLHF (M-RLHF). These methodologies illustrate the diverse ways in which LLMs can be leveraged in multimodal contexts.

Models	Perception										Cognition			
	Existence	Count	Position	Color	Poster	Celebrity	Scene	Landmark	Artwork	OCR	Commonsense Reasoning	Numerical Calculation	Text Translation	Code Reasoning
mPLUG-Owl (Ye et al., 2023)	120.00	50.00	50.00	55.00	136.05	100.29	135.5	159.25	96.25	65.00	78.57	60.00	80.00	57.50
LLaVA (Liu et al., 2023a)	185.00	155.00	133.00	170.00	160.54	152.94	161.25	170.50	117.75	125.00	127.86	42.50	77.50	47.50
MiniGPT-4 (Zhu et al., 2023)	68.33	55.00	43.33	75.00	41.84	54.41	71.75	54.00	60.50	57.50	59.29	45.00	-	40.00
MMICL (Zhao et al., 2023a)	170.00	160.00	81.67	156.67	146.26	141.76	153.75	136.13	135.50	100.00	136.43	82.50	132.50	77.50
Gemini Pro (Team et al., 2023)	175.00	131.67	90.00	163.33	164.97	147.35	144.75	158.75	135.75	185.00	129.29	77.50	145.00	85.00
LLAMA-Adapter V2 (Gao et al., 2023)	185.00	133.33	56.67	118.33	147.96	136.76	156.25	167.84	123.75	102.50	106.43	47.50	112.50	90.00
GPT-4V (Tasar and Tasar, 2023)	190.00	160.00	95.00	150.00	192.18	-	151.00	-	148.00	185.00	142.14	130.00	75.00	170.00

Table 3: Model performance on MME benchmark. MME measures both perception and cognition abilities on a total of 14 subtasks. Each of the 14 subtasks is worth 200 points. The score is the sum of the accuracy and the accuracy+. We adopted reported results from published literature (Fu et al., 2023).

3 Challenges and Future Directions in Multimodal Deep Learning

3.1 Addressing Imbalance in Multimodal Learning Environments

Within the realm of multimodal deep learning, the challenge of imbalance learning stands as a formidable obstacle, arising when the distribution of data across various modalities or classes assumes an uneven, skewed configuration. This intricate concern materializes in instances where certain modalities assert dominance, eclipsing others in frequency, or when specific classes are endowed with a surplus of samples or features. The repercussions of such imbalance reverberate through the learning process, potentially engendering bias and suboptimal performance. This manifests as an overemphasis on majority modalities or classes, and a corresponding neglect of their minority counterparts.

Interestingly, even the expanse of large-scale models is not immune to the clutches of this issue, as the constraints of data quality and quantity remain steadfast. In the tapestry of training data, certain modalities might be absent, marred by noise, or misaligned—a predicament that casts shadows on the model's capacity to glean meaningful representations and intermodal interactions. Furthermore, class disparities can introduce complexities of their own, ushering in intricate and diverse patterns that beset the model's predictive accuracy and confidence.

The endeavor to judiciously sample imbalanced data from diverse modalities while preserving coherence between them stands as an intricate quandary. Novel strategies are on the horizon, encompassing modality-specific sampling tactics, holistic joint sampling methodologies that take into account the intermodal relationships, and the integration of generative models to conjure synthetic samples. These prospective solutions beckon for rigorous investigation and refinement, sparking the evolution of effective approaches that can adeptly surmount this challenge.

3.2 Advancing Domain Generalization in AI with Multimodal Few-Shot Learning

Multimodal Few-Shot Learning stands at the forefront of artificial intelligence research, rapidly expanding as it delves into leveraging multiple data types like images, text, and audio. This field aims to equip intelligent systems with the ability to quickly comprehend new tasks or concepts from minimal data input. The ultimate goal is to develop agents that can effortlessly navigate and adapt to a wide array of environments, transcending the limitations of specific domains, languages, and modalities.

