A Survey on Multimodal Large Language Models

***Abstract*—In recent years, Multimodal Large Language Models (MLLMs) have gradually become an important research direction in the field of artificial intelligence. Traditional unimodal language models primarily rely on textual data, and although they are capable of handling language tasks, their performance is limited when dealing with non-text data such as images and audio. MLLMs integrate various forms of data, such as text, images, audio, and video, significantly improving performance in multimodal tasks, including visual-language understanding, cross-modal reasoning, and vision-based generation tasks. These models provide a more comprehensive ability to understand and reason with information, driving the diverse application of intelligent systems. This paper first reviews the basic architecture of current MLLM research, providing a detailed introduction to the model training strategies (pre-training, instruction-tuning, and alignment tuning) and data processing methods, while also exploring common evaluation criteria for multimodal tasks. Next, the paper discusses the potential for expanding MLLMs, including how to optimize models to tackle more complex tasks such as multimodal reasoning, unsupervised learning, and cross-modal reasoning. Additionally, we analyze key challenges in current MLLM research, focusing on issues like modality fusion and techniques for mitigating multimodal hallucination. Finally, the paper looks ahead to future research directions for MLLMs, proposing potential breakthroughs in technology.**

***Index Terms*—Multimodal, Large Language Models, Vision-Language Models**

I. Introduction

In the real world, information exists not only as text but also in multimodal forms including images, audio, and video. Traditional language models primarily rely on textual data for training, enabling them to understand and generate linguistic information to a certain extent. However, their limited capacity for processing other modalities such as visual and auditory information often hinders comprehensive understanding of cross-modal correlations and latent value in multimodal contexts. To address this limitation, researchers have recently proposed Multimodal Large Language Models (MLLMs), whose core objective is to establish intermodal connections and enable joint reasoning across textual, visual, and auditory information by fusing data from multiple modalities. Built upon Large Language Models (LLMs) as their foundation, MLLMs demonstrate the capability to process, reason with, and output information from diverse modalities. Their development not only significantly enhances model performance in multimodal tasks but also extends our understanding of information processing from unimodal to complex multimodal scenarios, thereby improving intelligent systems' generalization capabilities and application potential.

As an emerging research frontier, multimodal large language models have achieved notable progress in recent years, yet numerous challenges remain unaddressed for practical implementation. A primary challenge lies in developing effective cross-modal fusion mechanisms that ensure seamless interaction and information flow between modalities, which is an ongoing research difficulty. Additionally, model performance is constrained by the inherent complexity and diversity of multimodal data, coupled with limited availability of high-quality training datasets. Current multimodal models also suffer from weak interpretability, with opaque decision-making processes and reasoning rationales that hinder their reliable deployment in real-world applications. Despite these challenges, systematic surveys comprehensively analyzing the evolution, technical obstacles, and future directions of MLLMs remain relatively scarce.

To address these research gaps, this paper conducts a systematic literature review analyzing landmark studies in multimodal large language models from both domestic and international sources. Through comprehensive analysis of existing research, we construct a holistic knowledge framework to facilitate comparative understanding of different approaches, identify critical technological advancements, and highlight unresolved issues. Our survey focuses on key aspects including model architectures, training strategies (e.g. pretraining, instruction fine-tuning, and alignment fine-tuning), data processing methodologies, and evaluation protocols, ensuring theoretical rigor while providing references for subsequent research.

The paper first presents the fundamental architecture of MLLMs, detailing their core structures, training paradigms, data processing techniques, and standardized evaluation metrics. Subsequently, we discuss potential improvements and extensions of current MLLMs, emphasizing strategies to mitigate multimodal hallucination during complex task execution, thereby enhancing model robustness and trustworthiness. Finally, based on a comprehensive analysis of existing research, we systematically examine the key challenges in MLLM development, including effective modality fusion, data diversity requirements, and model interpretability. Additionally, we propose future research directions to advance this critical field.

