The Development and Application of Multimodal Large Models

Kejin Chen *

School of Electrical and Electronic Engineering, Nanyang Technological University, 639798, Singapore

* Corresponding Author Email: KEJIN001@e.ntu.edu.sg

Abstract. This paper aims to systematically review recent research on mainstream multimodal large-scale model structures, training strategies, multimodal fusion architectures, and practical applications. It focuses on the current state and future trends of multimodal large language models (MLLMs). By analyzing the structures of typical models such as CLIP, BLIP-2, PaLM-E, and GPT-4V, this paper summarizes their practical applications, common cross-modal alignment methods in multimodal large models, pre-training strategies, and standard task evaluation methods used during the modeling process. At the application level, we conducted a detailed study of multi-modal large language models (MLLMs) regarding typical use cases and performance metrics in key areas such as image-based content creation, cross-modal retrieval, visual question-answering, and multi-modal dialogue. The analysis found that MLLMs still encounter significant challenges in addressing modal hallucinations, maintaining semantic consistency, improving reasoning capabilities, and lowering training costs. This paper also summarizes the current research issues and emphasizes that future multimodal models should aim for greater generalization, modularity, controllability, and low-resource adaptability. Finally, building on existing research, the paper suggests several promising directions for further exploration, including multimodal context learning (M-ICL), visual-language chain reasoning, cross-domain knowledge transfer, and the miniaturization and deployment optimization of multimodal large models.

Keywords: Multimodal Large Language Model, Modality Fusion, Image-Text Generation, Embodied Intelligence.

1. Introduction

Language processing performance continues to advance with ongoing breakthroughs in artificial intelligence technology. However, human cognition relies not only on language patterns but also on integrating various information sources, such as images, sounds, and actions. To help AI systems better mimic human comprehensive perception, researchers have started expanding language models into multimodal models, which is fueling the rapid growth of multimodal large language models (MLMs).

The core feature of multimodal large models is their ability to integrate "perception-understanding-generation' capabilities. This allows them to "describe images," "recognize meaning from sound," or "think about images," demonstrating strong knowledge transfer and generation skills in cross-modal tasks. Currently, this core feature is widely applied in tasks such as text-image generation, cross-modal retrieval, image question-answering, speech interaction, and robot operation control, showing broad application prospects and widespread adoption across multiple fields.

Despite rapid advances in recent years, the future of multimodal large models still faces many challenges. For example, semantic alignment is more difficult in multimodal tasks than in unimodal ones; it is hard to obtain high-quality multimodal data; the model's ability to generalize needs further improvement; multimodal hallucination phenomena threaten reliability; and the extensive resources needed for training and deployment limit their use in edge devices and certain industries.

This paper systematically reviews and compares current mainstream MLLMs, summarizes the key technical approaches of existing achievements, and analyzes application outcomes, aiming to provide a reference for the future direction of this field.

2. Technical background of multimodal large models

As large language models (LLMs) become more widely used in natural language processing, researchers are expanding their architecture to include multimodal scenarios to improve models' overall perception and understanding of the world. Multimodal large language models (MLMs) have been developed, capable of processing different types of information such as language, images, speech, and video consistently, enabling natural cross-modal interaction.

This trend has already emerged with OpenAI's CLIP model. Using contrastive learning of text and images, CLIP links visual data to the linguistic semantic space in zero-shot text-image retrieval and classification, delivering impressive results. The BLIP and BLIP-2 models have further advanced in text-image question-answering and generation tasks, enhancing image understanding through techniques like visually guided question reconstruction and Q-Former [1].

PaLM-E is the first "specific intelligence model" that combines multimodal perception abilities with large language models, enabling efficient transfer from images and status information to natural language instructions by freezing the LLM backbone [2]. GPT-4V embeds multimodal input capabilities into a unified Transformer architecture [3], becoming a representative system for general text and image understanding.

Chen highlighted that crucial technologies in MLMs include modal alignment, pre-training task design, input fusion modes, and decoding strategies [4]. The ongoing development of these technologies has led to substantial growth in model size and capabilities, while also requiring higher data quality, more computing resources, and better model control.

3. Structure and classification of multimodal large models

3.1. Input fusion (Early Fusion)

In the input-level fusion mode, before passing data through the backbone network, information from various modalities is encoded into a single sequence of tokens, which are then input into a language model like Transformer for processing. Models such as BLIP and ALBEF use simple structures and methods suitable for lightweight understanding tasks, serializing image features and feeding them along with other inputs text [5]. The advantage is that it is easy to implement and has low computational costs, but because of the loss of modal specificity, it can easily lead to insufficient semantic depth.