The progression of Multimodal Few-Shot Learning relies heavily on pushing the limits of what's possible by venturing into complex scenarios that test the boundaries of this technology. Central to this progress is the ability to effectively align and integrate data from various modalities. This alignment is crucial for the success of few-shot learning initiatives, as it forms the basis for teaching intelligent agents to grasp new concepts with limited examples.

Moreover, the advancement of this field demands rigorous testing and evaluation of these multimodal few-shot learning models against the realities of practical, real-world data and applications. A significant challenge in this endeavor is to imbue these models with robust domain generalization capabilities. This involves preparing them to perform well in unfamiliar, out-of-domain situations. Enhancing the models' ability to generalize across varied domains is a key objective, aiming to forge truly intelligent agents that can perform effectively in diverse and unforeseen environments.

3.3 Advancing Multimodal Category Reasoning

Multimodal Category Reasoning, a vibrant field in AI, seeks to integrate diverse modalities like text, images, and audio for complex reasoning tasks including question answering and classification. This integration marks a crucial evolution towards a higher level of intelligence in AI, enhancing the scope and adaptability of multimodal deep learning across various domains. Key to advancing in this field is enhancing the interpretability of multimodal reasoning models. A pivotal approach here is the adoption of multimodal thought chains, designed to minimize reasoning errors and create a coherent, interconnected thought process across modalities, fostering a unified reasoning framework. Additionally, integrating high-quality external multimodal knowledge graphs into these models is essential for addressing model hallucinations and ensuring accuracy in reasoning. This integration not only bolsters the models' cognitive capabilities but also guarantees precision and reliability in their outcomes, significantly enriching their reasoning potential.

3.4 Navigating Towards Integrated Multimodal Processing

In the current landscape of large-scale multimodal models, a predominant approach involves a multi-tiered process for interpreting user inputs. Typically, this starts with converting a user's query into a text-based format, followed by applying various visual or other modal tools for generating results. While this method effectively combines different unimodal models for multimodal capabilities, its sequential or multi-tiered nature is prone to the compounding of errors, potentially leading to misleading or incorrect outcomes.

In contrast, human cognition naturally processes multiple modalities simultaneously, using this integration for judgment and logical reasoning. Thus, the quest for true artificial general intelligence challenges us to develop a unified model within a single framework. This model would need to efficiently blend information from various modalities, allowing for comprehensive reasoning and consistent decision-making while avoiding the introduction of noise from additional modalities.

The goal is to mimic the human brain's skill in merging information from different sensory inputs with a similar level of accuracy in aligning different modalities in large-scale multimodal models. Pursuing this path opens the door to a new generation of computational systems that reflect the human ability to synergistically process and incorporate information from multiple sources. In different scenarios, careful consideration of the appropriate backbones is essential, as is the thoughtful selection of prompts and their embedding strategies. This nuanced approach is crucial for the development of versatile and effective multimodal AI systems.

3.5 Prioritizing Safety, Explainability, and Modality-Specific Considerations in Large-Scale Multimodal Models

In the realm of large-scale multimodal models, safety and explainability are paramount, but equally important is the need to address the unique limitations of different modalities. Each modality – be it text, image, or audio – has its own set of constraints and potential biases. Addressing these modality-specific challenges is crucial for enhancing the model's overall effectiveness and reliability.

A key strategy is to equip these models with capabilities to identify and counteract the weaknesses inherent in individual modalities. This approach enhances the model's overall safety and reliability, and ensures more controlled and accurate content generation. Furthermore, incorporating these considerations into the model's framework involves creating metrics that assess how well these challenges are managed. Such metrics will guide the development of multimodal models that are not only high-performing but also adhere to strict safety and ethical standards, resulting in responsible, user-centric, and trustworthy AI systems.

4 Conclusion

In this paper, we provide a focused synthesis of multimodal deep learning, emphasizing its integration of diverse data types like text, speech, and images, especially in the context of large language models. Our exploration of evolving MMDL methodologies, including advanced model architectures and data handling techniques, offers a clear view of the field's current state and future potential. We also highlighted the diverse applications of MMDL, identifying challenges and opportunities in the LLM era. Our comparative analysis of current models provides a framework for understanding their performance and guides future research directions.