II. ARCHITECTURE

A typical Multimodal Large Language Model (MLLM) comprises three core modules: a pre-trained modality encoder, a pre-trained LLM, and a modality interface to connect them. Certain architectures may additionally incorporate a generator module to synthesize non-textual outputs such as images or video frames. Fundamentally, the LLM serves as the cognitive engine of the MLLM architecture, analogous to the human brain's reasoning system, responsible for cross-modal understanding and logical inference. In most implementations, the

LLM component remains pretrained and kept frozen during multimodal adaptation, thereby preserving its acquired linguistic capabilities without further parameter updates under multimodal inputs.

The multimodal encoder operates by projecting diverse modality inputs into a unified embedding space. For visual processing, standard implementations employ Vision Transformers (ViTs) from frameworks like CLIP [1], which effectively map image content into vector representations congruent with linguistic expressions. These encoders transform input images or other modality-specific data into high-dimensional embeddings, enabling joint processing across modalities within a shared semantic space. However, inherent disparities between multimodal encoders (e.g. CLIP's tokenizers and pretraining schemas) and LLMs create semantic gaps in cross-modal alignment. Consequently, the critical role of the multimodal interface lies in translating encoder-generated embeddings into LLM-compatible textual representations. This conversion necessitates not only geometric alignment in the embedding space but also preservation of cross-modal semantic consistency to ensure the LLM can reliably leverage these representations for reasoning and generation tasks.

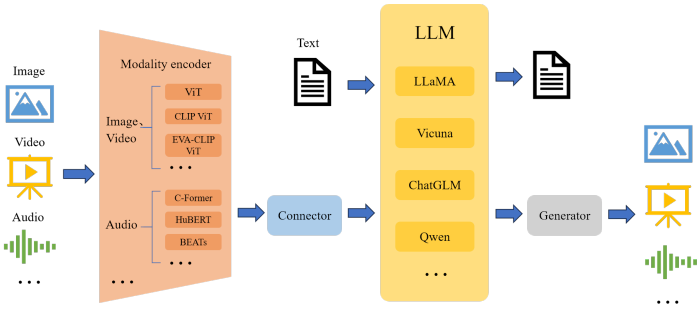

Fig. 1. An illustrative diagram of a typical MLLM architecture, consisting of an encoder, a connector, and an LLM. An optional generator can be appended to the LLM to generate additional modalities (e.g. images, audio) beyond text.

A. Modality encoder

A modality encoder serves as a projection module that transforms raw data from heterogeneous modalities (e.g. images, audio, video) into structured, semantically meaningful embeddings within a unified latent space. By bridging the "modality gap," these encoders enable Large Language Models (LLMs) to process non-linguistic inputs via their pretrained text-centric architectures.

a) Image Encoders: Vision Transformers (ViTs) and their derivatives dominate image encoding. ViTs partition images into fixed-size patches, linearly projected into token sequences processed through self-attention layers, effectively capturing global visual relationships. CLIP-ViT [1], an extension of ViT, is co-trained with a text encoder via contrastive learning on image-text pairs, aligning visual and textual embeddings in a shared space. This dual-training strategy empowers CLIP-ViT to generate image representations semantically congruent with linguistic contexts, making it a cornerstone for vision-language integration in many MLLMs.

b) Audio Encoders: Transformer-based models like Wav2Vec 2.0 [2] and CLAP (Contrastive Language-Audio Pretraining) [3] are state-of-the-art. Unlike CNNs for spectrogram processing, Wav2Vec 2.0 operates directly on raw waveforms using self-supervised learning. It masks segments of audio signals and trains the model to predict latent representations, learning robust acoustic features. For multimodal alignment, CLAP extends CLIP's paradigm to audio, jointly training audio and text encoders via contrastive loss on audio-caption pairs, thereby mapping audio signals to language-compatible embeddings.

c) Video Encoders: Video encoding demands modeling spatiotemporal dynamics. The TimeSformer [4] architecture adapts ViTs for video by factorizing self-attention into spatial and temporal dimensions, capturing both intra-frame features and inter-frame motion. For efficiency, VideoMAE [5] applies masked autoencoding to spatiotemporal patches, reconstructing corrupted video content during pretraining. These approaches enable holistic video representation learning while maintaining compatibility with transformer-based MLLMs.

