Computers and Artificial Intelligence；Vol.3, No.3; 2026
E-ISSN: 3006 -3728 P-ISSN：3006-371X
DOI: https://doi.org/10.70267/cai.26v3n3.167176
Published by: Zeus Press

# A Review of Multimodal Large Language Models: Fusion Mechanisms and Capability Evolution

**Zibo Xu and Shuqiang Gao***

*School of Computer Science and Artificial Intelligence, Zhengzhou University, Zhengzhou 450001, China*

**Corresponding author: Shuqiang Gao.*

## Abstract

With the continuous advancement of large language models, MLLMs have gradually emerged as a major research focus in the field of artificial intelligence. Although existing studies have reviewed relevant progress from different perspectives, such as multimodal learning, vision-language models, and domain-specific applications, there is still a lack of systematic analysis of the intrinsic relationships among fusion mechanisms, capability formation, and application expansion in MLLMs. This paper reviews the principal technical paradigms of multimodal fusion, including early fusion, intermediate fusion, late fusion, and hybrid fusion, and compares the structural characteristics and applicable scenarios of different fusion approaches. Building on this analysis, the paper further explores the development of MLLMs from the perspectives of vision-language understanding, multimodal content generation, multimodal interaction and agent-oriented tasks, as well as domain-specific applications, thereby outlining the evolutionary trajectory of their core capabilities and broader application trends.

### Keywords

multimodal large language models, fusion mechanisms, multimodal understanding, multimodal generation

## 1. Introduction

The rapid development of large language models (LLMs) has significantly transformed artificial intelligence research, leading to notable improvements in language understanding, knowledge-based question answering, reasoning, and content generation. However, information in real-world scenarios is inherently multimodal, often involving images, text, speech, video, and sensor data. A purely text-based approach struggles to meet the demands of processing complex, multi-source, and heterogeneous data.

Against this background, MLLMs have gradually become an important research direction in artificial intelligence. By incorporating visual, auditory, video, and other modalities, MLLMs enable richer perception, understanding, reasoning, and generation within a unified framework [1, 2]. With advances in visual encoders, large language models, and cross-modal alignment methods, multimodal systems have evolved from simple feature concatenation and shallow alignment into more sophisticated architectures capable of unified representation learning and deeper cross-modal interaction. These models have shown broad application potential in tasks such as intelligent question answering, content generation, human–computer interaction, and domain-specific applications.

Nevertheless, MLLMs still face several critical challenges. First, substantial differences among modalities in data distribution, representational granularity, and semantic structure make efficient and stable cross-modal fusion a major challenge in model design. Second, issues such as insufficient cross-modal consistency, factual inaccuracies, and hallucinations persist in understanding and generation tasks, constraining deployment in high-stakes settings. Third, as multimodal models are increasingly deployed in real-world settings, improving their reliability in interactive tasks, tool use, and domain-specific applications has become an important issue.

Prior surveys have examined progress through the lenses of multimodal machine learning, vision–language models, and domain-specific applications. However, a systematic analysis of the interplay among fusion mechanisms, capability development, and application expansion in MLLMs remains relatively limited. Given the rapid growth of new models, tasks, and application scenarios, a more systematic review is needed to clarify the design principles and future development directions of modern MLLMs.

This survey reviews the key technologies and capability development of multimodal large language models. Its main contributions are as follows:

(1) We review major fusion paradigms—including early, mid-level, late, and hybrid fusion—analyzing the structural characteristics, advantages, and limitations of each approach.

(2) We summarize the core capabilities and emerging trends of MLLMs across vision–language understanding, multimodal content generation, multimodal interaction and agentic tasks, and vertical-domain applications.

(3) Building on the above, we outline overall development trends and future directions.

The remainder of the paper is organized as follows: Section 2 examines the main classifications and evolutionary paths of multimodal fusion techniques; Section 3 discusses the capabilities and application expansion of MLLMs in understanding, generation, interaction, and vertical domains; and the final section summarizes the survey and offers prospects for future research.

## 2. Classification of Multimodal Fusion Techniques

### 2.1 Early Fusion

Early fusion combines features or initial representations from different modalities at the early stages of the model architecture to form a unified representation. This fusion process usually takes place at the input level or within shallow feature layers before deeper semantic modeling is performed.