5 Limitations

Despite our comprehensive approach, this survey is subject to several limitations. Firstly, due to the rapidly evolving nature of the field of multimodal deep learning, our review may not encompass all recent developments and methodologies. The field's rapid progression often leads to the emergence of new models and techniques shortly after a literature review is conducted. Secondly, while we strive for a thorough comparative analysis of current models, the assessment is limited by the availability and accessibility of benchmark datasets and evaluation metrics. As a result, some potentially impactful models might not be included in our analysis due to the lack of comprehensive evaluation data.

References

Amal Abdulrahman and Debbie Richards. 2022. Is natural necessary? human voice versus synthetic voice for intelligent virtual agents. *Multimodal Technol. Interact.*, 6(7):51.

Axel Abels, Diederik Roijers, Tom Lenaerts, Ann Nowé, and Denis Steckelmacher. 2019. Dynamic weights in multi-objective deep reinforcement learning. In *International Conference on Machine Learning*, pages 11–20. PMLR.

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

Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. 2020. Self-supervised multimodal versatile networks. *Advances in Neural Information Processing Systems*, 33:25–37.

Dora E Angelaki and Kathleen E Cullen. 2008. Vestibular system: the many facets of a multimodal sense. *Annu. Rev. Neurosci.*, 31:125–150.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Yitzhak Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *CoRR*, abs/2308.01390.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912.

Assaf Ben-Kish, Moran Yanuka, Morris Alper, Raja Giryes, and Hadar Averbuch-Elor. 2023. Mocha: Multi-objective reinforcement mitigating caption hallucinations. *CoRR*, abs/2312.03631.

Luisa Brinkschulte, Stephan Schlögl, Alexander Monz, Pascal Schöttle, and Matthias Janetschek. 2022. Perspectives on socially intelligent conversational agents. *Multimodal Technol. Interact.*, 6(8):62.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. 2023. Honeybee: Locality-enhanced projector for multimodal LLM. *CoRR*, abs/2312.06742.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023a. Shikra: Unleashing multimodal llm's referential dialogue magic. *CoRR*, abs/2306.15195.

Shizhe Chen, Qin Jin, Jinming Zhao, and Shuai Wang. 2017. Multimodal multi-task learning for dimensional and continuous emotion recognition. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 19–26.

Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. 2023b. LL3DA: visual interactive instruction tuning for omni-3d understanding, reasoning, and planning. *CoRR*, abs/2311.18651.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2022. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Zhi Chen, Yongguo Liu, Yun Zhang, Qiaoqin Li, and Alzheimer's Disease Neuroimaging Initiative. 2023c. Orthogonal latent space learning with feature weighting and graph learning for multimodal alzheimer's disease diagnosis. *Medical Image Anal.*, 84:102698.

Zhiyang Chen, Yousong Zhu, Yufei Zhan, Zhaowen Li, Chaoyang Zhao, Jinqiao Wang, and Ming Tang. 2023d. Mitigating hallucination in visual language models with visual supervision. *CoRR*, abs/2311.16479.

Xiangxiang Chu, Limeng Qiao, Xinyang Lin, Shuang Xu, Yang Yang, Yiming Hu, Fei Wei, Xinyu Zhang, Bo Zhang, Xiaolin Wei, and Chunhua Shen. 2023. Mobilevlm: A fast, strong and open vision language assistant for mobile devices.

John M Cioffi and John AC Bingham. 1994. A data-driven multitone echo canceller. *IEEE transactions on communications*, 42(10):2853–2869.

Neil Cohn. 2016. A multimodal parallel architecture: A cognitive framework for multimodal interactions. *Cognition*, 146:304–323.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Weixi Feng, Wanrong Zhu, Tsu-Jui Fu, Varun Jampani, Arjun R. Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2023. Layoutgpt: Compositional visual planning and generation with large language models. *CoRR*, abs/2305.15393.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

Eleanor Jack Gibson. 1969. Principles of perceptual learning and development.