B. Pre-trained LLM

Large Language Models (LLMs) serve as the cognitive core of Multimodal Large Language Models (MLLMs). This enables MLLMs to inherit critical LLM capabilities such as zero-shot generalization, few-shot learning, chain-of-thought reasoning, and instruction-following abilities. In practice, fine-tuning pre-trained LLMs is typically more efficient and practical for MLLM development, rather than training LLMs from scratch. During pre-training, LLMs acquire extensive world knowledge through unsupervised learning on massive text corpora, resulting in robust generalization and reasoning capabilities.

Among mainstream publicly available LLMs, Decoder-Only architectures dominate due to their superior generative capacity, particularly suited for tasks requiring autoregressive text generation. Representative open-source LLMs such as the LLaMA [6] and Vicuna [7] families heavily rely on English-centric training corpora, which limits their multilingual competence—especially under non-English scenarios (e.g. Chinese). In contrast, the Qwen [8] series exemplifies a bilingual LLM explicitly optimized for both English and Chinese, demonstrating enhanced potential in multilingual applications.

Notably, scaling LLM parameters has proven instrumental for performance gains, analogous to increasing input resolution. Specifically, Liu et al. [9] demonstrated that expanding model parameters from 7B to 13B yields comprehensive performance improvements across multiple standard benchmarks. This scaling enhances not only NLP task accuracy but also contextual understanding, text fluency, and problem-solving versatility. As LLMs grow larger, they capture richer linguistic patterns and contextual nuances, thereby achieving stronger generalization. Consequently, parameter scaling remains pivotal for advancing LLM capabilities.

C. Modality interface

The modality interface serves as the critical bridge between raw non-linguistic data (e.g. images, audio) and the textual processing framework of large language models (LLMs). Its primary role is to project heterogeneous modality-specific features into a unified embedding space that aligns with the LLM's linguistic understanding, enabling seamless multimodal reasoning. Two dominant methodologies have emerged: **learnable neural connectors** and **expert-driven language conversion**.

a) Learnable connectors: This approach introduces lightweight trainable adapter networks to align multimodal embeddings with LLM token spaces. A prevalent strategy employs learnable query tokens paired with attention-based architectures to extract cross-modal semantic relationships. For example, BLIP-2 [10] introduces a Transformer-based Q-Former, where a set of initialized query tokens interacts with visual features via self-attention and cross-attention layers. During pretraining, frozen pretrained encoders (e.g. CLIP-ViT for images and OPT for text) provide initial visual and textual representations. The Q-Former, initialized with randomly sampled query tokens, iteratively refines these queries via cross-attention layers to identify semantically relevant regions in visual features. In the alignment phase, the model optimizes a contrastive loss that minimizes the distance between learned visual queries and their corresponding textual embeddings, effectively creating a shared vision-language latent space. This approach balances flexibility and efficiency, as the frozen pretrained vision and language encoders retain their domain expertise, while the lightweight Q-Former handles cross-modal alignment with minimal trainable parameters.

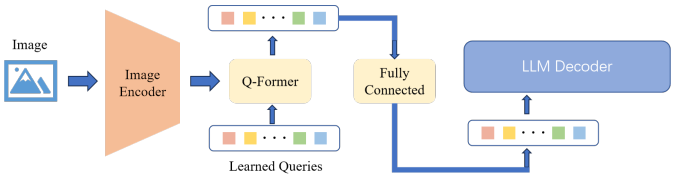

Fig. 2. An illustration of the BLIP-2 architecture, composed of a vision encoder, the Q-Former (a query-based Transformer), and a decoder-based LLM (e.g. OPT). A fully-connected layer bridges the Q-Former's output dimension to the LLM's input space, enabling efficient cross-modal alignment while keeping both the encoder and LLM frozen.