In traditional multimodal learning, early fusion primarily takes the form of input-level or feature-level integration: initial representations or shallow features are extracted from each modality, then concatenated or transformed into a unified input that is processed jointly by the downstream model. Early audio–video joint learning already reflected the idea of integrating multimodal information through shared representations [1]. Subsequent work on language, vision, and acoustic modalities further formalized this as a canonical feature-level fusion strategy [2].

With the development of vision–language pre-training, early fusion gradually shifted from simple feature concatenation to joint encoding within shared backbone architectures. For instance, VisualBERT feeds image region features and text tokens together into a Transformer, allowing self-attention to capture correspondences between vision and language [3]. ViLT further reduces reliance on a dedicated visual encoder by passing both visual patches and text tokens directly to the same Transformer, exhibiting a stronger end-to-end modeling character [4].

In the era of multimodal large models, early fusion has extended to token-level unified modeling. Chameleon represents both images and text as discrete tokens and processes interleaved multimodal sequences within a single Transformer [5]. Emu3 unifies text, images, and video in the same autoregressive framework, performing joint modeling via next-token prediction [6]. Although these approaches differ from traditional feature concatenation methods, they still map different modalities into a shared sequence space before processing by the backbone network and can therefore be regarded as an extension of early fusion in large-model architectures.

A major advantage of early fusion is that it enables different modalities to interact at an early stage, which helps the model learn cross-modal correspondences and complementary information. However, differences in data distribution, representational granularity, and information density across modalities are introduced into the same modeling process at an early stage, often leading to alignment difficulties and noise interference. Its development has evolved from traditional feature-level combination to shared-backbone joint encoding and, more recently, to end-to-end sequence modeling in a unified token space.

## 2.2 Mid-Level Fusion

Mid-level fusion performs cross-modal interaction and joint modeling at intermediate representation layers after each modality has undergone initial encoding. Unlike early fusion, mid-level fusion first learns modality-specific representations and then performs cross-modal interaction through mechanisms such as cross-attention, query-based bridging, or adapters.

Implementation generally follows the pipeline of "unimodal encoding → intermediate interaction → joint representation output." Visual, linguistic, and other modalities are first encoded by their respective encoders to produce relatively stable unimodal representations. Cross-modal interaction modules are then introduced at intermediate layers to align, filter, and fuse information from different modalities. The fused representations are then fed into downstream decoders or task-specific modules for prediction or generation.

Flamingo [7] is a representative example of mid-level fusion. It combines a frozen visual encoder with a frozen language model and inserts cross-modal interaction modules into the language backbone. A Perceiver Resampler compresses visual features into a fixed number of visual tokens, which are then integrated through GATED XATTN-DENSE blocks inserted between language model layers. This design allows visual information to be incorporated while largely preserving the original generative ability of the language model, and it supports interleaved image, video, and text inputs.

BLIP-2 [8] exemplifies a bridging-module-centric approach. It introduces a lightweight Q-Former between a frozen image encoder and a frozen LLM. Learnable queries extract vision-relevant representations from visual features and project them into a form suitable for the language model. Unlike Flamingo, which performs continuous layer-wise interaction, BLIP-2 focuses more on compressing, selecting, and aligning visual information through an independent bridging module.

From an evolutionary perspective, Frozen first demonstrated the insertion of a bridging mechanism between pretrained unimodal models by treating image representations as soft prompts for a frozen language model [9]. Building on this, Flamingo extended visual injection to continuous interaction within language model layers, while BLIP-2 improved stability and efficiency via a dedicated bridging module. As a result, mid-level fusion has evolved from simple visual information injection to more stable and controllable forms of cross-modal interaction.

Mid-level fusion retains strong unimodal representation ability while allowing more flexible cross-modal integration. It shows strong adaptability for complex semantic alignment and deep collaborative tasks. However, the additional interaction or bridging modules also increase architectural complexity, training cost, and computational overhead, and model performance often depends heavily on the effectiveness of these intermediate components.

## 2.3 Late Fusion

Late fusion combines the outputs of independently modeled modalities at the final prediction or decision stage. A key characteristic of late fusion is that each modality is processed separately for representation learning or task prediction, while cross-modal information is merged only at the final stage to produce a unified result.