Dimitrios Gkoumas, Qiuchi Li, Yijun Yu, and Dawei Song. 2021a. An entanglement-driven fusion neural network for video sentiment analysis. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 1736–1742. International Joint Conferences on Artificial Intelligence Organization.

Dimitris Gkoumas, Qiuchi Li, Shahram Dehdashti, Massimo Melucci, Yijun Yu, and Dawei Song. 2021b. Quantum cognitively motivated decision fusion for video sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 827–835.

Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 14953–14962. IEEE.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. 2023. Onellm: One framework to align all modalities with language. *CoRR*, abs/2312.03700.

Xiaoshuai Hao, Yi Zhu, Srikar Appalaraju, Aston Zhang, Wanqian Zhang, Bo Li, and Mu Li. 2023. Mixgen: A new multi-modal data augmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 379–389.

Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. 2022. Language models are general-purpose interfaces. *arXiv preprint arXiv:2206.06336*.

Jiabang He, Lei Wang, Yi Hu, Ning Liu, Hui Liu, Xing Xu, and Heng Tao Shen. 2023. ICL-D3IE: in-context learning with diverse demonstrations updating for document information extraction. *CoRR*, abs/2303.05063.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxuan Zhang, Juanzi Li, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. Cogagent: A visual language model for GUI agents. *CoRR*, abs/2312.08914.

Di Hu, Xuhong Li, Lichao Mou, Pu Jin, Dong Chen, Liping Jing, Xiaoxiang Zhu, and Dejing Dou. 2020. Cross-task transfer for geotagged audiovisual aerial scene recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 68–84. Springer.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2022. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*.

Ziniu Hu, Ahmet Iscen, Chen Sun, Kai-Wei Chang, Yizhou Sun, David A. Ross, Cordelia Schmid, and Alireza Fathi. 2023. AVIS: autonomous visual information seeking with large language models. *CoRR*, abs/2306.08129.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2023. OPERA: alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *CoRR*, abs/2311.17911.

Ray Jackendoff. 1987. On beyond zebra: The relation of linguistic and visual information. *Cognition*, 26(2):89–114.

Joonhyun Jeong. 2023. Hijacking context in large multi-modal models. *CoRR*, abs/2312.07553.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 655–665. The Association for Computer Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR.

Shashidhar G Koolagudi and K Sreenivasa Rao. 2012. Emotion recognition from speech: a review. *International journal of speech technology*, 15:99–117.

Shruti Rajendra Kshirsagar, Anurag Pendyala, and Tiago H. Falk. 2023. Task-specific speech enhancement and data augmentation for improved multimodal emotion recognition under noisy conditions. *Frontiers Comput. Sci.*, 5.

Santosh Kumar, Mithilesh Kumar Chaube, Saeed Hamood Alsamhi, Sachin Kumar Gupta, Mohsen Guizani, Raffaele Gravina, and Giancarlo Fortino. 2022. A novel multimodal fusion framework for early diagnosis and accurate classification of COVID-19 patients using x-ray images and speech signal processing techniques. *Comput. Methods Programs Biomed.*, 226:107109.

Dana Lahat, Tülay Adali, and Christian Jutten. 2015. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023a. MIMIC-IT: multi-modal in-context instruction tuning. *CoRR*, abs/2306.05425.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021a. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. 2023b. Silkie: Preference distillation for large visual language models.

Qiuchi Li, Dimitris Gkoumas, Alessandro Sordoni, Jian-Yun Nie, and Massimo Melucci. 2021b. Quantum-inspired neural network for conversational emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13270–13278.

Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. 2022b. Scaling language-image pre-training via masking. *arXiv preprint arXiv:2212.00794*.

Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramanan. 2023. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. *CoRR*, abs/2301.06267.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Zhaoyang Liu, Zeqiang Lai, Zhangwei Gao, Erfei Cui, Zhiheng Li, Xizhou Zhu, Lewei Lu, Qifeng Chen, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2023b. Controlllm: Augment language models with tools by searching on graphs. *CoRR*, abs/2310.17796.