In contrast, linear projection methods offer a computationally frugal solution by mapping raw multimodal features directly to LLM spaces via fully connected layers. LLaVA [11] exemplifies this paradigm, employing a single linear MLP to transform CLIP-extracted image features into vectors that match the dimension of the LLM's text embeddings. Although less expressive than attention-based mechanisms, this method achieves remarkable efficiency and simplicity, enabling seamless integration of visual and textual tokens for multimodal prompts.

b) Expert-driven language conversion: Here, domain-specific expert models (e.g. image-to-text captioners) first convert non-text data into linguistic descriptions. The generated text is then fed directly to the LLM. For instance, PaLI [12] leverages an image captioning model to transform visual inputs into textual prompts that guide LLM reasoning. This method bypasses direct feature alignment but depends on the accuracy of intermediate linguistic representations.

III. TRAINING STRATEGY

Multimodal Large Language Models (MLLMs) generally undergo three primary training stages: pretraining, instruction tuning, and alignment tuning. Each stage addresses distinct objectives and employs different data types, enabling progressive optimization of cross-modal capabilities. This section delves into the goals, data requirements, and characteristics of each stage, followed by a discussion of their synergistic effects on model performance.

A. Pretraining

Pretraining serves as the foundational stage for MLLMs, aiming to align heterogeneous modalities (e.g. vision and language) and establish robust cross-modal representations. Typically, MLLMs integrate a pretrained language model (LLM) and a multimodal encoder (e.g. a vision encoder during image-text prtraining), both of which remain frozen during this phase. While the LLM excels at text understanding and reasoning, and the vision encoder captures rich visual features, aligning their latent spaces is critical for enabling joint multimodal understanding. The primary objective of pretraining is to optimize lightweight multimodal connectors that bridge these frozen modules, facilitating cross-modal interaction without compromising their pretrained knowledge.

a) Strategy: A widely adopted pretraining strategy involves freezing both the vision encoder and LLM while training only the connector. A common training objective is to minimize the cross-entropy loss, enabling the model to autoregressively generate accurate textual descriptions (e.g., image captions) conditioned on visual inputs. Recent studies, however, have explored partial unfreezing of the vision encoder to refine image-text alignment. For instance, the ShareGPT-4V model [13] demonstrated that selectively unfreezing the vision encoder while leveraging high-quality caption data significantly improves multimodal alignment, yielding more accurate and coherent text outputs.

b) Data: This stage primarily relies on large-scale image-text pairs, where natural language descriptions are associated with corresponding images, audio, or videos. Two types of data are commonly utilized: coarse-grained and fine-grained pairs. Coarse-grained data is typically scraped from the web, offering vast scale and diversity at the cost of noise. These datasets often suffer from mismatched or inaccurate image-text pairs, with captions being brief and imprecise. In contrast, fine-grained data features longer, detailed descriptions that enable precise alignment between modalities. Such data is often human-annotated or synthesized by powerful MLLMs (e.g. GPT-4V). For example, using high-quality, small-scale

MLLM-generated captions to train a compact caption generator, which is then scaled up to produce large, refined datasets. While fine-grained data enhances cross-modal understanding, its creation requires expensive commercial MLLMs, resulting in limited dataset sizes.

B. Instruction tuning

Instruction tuning refines MLLMs to interpret and execute diverse tasks specified via natural language instructions. This phase significantly enhances zero-shot generalization by exposing models to a broad spectrum of tasks during training.

a) Strategy: Instruction samples typically consist of a task instruction paired with multimodal input-output pairs. For instance, in Visual Question Answering (VQA) [14], an instruction may be formulated as "Answer the question: How many objects are present in the image?" The input can include images, text, or their combination, while the output is the task-specific response.

Formally, an instruction-based sample is represented as a triplet (I, X, Y), where:

- I: Task instruction (natural language description).
- X: Multimodal input (e.g., image, text, or their fusion).
- Y: Ground-truth response.