Rather than representing a specific model architecture, late fusion is better understood as a decision-level integration strategy. Early work on Stacked Generalization (Stacking) [10] offered a representative learning-based integration strategy: multiple base learners generate predictions, which are then synthesized by a meta-learner. In multimodal tasks, different modality branches function as independent learners, while the fusion module combines their outputs at the decision level. In the deep learning era, late fusion typically involves

separate branches for representation learning or task prediction per modality, followed by weighted averaging, voting, or a learned fusion module at the output stage [2, 11].

Late fusion provides strong modularity and is highly compatible with heterogeneous data and different unimodal models, making it relatively easy to extend existing unimodal systems into multimodal frameworks. Because it relies less on precise intermediate cross-modal alignment, late fusion remains effective in scenarios involving missing modalities, uneven modality quality, or limited paired multimodal data. Still, since interaction between modalities occurs only at later stages, late fusion often struggles to capture fine-grained cross-modal relationships, which limits its performance on tasks requiring deep semantic alignment. As a result, late fusion is more suitable for scenarios with strong modality heterogeneity, limited paired data, or a need for independent modality modules. Its development has evolved from traditional decision-level combination methods to output-stage integration within deep learning architectures.

## 2.4 Hybrid Fusion

Hybrid fusion performs multimodal processing and integration across multiple stages, emphasizing information interaction across different layers and stages. These approaches typically incorporate characteristics of early, mid-level, and late fusion, enabling multimodal information to participate jointly in representation learning, interaction modeling, and decision-making.

Hybrid fusion does not refer to a single fixed architecture but instead encompasses various multi-stage and multi-level fusion designs. Depending on fusion locations, interaction mechanisms, and joint modeling paths, implementations may combine feature-level and decision-level integration, hierarchical interaction structures, representation space decomposition, or joint optimization of learning objectives. The main emphasis is on how different fusion mechanisms are coordinated within a unified framework rather than on any single module itself.

Early hybrid approaches often combined feature-level integration with decision-level fusion [11]. These methods typically first normalize and combine visual, textual, and other features at the early stage, and then use downstream classifiers to model higher-level semantic relationships, thereby integrating information at both feature and decision levels.

Among multi-stage methods, the Hierarchical Feature Fusion Network (HFFN) [12] is a notable example. Adopting a “divide, conquer, and combine” strategy, it first models cross-modal relationships within local windows and then performs global integration of the local results. Local fusion modules capture fine-grained correspondences, while global modules model dependencies among local states, with two-level attention highlighting important local and global information. Compared with single-stage methods, this progressive “local-to-global” design more clearly reflects the multi-level integration characteristics of hybrid fusion.

In addition to hierarchical interaction structures, hybrid fusion can also be reflected in the joint design of representation spaces and learning objectives. The Multimodal Factorization Model (MFM) [13] decomposes multimodal representations into cross-modal shared discriminative factors (used for label prediction) and modality-specific generative factors (used to preserve individual modality information and support reconstruction). By jointly optimizing generative and discriminative objectives, the model can learn cross-modal relationships while improving robustness to missing or noisy modalities.

Hybrid fusion enables multimodal integration at different levels, where lower stages capture fine-grained correspondences and higher stages combine information from multiple sources to support downstream tasks. Compared with single-stage fusion methods, hybrid fusion is generally more suitable for tasks involving complex semantic relationships and multi-step reasoning. However, multi-stage structures increase model complexity, computational cost, and implementation overhead. Noise or alignment errors introduced in early stages can also propagate along the fusion pipeline. For more complex designs involving hierarchical interaction or joint objective optimization, it is necessary to balance model performance, computational efficiency, and system complexity.

## 2.5 Summary

This section reviewed the major paradigms of early, mid-level, late, and hybrid fusion. These approaches mainly differ in terms of fusion stage, modality interaction mechanisms, and overall system organization: early

fusion emphasizes front-end joint representation, mid-level fusion focuses on intermediate cross-modal interaction, late fusion highlights output- or decision-level integration, and hybrid fusion achieves multi-stage synergy that balances low-level complementarity with high-level decision optimization. Each fusion paradigm reflects different trade-offs among representation capability, system flexibility, robustness, and computational cost.