3.2. Middle fusion (Cross-Attention or Modality Alignment)

Such as cross-attention, shared spatial projection, and contrastive learning between text and images. Representative models like CLIP use contrastive loss to align the outputs of text and image encoders within a unified semantic space [1]; BLIP-2 uses the Q-Former module to extract Visual Tokens as language input prompts [6]; Flamingo uses Gated Cross-attention Block to link text and image sequences [7]. This structure improves the intermodal semantic interaction abilities of current popular solutions for multimodal question-answering and dialogue tasks.

3.3. Decoding uniformity (Late Fusion / Unified Decoder)

In such structures, modal information is first processed by separate encoders, then combined and generated by a unified language model decoder, highlighting the "generation-driven" approach. Models like PaLM-E freeze the language model backbone and introduce modal information solely through a visual state encoding module, enabling task transfer [2]; GPT-4V directly integrates image tokens into the main Transformer of the LLM, allowing for unified processing with modal symmetry [3]. It is characterized by its high structural versatility and can be expanded into a multi-modal input, multi-task output system, but it is challenging to train.

3.4. Generalist MLLMs

The latest trend is the development of general-purpose large models that have unified capabilities across multimodal, multitask, and multilingual domains. Examples include Kosmos-2, GEMINI, and GPT-4o. These models integrate language, images, speech, and even code patterns structurally through a combination of instruction fine-tuning and context learning, enabling generalization in task execution. Future models will continue to evolve toward "multimodal agents" that focus on the role of language and perception in decision-making processes [8].

4. Introduction and comparative analysis of typical models

4.1. CLIP and BLIP Series Models: From Modal Alignment to Task Generation

Since Radford et al. proposed CLIP (Contrastive Language–Image Pretraining) in 2021, image-text models based on contrastive learning have officially become one of the mainstream directions in multimodal research[1]. As shown in Figure 1, CLIP trains on 400 million image-text pairs using two encoders (such as VIT) and a text encoder (like Transformer). It maximizes the cosine similarity between positive sample pairs to align the semantic space, thereby gaining strong zero-shot capabilities.

Chen et al. pointed out in subsequent studies that the structural advantage of the CLIP model lies in its "modular transferability," which makes it suitable for a variety of cross-modal retrieval and classification tasks. However, it still has limitations in terms of generative capabilities and semantic depth modeling [4].

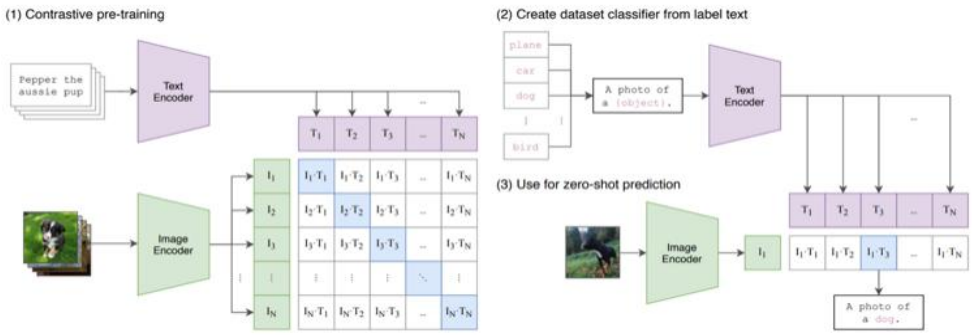


Figure 1. CLIP contrastive learning structure diagram.

To address the limitations of CLIP in generation and understanding tasks, Li et al. proposed BLIP, which introduces image-grounded prompting and a language reconstruction mechanism derived from the CAPTIONING/VQA pre-training task, thereby significantly enhancing the model's ability to convey visual information in language [5]. As shown in Figure 2, in BLIP-2, they further innovatively introduced the Q-Former module, which enables key information to be extracted from visual features using a set of LEARNABLE QUERIES and projected into a frozen language model for generation, thereby effectively achieving "perception-language decoupling" and task generalization [6].