The MLLM predicts an answer $\hat{Y}$ based on I and X, parameterized by θ :

$$\hat{Y} = f_\theta(I, X) \tag{1}$$

where the training objective minimizes the autoregressive generation loss:

$$\mathcal{L} = -\sum_{t=1}^{T} \log P_\theta \left(y_t \mid y_{<t}, I, X \right) \tag{2}$$

Here, T denotes the response length, and y_t is the t-th token in Y. This objective maximizes the likelihood of generating the ground-truth tokens incrementally.

b) Data: Instruction datasets demand flexible formats and task diversity, posing challenges in data collection. Three primary approaches are adopted: data adaptation, self-instruction, and data blending.

- Data Adaptation: Existing datasets (e.g. VQA) are reformatted into instruction-answer pairs. Instructions are generated manually or semi-automatically via LLMs (e.g. GPT-4) using seed templates. For tasks with short answers, strategies include enforcing output length constraints or augmenting contextual information.
- Self-Instruction: LLMs synthesize new instructions to address specialized needs (e.g. multi-turn dialogue). For example, LLaVA converts images into textual descriptions and employs GPT-4 to generate multimodal instruction datasets. Similar approaches are adopted in MiniGPT-4 and ChatBridge [15]. Recent advancements leverage powerful MLLMs like GPT-4V to automate high-quality data generation.
- Data Blending: Combining multimodal and language-only data improves dialogue fluency and instruction-following. LaVIN [16] randomly samples from both data types during training.In contrast, MultiInstruct [17] explores hybrid strategies, such as mixed tuning, which involves random blending, and sequential tuning, where language data is followed by multimodal data.

C. Alignment tuning

Alignment tuning is critical for adapting MLLMs to human preferences, particularly in reducing hallucinations (e.g. generating unsupported visual claims). Two dominant approaches are Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO), which we detail below.

a) Reinforcement Learning from Human Feedback (RLHF): RLHF [18] aligns models with human preferences via three stages:

- Supervised Fine-tuning (SFT): Initializes a policy model using labeled data or pretrained instruction-tuned models. This stage is optional if the base model already exhibits instruction-following capabilities.
- Reward Modeling: Trains a reward model to score candidate responses based on human preference data. Given a multimodal input (e.g. image-text pair) and two responses, the model assigns higher rewards to favorable outputs.
- Reinforcement Learning: Optimizes the policy model via Proximal Policy Optimization (PPO), while constraining divergence from the initial policy using a KL divergence penalty [19].

b) Direct Preference Optimization (DPO): DPO [23] bypasses explicit reward modeling by directly learning from pairwise human preferences via binary classification loss. This approach reduces computational complexity and hallucination rates. Variants like RLHF-V [20] refine this process by collecting fine-grained (e.g. segment-level) preference data and optimizing with DPO.

Alignment tuning relies on high-quality pairwise comparison datasets where humans rank model responses. Due to high annotation costs, datasets are typically small but meticulously curated (Table 1). Common modalities include image (I) and text (T) inputs.

TABLE I
A SUMMARY OF DATASETS FOR ALIGNMENT-TUNING. FOR INPUT/OUTPUT MODALITIES

Dataset	Sample	Modality	Source
LLaVA-RLHF [21]	10K	$I + T \rightarrow T$	Human
RLHF-V [20]	5.7K	$I + T \rightarrow T$	Human
VLFeedback [22]	380K	$I + T \rightarrow T$	GPT-4V

IV. EVALUATION

Evaluation is a critical component of MLLM development, offering actionable feedback for optimization and enabling systematic comparison across models. Unlike traditional multimodal models, which are often task-specific, MLLMs are

inherently general-purpose, necessitating comprehensive evaluation frameworks to assess their broad capabilities. Additionally, MLLMs exhibit emerging abilities (e.g. OCR-free mathematical reasoning), demanding novel evaluation protocols. We categorize existing approaches into the following dimensions:

A. Task-Specific Benchmarks

These benchmarks focus on classical multimodal tasks:

- Visual Question Answering (VQA): Accuracy on datasets like VQA-v2 and OK-VQA [23] measures visual grounding and commonsense reasoning.
- Image Captioning: Metrics such as CIDEr [24] and BLEU-4 evaluate fluency and relevance of generated descriptions.
- Cross-Modal Retrieval: Precision@K on datasets like COCO quantifies alignment between modalities.