From a developmental perspective, multimodal fusion techniques have gradually evolved from simple feature combination and decision-level integration toward unified representation learning, deeper cross-modal interaction, and multi-stage collaborative modeling. In multimodal large models, fusion mechanisms increasingly emphasize efficient alignment and coordinated modeling of different modalities within unified frameworks, reflecting a transition from simple information integration to more collaborative forms of multimodal modeling.

## 3. Core Capabilities and Applications of Multimodal Large Language Models

### 3.1 Multimodal Understanding and Generation Capabilities

Multimodal understanding and generation are among the core capabilities of MLLMs. Current MLLMs have gradually formed a capability framework centered on multimodal understanding and content generation. This section discusses these capabilities from the perspectives of vision–language understanding and multimodal content generation.

#### 3.1.1 Vision–language Understanding Tasks

Vision–language understanding tasks differ from traditional methods that focus only on image classification or text understanding. These tasks require models to align visual and linguistic information within a shared semantic space in order to support more complex cross-modal reasoning. Existing studies suggest that current vision–language understanding capabilities have expanded to tasks such as visual question answering, multimodal dialogue, OCR, and knowledge reasoning.

The main types of vision–language understanding tasks are summarized in Table 1.

*Table 1: Major types of vision–language understanding tasks*

| Task Type | Typical Tasks | Core Capability Requirements |
|---|---|---|
| General image-text understanding | Visual question answering, image captioning, multimodal dialogue | Basic semantic alignment between visual content and language instructions |
| Scene text and document understanding | OCR, scene text VQA, document VQA | Integrated recognition of image text, layout structure, and semantic relationships |
| Structured visual understanding and complex reasoning | Chart understanding, multi-image relation modeling, cross-modal evidence integration | Modeling of structured information and complex cross-modal semantic relationships |
| Video understanding | Video question answering, event understanding, temporal reasoning | Understanding of action dynamics, event continuity, and temporal semantic relationships |

In general image–text understanding, studies such as MiniCPM-V [14] suggest that vision–language understanding has gradually evolved from single-task performance to a more comprehensive capability framework covering multiple scenarios and dimensions. In scene text and document understanding, TextVQA [15] and DocVQA [16] extend model capabilities to the joint understanding of textual content, layout structures, and semantic information. For structured visual understanding and complex reasoning, ChartLLaMA [17] and MMICL [18] advance chart and multi-image comprehension. Video-MME [19] further expands vision–language understanding to video scenarios, with particular emphasis on temporal relationships and event dynamics. Representative models and their major capability characteristics are summarized in Table 2. Overall, vision–language understanding tasks are gradually extending from natural-image scenarios to more complex settings involving dense text, structured information, and dynamic content.

*Table 2: Representative studies on vision–language understanding tasks and their capability features*

| Representative Study | Primary Task Focus | Core Capabilities and Task Characteristics | Development Trend |
|---|---|---|---|
| MiniCPM-V[14] | General image-text understanding | Comprehensive evaluation of VQA, multimodal dialogue, OCR, document understanding, and knowledge reasoning | From single-task to multi-scenario, multi-dimensional capabilities |
| TextVQA[15] | Scene text understanding | Joint reasoning using visual context, image text, and question semantics | From image recognition to text reading and joint reasoning |
| DocVQA[16] | Document understanding | Integrated modeling of document text, layout, and semantic relationships | Text-dense visual content becoming a key research direction |
| ChartLLaMA[17] | Structured visual understanding | Chart understanding, semantic parsing, and related generation tasks | Expansion to structured visual content such as charts |
| MMICL[18] | Complex relation reasoning | Modeling cross-image relationships and contextual dependencies with multiple images | From single-image recognition to multi-image relational reasoning |
| Video-MME[19] | Video understanding | Assessment of action dynamics, event processes, and temporal reasoning | From static image analysis to dynamic video understanding |

### 3.1.2 Multimodal Content Generation

Whereas vision–language understanding mainly focuses on accurately interpreting multimodal inputs, multimodal content generation places greater emphasis on semantic consistency, content coherence, and controllability in generated outputs. Generation tasks in MLLMs have expanded from traditional text generation to images, speech, music, and other modalities, gradually evolving from one-way generation to more unified cross-modal and any-to-any generation frameworks [20]. Typical tasks are shown in Table 3.