Zhijian Liu, Zhanghao Wu, Chuang Gan, Ligeng Zhu, and Song Han. 2020. Datamix: Efficient privacy-preserving edge-cloud inference. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 578–595. Springer.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2022a. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*.

Jiaying Lu, Jinmeng Rao, Kezhen Chen, Xiaoyuan Guo, Yawen Zhang, Baochen Sun, Carl J. Yang, and Jie Yang. 2023. Evaluation and mitigation of agnosia in multimodal large language models. *CoRR*, abs/2309.04041.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022b. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022c. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. 2023. Dolphins: Multimodal language model for driving. *CoRR*, abs/2312.00438.

Petra Magnusson and Anna-Lena Godhe. 2019. Multimodality in language education–implications for teaching. *Designs for Learning*, 11(1):127–137.

Weiguang Mao, Javad Rahimikollu, Ryan Hausler, and Maria Chikina. 2021. Dataremix: a universal data transformation for optimal inference from gene expression datasets. *Bioinformatics*, 37(7):984–991.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. 2023. Med-flamingo: a multimodal medical few-shot learner. *CoRR*, abs/2307.15189.

Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. 2023. Embodiedgpt: Vision-language pre-training via embodied chain of thought. *CoRR*, abs/2305.15021.

Bence Nanay. 2018. Multimodal mental imagery. *Cortex*, 105:125–134.

Shaylene E. Nancekivell, Xin Sun, Susan A. Gelman, and Priti Shah. 2021. A slippery myth: How learning style beliefs shape reasoning about multimodal instruction and related scientific evidence. *Cogn. Sci.*, 45(10).

Maja Pantic and Leon JM Rothkrantz. 2003. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390.

Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. 2022. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*.

R Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. 2005. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107.

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

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. 2022. A generalist agent. *arXiv preprint arXiv:2205.06175*.

Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 14974–14983. IEEE.

Linda Smith and Michael Gasser. 2005. The development of embodied cognition: Six lessons from babies. *Artificial life*, 11(1-2):13–29.

Cees GM Snoek, Marcel Worring, J Geusebroek, Dennis C Koelma, Frank J Seinstra, and Arnold WM Smeulders. 2006. The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1678–1689.

Jingkuan Song, Yuyu Guo, Lianli Gao, Xuelong Li, Alan Hanjalic, and Heng Tao Shen. 2018. From deterministic to generative: Multimodal stochastic rnns for video captioning. *IEEE transactions on neural networks and learning systems*, 30(10):3047–3058.

Jabeen Summaira, Xi Li, Amin Muhammad Shoib, and Abdul Jabbar. 2022. A review on methods and applications in multimodal deep learning. *CoRR*, abs/2202.09195.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023. Aligning large multimodal models with factually augmented RLHF. *CoRR*, abs/2309.14525.

Krunoslav Michael Sveric, Ivan Platzek, Elena Golgor, Ralf-Thorsten Hoffmann, Axel Linke, and Stefanie Jellinghaus. 2022. Purposeful use of multimodality imaging in the diagnosis of caseous mitral annular calcification: a case series report. *BMC Medical Imaging*, 22(1):7.

Davut Emre Tasar and Ceren Ocal Tasar. 2023. Bridging history with AI A comparative evaluation of GPT 3.5, gpt4, and googlebard in predictive accuracy and fact checking. *CoRR*, abs/2305.07868.

Gemini Team, Rohan Anil, and Sebastian Borgeaud et. al. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. 2022. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17918–17928.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Miguel Vasco, Hang Yin, Francisco S. Melo, and Ana Paiva. 2022. Leveraging hierarchy in multimodal generative models for effective cross-modality inference. *Neural Networks*, 146:238–255.

Gurjit Singh Walia, Gaurav Jain, Nipun Bansal, and Kuldeep Singh. 2019. Adaptive weighted graph approach to generate multimodal cancelable biometric templates. *IEEE transactions on information forensics and security*, 15:1945–1958.

Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. 2023. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. *CoRR*, abs/2312.01701.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2022. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*.

Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2023. Vary: Scaling up the vision vocabulary for large vision-language models. *CoRR*, abs/2312.06109.

Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023a. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.

Penghao Wu and Saining Xie. 2023. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*.

Tsung-Han Wu, Giscard Biamby, David Chan, Lisa Dunlap, Ritwik Gupta, Xudong Wang, Joseph E. Gonzalez, and Trevor Darrell. 2023b. See, say, and segment: Teaching lmms to overcome false premises. *CoRR*, abs/2312.08366.

Jiarui Xu, Xingyi Zhou, Shen Yan, Xiuye Gu, Anurag Arnab, Chen Sun, Xiaolong Wang, and Cordelia Schmid. 2023a. Pixel aligned language models. *CoRR*, abs/2312.09237.

Ruize Xu, Ruoxuan Feng, Shi-Xiong Zhang, and Di Hu. 2023b. Mmcosine: Multi-modal cosine loss towards balanced audio-visual fine-grained learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, and Nan Duan. 2022. Bridge-tower: Building bridges between encoders in vision-language representation learning. *arXiv preprint arXiv:2206.08657*.

Meichao Yan, Yu Wen, Qingxuan Shi, and Xuedong Tian. 2022. A multimodal retrieval and ranking method for scientific documents based on HFS and xlnet. *Sci. Program.*, 2022:5373531:1–5373531:11.

Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. 2021. Taco: Token-aware cascade contrastive learning for video-text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11562–11572.

Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Xin Geng. 2023a. Exploring diverse in-context configurations for image captioning. *CoRR*, abs/2305.14800.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. *Proceedings of the AAAI Conference on Artificial Intelligence*, page 3081–3089.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023b. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. 2023a. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. *CoRR*, abs/2311.13614.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. 2023b. RLHF-V: towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *CoRR*, abs/2312.00849.

Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. 2021. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 558–567.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. 2022. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113.

Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Yu Qiao, Peng Gao, and Hongsheng Li. 2023. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 15211–15222. IEEE.

Yazhou Zhang, Dawei Song, Xiang Li, Peng Zhang, Panpan Wang, Lu Rong, Guangliang Yu, and Bo Wang. 2020. A quantum-like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis. *Information Fusion*, 62:14–31.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2023a. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2023b. MMICL: empowering vision-language model with multi-modal in-context learning. *CoRR*, abs/2309.07915.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *CoRR*, abs/2310.16436.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Jinguo Zhu, Xizhou Zhu, Wenhai Wang, Xiaohua Wang, Hongsheng Li, Xiaogang Wang, and Jifeng Dai. 2022a. Uni-perceiver-moe: Learning sparse generalist models with conditional moes. *arXiv preprint arXiv:2206.04674*.

Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyao Zeng, Shanghang Zhang, and Peng Gao. 2022b. Pointclip V2: adapting CLIP for powerful 3d open-world learning. *CoRR*, abs/2211.11682.

Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. 2022c. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16804–16815.

Mingchen Zhuge, Haozhe Liu, Francesco Faccio, Dylan R. Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, Hasan Abed Al Kader Hammoud, Vincent Herrmann, Kazuki Irie, Louis Kirsch, Bing Li, Guohao Li, Shuming Liu, Jinjie Mai, Piotr Piekos, Aditya Ramesh, Imanol Schlag, Weimin Shi, Aleksandar Stanic, Wenyi Wang, Yuhui Wang, Mengmeng Xu, Deng-Ping Fan, Bernard Ghanem, and Jürgen Schmidhuber. 2023. Mindstorms in natural language-based societies of mind. *CoRR*, abs/2305.17066.

A Key Technologies and Applications of the Multimodal Large Language Model

Table 4 presents an overview of the key technologies and applications pertinent to Multimodal Large Language Models.