B. General Capability Assessment

To evaluate broader abilities inherent to MLLMs:

- Open-Ended Generation: Tools like GPT-4 as a judge score creativity and coherence in open scenarios (e.g. story generation from images).
- In-Context Learning: Performance on few-shot tasks using benchmarks like MMBench [25] tests adaptability to new instructions.
- Multimodal Dialogue: Metrics for multi-turn interaction (e.g. relevance, consistency) are evaluated using datasets such as VisDial.

C. Emerging Ability Evaluation

Novel protocols address unique MLLM capabilities:

- OCR-Free Reasoning: Math-focused benchmarks (e.g. MathVista, TabMWP [26]) assess symbolic and numerical reasoning without explicit text detection.
- Hallucination Suppression: The POPE benchmark quantifies hallucination rates in object existence verification tasks.

V. Multimodal Hallucination

Multimodal hallucinations refer to incongruences between generated text and visual input. These can include incorrect object recognition, attribute misdescription, or flawed relational reasoning. Such hallucinations severely degrade the performance and trustworthiness of Multimodal Large Language Models (MLLMs). Therefore, addressing these hallucinations is critical for enhancing model robustness and reliability. This can be achieved through three strategies: pre-rectification, process rectification, and post-rectification.

A. Pre-correction

Pre-correction mitigates hallucinations by optimizing training data and fine-tuning strategies. This involves augmenting standard datasets with adversarially crafted negative samples, such as image-text pairs with deliberate inconsistencies (e.g., mismatched object attributes), to strengthen the model's cross-modal alignment. Frameworks like LRV-Instruction [27] further enhance grounding through negative semantic instructions (e.g. "Avoid excessive reasoning or associations when describing images."), directing the model to prioritize factual fidelity. Similarly, LLaVA-RLHF [21] employs reinforcement learning from human feedback (RLHF) to align outputs with human preferences, reducing over-imaginative outputs in complex multimodal reasoning tasks. These data-driven approaches strengthen cross-modal fidelity by training models to prioritize visual evidence over unwarranted inferences.

B. In-process correction

In-process correction strategies alleviate hallucination generation by improving model architectures and inference processes. These methods identify the root causes of hallucinations and design corresponding adjustment mechanisms. For instance, HallE-Switch [28] analyzes the sources of object existence hallucinations and proposes that hallucination generation is closely related to the inherent knowledge reasoning in LLMs rather than solely the visual encoder's output. By introducing continuous control factors, it effectively constrains hallucination generation during inference, thereby reducing over-speculation. Additionally, VCD (Visual Contrastive Decoding) [29] suggests prioritizing reliance on visual information over potential inferences from the language model when image content undergoes noise processing, effectively mitigating hallucinations caused by linguistic interference. HACL [30] optimizes visual and linguistic embedding spaces by employing contrastive learning to enhance cross-modal representation similarity, enabling the model to better distinguish real content from hallucinated content.

C. Post-correction

Post-correction methods correct hallucinations in generated text through specific mechanisms after model inference. These methods typically integrate expert models to supplement image contextual information and iteratively adjust model-generated descriptions. Woodpecker [31] proposes a general hallucination correction framework that revises erroneous descriptions by stepwise examination of intermediate results at each reasoning stage. A key strength of this method lies in its interpretability, allowing researchers to clearly trace and understand the causes of hallucinations and perform corrections by augmenting image context. Similarly, LURE [32] trains specialized revisers to regenerate descriptions for high-uncertainty objects, ensuring the final output is more accurate and trustworthy.

VI. Challenges and Future Directions

As a cutting-edge technology in artificial intelligence, Multimodal Large Language Models (MLLMs) have demonstrated significant potential in various application scenarios. However, they still face a series of challenges that limit their widespread adoption and in-depth development in real-world applications.