*Table 3: Major types of multimodal content generation tasks*

| Task Type | Typical Tasks | Core Capability Requirements |
|---|---|---|
| Unidirectional cross-modal generation | Vision-to-text, text-to-image, text-to-speech, text-to-music | Modality conversion and content generation in specific directions |
| Unified multimodal generation | Multimodal input-output combinations, arbitrary modality generation | Unified modeling of multiple modalities within a shared framework |
| Context-driven and interactive generation | Multi-turn generation, interleaved input generation, editable generation | Maintaining coherence, controllability, and editability in complex contexts |
| Integrated understanding-generation and structured generation | Chart generation, visual generation, understanding-driven generation | Joint modeling of understanding and generation, extending to structured content |

In unidirectional cross-modal generation, research has progressed from traditional vision-to-text description to text-to-image, speech, music, and structured content generation. ChartLLaMA [17], for example, demonstrates structured generation capabilities in chart description, creation, and modification. In unified multimodal generation, NExT-GPT [21], AnyGPT [20], and Unified-IO 2 [22] adopt multimodal adapters, discrete token representations, and shared semantic spaces, respectively, to support unified input–output modeling. In context-driven and interactive generation, CoDi-2 [23] highlights interleaved inputs, multi-turn interaction, and editable outputs. Studies such as MetaMorph [24] further suggest a trend toward integrating understanding and generation tasks within unified multimodal architectures. Representative models and their capability features are summarized in Table 4. Research on multimodal content generation has gradually shifted from simply expanding output modalities toward developing unified generation frameworks, interactive generation processes, and more integrated model capabilities.

*Table 4: Representative studies on multimodal content generation and their capability features*

| Representative Study | Primary Task Focus | Core Capabilities and Task Characteristics | Development Trend |
|---|---|---|---|
| ChartLLaMA[17] | Structured content generation | Chart generation, modification, and description tasks | Generation targets extended to structured content |
| AnyGPT[20] | Unified multimodal generation | Unifies images, speech, music, and text as discrete token sequences for autoregressive generation | Unified representation as a key technical path |
| NExT-GPT[21] | Unified multimodal generation | Connects LLMs, multimodal adapters, and diffusion decoders for input-output mapping | From unidirectional conversion to unified generation |
| Unified-IO 2[22] | Unified multimodal generation | Unified modeling of images, text, audio, actions, and bounding boxes in a shared semantic space | From modular combination to integrated architecture |
| CoDi-2[23] | Context-driven and interactive generation | Supports interleaved inputs, multi-turn interaction, editable and composable outputs | From static generation to interactive generation |
| MetaMorph[24] | Understanding-generation integration | Simultaneous support for multimodal understanding and visual generation in a unified autoregressive framework | Understanding and generation moving toward collaborative modeling |

## 3.2 Multimodal Interaction and Vertical Applications

As MLLMs continue to improve, their application paradigms are gradually shifting from static task processing toward continuous interaction and adaptation to complex real-world environments. Multimodal interaction emphasizes task execution in continuous contexts, while vertical-domain applications focus on adaptation to domain-specific knowledge, workflows, and high-stakes constraints. Together, these directions reflect the growing movement of MLLMs toward practical real-world deployment.

### 3.2.1 Multimodal Interaction and Agentic Tasks

Multimodal interaction and agentic tasks focus on the ability of models to execute tasks in continuous interactive settings involving complex contexts and external tool integration. These tasks require models to integrate multimodal inputs, historical context, and external feedback while supporting functions such as context maintenance, task planning, tool invocation, and collaborative execution.

In continuous interaction, Visual ChatGPT [25] combines large language models with multiple vision foundation models, enabling multi-turn image-based dialogue and supporting complex visual tasks that require cooperation among different models. In tool calling and task execution, LLaVA-Plus [26] builds a skill library that connects pretrained vision and vision–language tools, allowing the model to select and combine different tools according to task requirements. MM-REACT [27] and ViperGPT [28] represent two different approaches: visual expert invocation and program-based task execution, respectively: the former supports multimodal reasoning and action through expert modules, while the latter transforms complex visual tasks into executable subroutine workflows via code generation. In more advanced collaborative scenarios, multimodal interaction further extends to multi-agent systems, where the focus shifts from improving a single model to mechanisms such as task allocation, context sharing, memory management, and collaboration among agents [29]. Representative systems are summarized in Table 5.