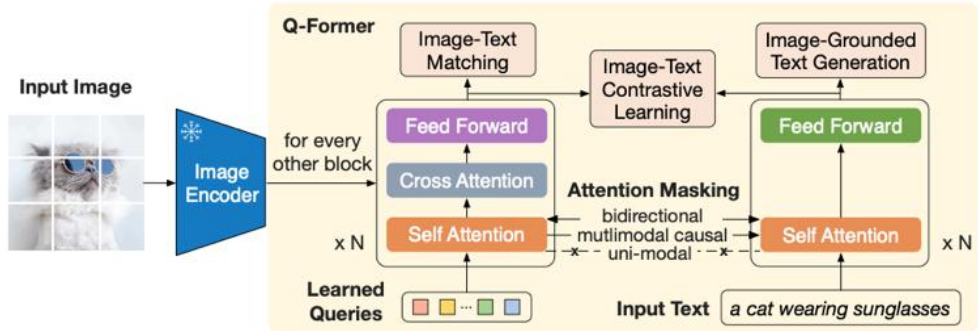


Figure 2. Q-former structure diagram [5].

Yu et al.'s comparative experiments show that BLIP-2 significantly outperforms current methods on metrics such as CIDEr, BLEU-4, and recall @1 across datasets, including MS-COCO Caption, Flickr30K, and VQAv2, demonstrating strong generalizability and scalability [8] [9]. A comparison of the BLIP and CLIP models is shown in Table 1.

Table 1. Differences between CLIP and BLIP models.

model	core mechanism	modal processing	typical task	Zero-shot capability	Structural characteristics
CLIP	Comparative learning with text and images	parallel embedding space	Retrieval, classification	Extremely strong	Decoupled dual decoding
BLIP	Image-guided text generation	middle fusion	Image and text generation, question and answer	general	Unified architecture for single entities
BLIP-2	Q-former + language model	Unified decoding	Multimodal dialogue and understanding	Medium strength	Separation of encoding and decoding structures

4.2. PaLM-E: A General-Purpose Agent Combining Language Models and Embodied Intelligence

As research on embodied intelligence advances, determining how to extend multimodal language models to robot perception and decision-making in real environments has become a crucial challenge. To tackle this, Driess et al. introduced the first general agent architecture that combines visual perception, state information, and large language models in 2023—PaLM-E (Embodied PaLM) [2].

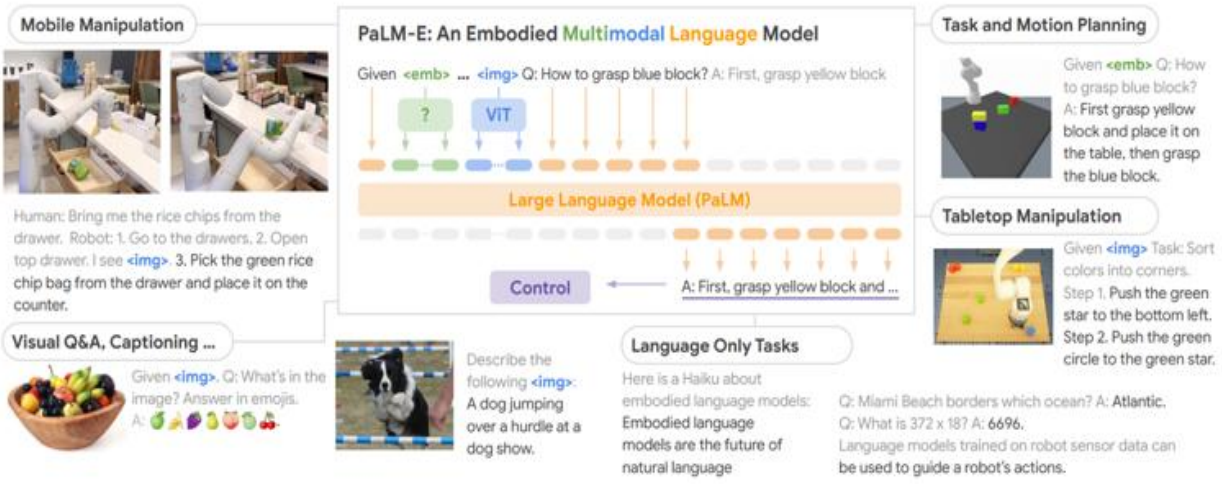


Figure 3. PALM-E structure diagram [2].

The core innovation of PaLM-E is its integration of a visual encoder and state, while utilizing a frozen language model (such as PaLM 540B) as the foundation for knowledge and reasoning. Its structure is illustrated in Figure 3. After the state-aware module transmits information to the language model, it encodes environmental data into tokens. This design ensures the model's versatility and controllability while greatly simplifying it. The model can be directly used in various scenarios, such as robot operations, grasping tasks, and language navigation, without the need for end-to-end fine-tuning.