A. Limitations in Long Context Processing

Current MLLMs exhibit notable difficulties in processing information with long temporal spans and extended contextual

dependencies. This is particularly evident in tasks such as video understanding or the analysis of lengthy documents interleaved with images and text, where models often underperform. Although recent advancements, including extended context windows and efficient memory mechanisms, have partially addressed these challenges, most models still struggle to handle the complex interactions of massive multimodal information. Therefore, designing models capable of efficiently processing long-span, multimodal interleaved information while balancing computational efficiency with performance remains a critical direction for future research.

B. Effectiveness of Modality Fusion

A core challenge in multimodal learning lies in effectively fusing information from heterogeneous modalities. While existing models can partially handle mappings between images and text, their performance degrades significantly when faced with more complex modality combinations (e.g. image-audio pairs or long videos containing diverse information). Different modalities exhibit distinct data structures and semantic representations, making it inherently difficult to design a unified representation that effectively captures and integrates multimodal signals. Future research may prioritize developing efficient cross-modal alignment mechanisms, attention architectures, and deep fusion methods to enhance the synergy of multimodal information.

C. Security and Defensive Mechanisms

Like generic Large Language Models (LLMs), MLLMs are vulnerable to security threats. Carefully designed adversarial attacks can mislead models to generate inappropriate, biased, or harmful content. Furthermore, since MLLMs integrate multiple data modalities, their attack surfaces are broader, potentially making them prime targets for malicious actors. Consequently, constructing models with enhanced robustness and security to prevent misuse or exploitation has become an urgent priority. Beyond improving security guarantees, future research must also ensure that models provide accurate and reliable judgments and responses in complex real-world tasks.

VII. CONCLUSION

This paper provides a comprehensive review of the development of Multimodal Large Language Models (MLLMs) and offers an in-depth analysis of current research advancements and challenges. By detailing the architectural frameworks, training strategies, data processing techniques, and evaluation standards of MLLMs, we establish a systematic foundation for understanding their capabilities and limitations. Furthermore, we identify critical shortcomings in current models when handling complex tasks, particularly multimodal hallucinations, and emphasize the urgent need to enhance model robustness and trustworthiness. We hope this survey will serve as a valuable reference for both academia and industry, fostering advancements and novel insights in MLLM research while inspiring broader interdisciplinary contributions to the field.

REFERENCES

[1] A. Radford, J. W. Kim, C. Hallacy et al., "Learning Transferable Visual Models From Natural Language Supervision," arXiv, 2021. [Online]. Available: arXiv:2103.00020

[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 29, pp. 1935–1948, 2021, doi: 10.1109/TASLP.2021.3096775.

[3] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP: Learning Audio Concepts from Natural Language Supervision," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Rhodes Island, Greece, 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10096102.

[4] G. Bertasius, H. Wang, and L. Torresani, "TimeSformer: Is Space-Time Attention All You Need for Video Understanding?," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Montreal, Canada, 2021, pp. 2278–2288, doi: 10.1109/ICCV48922.2021.00232.

[5] C. Tong, Z. Liu, H. Liu, and X. Wang, "VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training," arXiv, 2022, [Online]. Available: https://arxiv.org/abs/2209.12573.

[6] Meta AI, "Llama: Open and Efficient Foundation Language Models," arXiv, 2023, [Online]. Available: https://arxiv.org/abs/2305.12345.

[7] Tong Z , Song Y , Wang J et al., "VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training," arXiv, 2022. [Online]. Available: arXiv:2203.12602.

[8] J. Bai, S. Bai, Y. Chu et al., "Qwen technical report," arXiv, 2023. [Online]. Available: arXiv:2309.16609.

[9] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," arXiv, 2023. [Online]. Available: arXiv:2310.03744.

[10] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," arXiv, 2023. [Online]. Available: arXiv:2301.12597.

[11] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," arXiv, 2023. [Online]. Available: arXiv:2304.08485.

[12] X. Chen, X. Wang, S. Changpinyo et al. "PaLI: A Jointly-Scaled Multilingual Language-Image Model," arXiv, 2022. [Online]. Available: arXiv:2209.06794.