Categories	Title	Venue	Code
Multimodal Instruction Tuning	**MobileVLM** (Chu et al., 2023)	**arXiv**	✓
	Vary (Wei et al., 2023)	**arXiv**	✓
	CogAgent (Hong et al., 2023)	**arXiv**	✓
	Pixel Aligned Language Models (Xu et al., 2023a)	**ICLR**	-
	See, Say, and Segment (Wu et al., 2023b)	**arXiv**	-
	Honeybee (Cha et al., 2023)	**arXiv**	✓
	Gemini (Team et al., 2023)	**Google**	✓
	OneLLM (Han et al., 2023)	**AAAI**	✓
	Dolphins (Ma et al., 2023)	**arXiv**	✓
	LL3DA (Chen et al., 2023b)	**arXiv**	✓
Multimodal In-Context Learning	**Hijacking Context in Large Multi-modal Models** (Jeong, 2023)	**arXiv**	-
	Prompting Large Language Models with Answer Heuristics for Knowledge-based Visual Question Answering (Shao et al., 2023)	**CVPR**	✓
	MMICL (Zhao et al., 2023b)	**arXiv**	✓
	OpenFlamingo (Awadalla et al., 2023)	**arXiv**	✓
	Med-Flamingo (Moor et al., 2023)	**arXiv**	✓
	AVIS (Hu et al., 2023)	**arXiv**	-
	MIMIC-IT (Li et al., 2023a)	**arXiv**	✓
	Exploring Diverse In-Context Configurations for Image Captioning (Yang et al., 2023a)	**NeurIPS**	✓
	ICL-D3IE (He et al., 2023)	**ICCV**	✓
	Visual Programming (Gupta and Kembhavi, 2023)	**CVPR**	✓
Multimodal Chain-of-Thought	**DDCoT** (Zheng et al., 2023)	**NeurIPS**	✓
	Shikra (Chen et al., 2023a)	**arXiv**	✓
	EmbodiedGPT (Mu et al., 2023)	**arXiv**	✓
	Learn to Explain (Lu et al., 2022c)	**NeurIPS**	✓
LLM-Aided Visual Reasoning	**V*** (Wu and Xie, 2023)	**arXiv**	✓
	Prompt, Generate, then Cache (Zhang et al., 2023)	**CVPR**	✓
	LayoutGPT (Feng et al., 2023)	**arXiv**	✓
	ControlLLM (Liu et al., 2023b)	**arXiv**	✓
	Mindstorms in Natural Language-Based Societies of Mind (Zhuge et al., 2023)	**NeurIPS**	✓
	PointCLIP V2 (Zhu et al., 2022b)	**CVPR**	✓
Multimodal Hallucination	**MOCHa** (Ben-Kish et al., 2023)	**EUSIPCO**	✓
	Mitigating Fine-Grained Hallucination by Fine-Tuning Large Vision-Lanquage Models with Caption Rewrites (Wang et al., 2023)	**arXiv**	✓
	RLHF-V (Yu et al., 2023b)	**arXiv**	✓
	OPERA (Huang et al., 2023)	**arXiv**	✓
	Mitigating Hallucination in Visual Language Models with Visual Supervision (Chen et al., 2023d)	**arXiv**	-
	HalluciDoctor (Yu et al., 2023a)	**arXiv**	✓
	Evaluating Object Hallucination in Large Vision-Language Models (Lu et al., 2023)	**EMNLP**	✓
Multimodal RLHF	**Silkie** (Li et al., 2023b)	**arXiv**	✓
	RLHF-V (Yu et al., 2023b)	**arXiv**	✓
	Aligning Large Multimodal Models with Factually Augmented RLHF (Sun et al., 2023)	**arXiv**	✓

Table 4: Key technologies and applications of the Multimodal Large Language Model, including Multimodal Instruction Tuning (M-IT), Multimodal In-Context Learning (M-ICL), Multimodal Chain-of-Thought (M-CoT), LLM-Aided Visual Reasoning (LAVR), Multimodal Hallucination (MMH), and Multimodal RLHF (M-RLHF).