*Table 5: Representative systems for multimodal interaction and agentic tasks*

| Model Name | Main Characteristics |
|---|---|
| Visual ChatGPT[25] | Combines LLMs with multiple vision foundation models to support multi-turn interaction around images and complex visual tasks requiring model collaboration. |
| LLaVA-Plus[26] | Builds a skill library to access pretrained vision and vision–language tools, enabling dynamic selection, invocation, and combination of tools for complex tasks. |
| MM-REACT[27] | Integrates LLMs with visual expert modules, emphasizing reasoning and action through expert calling in multimodal tasks. |
| ViperGPT[28] | Organizes vision and language modules via code generation and program execution, converting complex visual tasks into executable subroutine flows. |

Multimodal interaction and agentic tasks reflect a further evolution of MLLM applications, moving from single-turn perception and response toward sustained interaction and more complex task execution. This trend is driving the transition from static understanding systems to more executable, scalable, and collaborative agent-based systems.

### 3.2.2 Vertical Domain Applications

As model capabilities continue to improve, vertical-domain applications have become an important development direction for MLLMs. These scenarios usually involve specialized knowledge boundaries, professional standards, and high-stakes requirements, placing stricter demands on model reliability, interpretability, and safety. Models are required not only to perform multimodal understanding and generation but also to adapt to domain-specific knowledge, professional terminology, workflow processes, and accountability requirements. Therefore, vertical applications highlight both the practical potential of MLLMs and the challenges they face in terms of data quality, trustworthiness, and safety [30].

Medicine and autonomous driving are representative vertical applications. Both scenarios involve heterogeneous inputs from multiple sources, strong domain knowledge constraints, and high requirements for reliability. In healthcare, models can integrate electronic health records, medical images, and laboratory indicators to support clinical decision assistance and report generation, yet they remain constrained by privacy protection, model bias, and hallucination risks [31,32]. In autonomous driving, models integrate visual perception, map information, sensor signals, and language instructions to support scene understanding, planning, decision-making, and human–vehicle interaction; however, system reliability, safety, and regulatory compliance remain critical deployment challenges [33]. Vertical applications are not mere direct transfers of general capabilities but require targeted adaptation using domain-specific data, task workflows, evaluation standards, and safety protocols to improve factual consistency, interpretability, and overall system reliability in complex real-world environments.

## 4. Conclusion and Outlook

This paper systematically reviews multimodal large language models, with a particular focus on their fusion mechanisms and capability development. It reviews major technical paradigms, including early, mid-level, late, and hybrid fusion, and summarizes the core capabilities and application development of MLLMs in vision–language understanding, multimodal content generation, multimodal interaction and agentic tasks, and vertical-domain applications. Despite significant progress, MLLMs still face a number of challenges, including cross-modal alignment, stable fusion, factual consistency, hallucination mitigation, and overall trustworthiness. Future research should further improve fusion mechanisms, model efficiency, and adaptation to domain-specific scenarios. At the same time, while advancing multimodal collaborative modeling, greater attention should also be paid to improving reliability in complex real-world applications.

## References

[1] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning* (pp. 689–696).

[2] Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423–443.

[3] Li, L. H., Yatskar, M., Yin, D., et al. (2019). VisualBERT: A simple and performant baseline for vision and language. arXiv. https://arxiv.org/abs/1908.03557

[4] Kim, W., Son, B., & Kim, I. (2021). ViLT: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 5583–5594).

[5] Chameleon Team. (2024). Chameleon: Mixed-modal early-fusion foundation models. arXiv. https://arxiv.org/abs/2405.09818

[6] Wang, X., Zhang, X., Luo, Z., et al. (2024). Emu3: Next-token prediction is all you need. arXiv. https://arxiv.org/abs/2409.18869

[7] Alayrac, J. B., Donahue, J., Luc, P., et al. (2022). Flamingo: A visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35, 23716–23736.

[8] Li, J., Li, D., Savarese, S., et al. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning* (pp. 19730–19742).