[13] L. Chen, J. Li, X. Dong et al., "Sharegpt4v: Improving large multi-modal models with better captions," arXiv, 2023. [Online]. Available: arXiv: 2311.12793.

[14] S. Antol, A. Agrawal, J. Lu et al., "VQA: Visual Question Answering," in 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 279-287, doi: 10.1109/ICCV.2015.279.

[15] Z. Zhao, L. Guo, T. Yue, S. Chen, S. Shao, X. Zhu, Z. Yuan, and J. Liu, "Chatbridge: Bridging modalities with large language model as a language catalyst," arXiv, 2023. [Online]. Available: arXiv:2305.16103.

[16] G. Luo, Y. Zhou, T. Ren, S. Chen, X. Sun, and R. Ji, "Cheap and quick: Efficient vision-language instruction tuning for large language models," arXiv, 2023. [Online]. Available: arXiv:2305.15023.

[17] Z. Xu, Y. Shen, and L. Huang, "Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning," arXiv, 2022. [Online]. Available: arXiv:2212.10773.

[18] D. M. Ziegler, N. Stiennon, J. Wu et al., "Fine-tuning language models from human preferences," arXiv, 2019. [Online]. Available: arXiv:1909.08593.

[19] L. Ouyang, J. Wu, X. Jiang et al., "Training language models to follow instructions with human feedback," arXiv, 2022. [Online]. Available: arXiv:2203.02155.

[20] T. Yu, Y. Yao, H. Zhang, T. He et al., "Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback," arXiv, 2023. [Online]. Available: arXiv:2312.00849.

[21] Z. Sun, S. Shen, S. Cao et al., "Aligning large multimodal models with factually augmented rlhf," arXiv, 2023. [Online]. Available: arXiv:2309.14525.

[22] L. Li, Z. Xie, M. Li et al., "Silkie: Preference distillation for large visual language models," arXiv, 2023. [Online]. Available: arXiv:2312.10665.

[23] K. Marino, M. Rastegari, A. Farhadi et al., "OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge," in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 5541-5550, doi: 10.1109/CVPR.2019.00331.

[24] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based Image Description Evaluation," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 4567-4575, doi: 10.1109/CVPR.2015.7299087.

[25] Y. Liu, H. Duan, Y. Zhang, et al., "MMBench: Is Your Multi-modal Model an All-Around Player?," arXiv, 2023. [Online]. Available: arXiv:2307.06281.

[26] P. Lu, L. Qiu, K.-W. Chang, Y. N. Wu, S. Zhu, T. Rajpurohit, P. Clark, and A. Kalyan, "Dynamic Prompt Learning via Policy Gradient for Semi-Structured Mathematical Reasoning," arXiv preprint arXiv:2209.14610, Sep. 2022. [Online]. Available: https://arxiv.org/abs/2209.14610.

[27] F. Liu, K. Lin, L. Li, J. Wang, Y. Yacoob, and L. Wang, "Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning," in Proc. 12th Int. Conf. Learn. Represent. (ICLR), Vienna, Austria, 2024, [Online].

[28] B. Zhai, S. Yang, X. Zhao et al., "Halle-switch: Rethinking and controlling object existence hallucinations in large vision language models for detailed caption," arXiv, 2023. [Online]. Available: arXiv:2310.01779.

[29] S. Leng, H. Zhang, G. Chen, et al., "Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding," in 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2024, pp. 1316–1325, doi: 10.1109/CVPR52733.2024.01316.

[30] C. Jiang, H. Xu, M. Dong et al., "Hallucination augmented contrastive learning for multimodal large language model," arXiv, 2023. [Online]. Available: arXiv:2312.06968.

[31] S. Yin, C. Fu, S. Zhao et al., "Woodpecker: Hallucination correction for multimodal large language models," arXiv, 2023. [Online]. Available: arXiv:2310.16045.

[32] Y. Zhou, C. Cui, J. Yoon et al., "Analyzing and mitigating object hallucination in large vision-language models," arXiv, 2023. [Online]. Available: arXiv:2310.00754.