[9] Tsimpoukelli, M., Menick, J. L., Cabi, S., et al. (2021). Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34, 200–212.

[10] Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2), 241–259.

[11] Atrey, P. K., Hossain, M. A., El Saddik, A., et al. (2010). Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems*, 16(6), 345–379.

[12] Mai, S., Hu, H., & Xing, S. (2019). Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 481–492).

[13] Tsai, Y. H. H., Liang, P. P., Zadeh, A., Morency, L.-P., & Salakhutdinov, R. (2019). Learning factorized multimodal representations. In International Conference on Learning Representations. https://openreview.net/forum?id=rygqqsA9KX

[14] Yao, Y., Yu, T., Zhang, A., et al. (2024). MiniCPM-V: A GPT-4V level MLLM on your phone. arXiv. https://arxiv.org/abs/2408.01800

[15] Singh, A., Natarajan, V., Shah, M., et al. (2019). Towards VQA models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8317–8326).

[16] Mathew, M., Karatzas, D., & Jawahar, C. V. (2021). DocVQA: A dataset for VQA on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 2200–2209).

[17] Han, Y., Zhang, C., Chen, X., et al. (2023). ChartLLaMA: A multimodal LLM for chart understanding and generation. arXiv. https://arxiv.org/abs/2311.16483

[18] Zhao, H., Cai, Z., Si, S., et al. (2023). MMICL: Empowering vision-language model with multi-modal in-context learning. arXiv. https://arxiv.org/abs/2309.07915

[19] Fu, C., Dai, Y., Luo, Y., et al. (2025). Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 24108–24118).

[20]Zhan, J., Dai, J., Ye, J., et al. (2024). AnyGPT: Unified multimodal LLM with discrete sequence modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (pp. 9637–9662).

[21] Wu, S., Fei, H., Qu, L., et al. (2024). NExT-GPT: Any-to-any multimodal LLM. In *Proceedings of the 41st International Conference on Machine Learning* (pp. 53366–53397).

[22] Lu, J., Clark, C., Lee, S., et al. (2024). Unified-IO 2: Scaling autoregressive multimodal models with vision, language, audio, and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 26439–26455).

[23] Tang, Z., Yang, Z., Khademi, M., et al. (2024). CoDi-2: In-context, interleaved, and interactive any-to-any generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 27425–27434).

[24] Tong, S., Fan, D., Zhu, J., et al. (2025). MetaMorph: Multimodal understanding and generation via instruction tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 17001–17012).

[25] Wu, C., Yin, S., Qi, W., et al. (2023). Visual ChatGPT: Talking, drawing and editing with visual foundation models. arXiv. https://arxiv.org/abs/2303.04671

[26] Liu, S., Cheng, H., Liu, H., et al. (2024). LLaVA-Plus: Learning to use tools for creating multimodal agents. In *Proceedings of the European Conference on Computer Vision* (pp. 126–142).

[27] Yang, Z., Li, L., Wang, J., et al. (2023). MM-REACT: Prompting ChatGPT for multimodal reasoning and action. arXiv. https://arxiv.org/abs/2303.11381

[28] Surís, D., Menon, S., & Vondrick, C. (2023). ViperGPT: Visual inference via Python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 11888–11898).

[29] Han, S., Zhang, Q., Yao, Y., et al. (2024). LLM multi-agent systems: Challenges and open problems. arXiv. https://arxiv.org/abs/2402.03578

[30] Meskó, B. (2023). The impact of multimodal large language models on health care's future. *Journal of Medical Internet Research*, 25, e52865.

[31] Yi, Z., Xiao, T., & Albert, M. V. (2025). A survey on multimodal large language models in radiology for report generation and visual question answering. *Information*, 16(2), 136.

[32] Huang, H., Zheng, O., Wang, D., et al. (2023). ChatGPT for shaping the future of dentistry: The potential of multi-modal large language model. *International Journal of Oral Science*, 15(1), 29.

[33] Cui, C., Ma, Y., Cao, X., et al. (2024). A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops* (pp. 958–979).

## Funding

This research received no external funding.

## Conflicts of Interest

The authors declare no conflict of interest.

## Acknowledgment

This paper is an output of the science project.

## Copyrights