Driess et al. noted in their empirical study that PaLM-E performed exceptionally well on the SayCan benchmark and R2-D2 robot platform, achieving high accuracy and reliability. It showed outstanding abilities in understanding natural language commands and turning them into actions [2].

This model also reflects the architectural trend of "multimodal modules + frozen LLM." Subsequent studies, such as RT-2, further integrate reinforcement learning and visual-language signals on this basis to achieve stronger concrete transfer capabilities [10].

The introduction of PaLM-E not only broadens the scope of LLM applications but also establishes the architectural groundwork for developing multimodal intelligent agents.

4.3. GPT-4V：Towards a unified intelligent entity: multimodal language models

GPT-4 with Vision (GPT-4V) is a versatile system released by OpenAI in 2023 that enables true "mixed text and image input, unified language model processing." [3], representing the current highest stage of multimodal model structure fusion.

Unlike traditional cross-attention mechanisms, GPT-4V uses a highly integrated approach: images are embedded as PatchTokens within the Transformer network and participate in the attention mechanism alongside text tokens. This unified input method (Natively Unified Input space) not only improves the quality of text-image interaction but also shows excellent transferability and generalization abilities when used in real-world tasks.

According to OpenAI's official report, GPT-4V can manage multi-round text-image conversations, handle complex visual reasoning tasks (like chart understanding and image question-answering), recognize mathematical expressions, and perform chained reasoning tasks. It has also set new records in several benchmark tests, such as TextVQA, DocVQA, and ChartQA [3].

This model continues with the 2024 GPT-4O, which enables real-time input and output loops across three modes: voice, image, and text. This represents the development of unified multimodal intelligent agents toward real-time interaction and low-latency reasoning [10].

As Jiang et al. point out, the architecture of GPT-4V essentially represents a "function-level modal alignment," with the goal no longer being feature space fusion, but rather task hierarchy of modal equivalence modeling (function-level equivalence) [11].

5. Application scenarios for multimodal large models

5.1. Image and text generation

Image and text generation is currently one of the most advanced application areas for multimodal models. Common tasks include image captioning, visual storytelling, and visual-grounded text generation.

For example, BLIP-2 achieved SOTA-level CIDER scores on the COCO CAPTION dataset [6], The Flamingo series model further supports multi-round image and text generation, demonstrating strong generalization capabilities in situations with limited samples [7]. Additionally, MiniGPT-4 can understand images after completing dialogue and description generation based on the combination of an image encoder and GPT-4, demonstrating high controllability and semantic understanding consistency [12] [13]. The primary evaluation metrics include BLEU, CIDER, SPICE, and others, which are used to assess the linguistic quality of text and the semantic alignment of images. As task requirements evolve, the future of text-image generation will shift from focusing on storytelling, reasoning, and perceptual skills in the "visual representation to sentence" process to a deeper understanding and contextual narrative transformation.

5.2. Cross-modal retrieval

Cross-modal retrieval is a common input-output cross-modal task, with the main goal of efficiently matching images and text, including image-to-text retrieval and text-to-image retrieval.

As a foundational model in this field, CLIP significantly enhances zero-shot retrieval capabilities and efficiency by creating a shared semantic space for text and images [1]. For example, in the Flickr30K dataset, CLIP's Recall@1 metric exceeds 75%, greatly surpassing traditional retrieval models. Additionally, Align (Jia et al., 2021) utilizes a larger-scale image-text pair dataset (1.6 billion) for pre-training, setting new performance benchmarks across multiple retrieval tasks [12-15]. Current technical efforts focus on improving the semantic consistency of long text retrieval, distinguishing complex details, and enabling cross-language retrieval.

Across various fields, including search engines, e-commerce recommendations, and video content analysis. For instance, Google has improved its image-text integration search and "search by image for products" technologies based on the CLIP/ALIGN series of models.

5.3. Multimodal dialogue and question-answering

Multimodal question answering (VQA, Visual Question Answering) and multimodal dialogue (Visual Dialogue) tasks require models to understand and respond to natural language questions based on images or videos, which is a common scenario involving multimodal perception and language reasoning.

Early representatives such as VILT (Kim et al., 2021) achieved end-to-end VQA models, while later models like LLaVA (Liu et al., 2023) and MiniGPT-4 introduced large language models with improved language generation abilities, such as Lava (Liu et al., 2023). VQA tasks usually measure accuracy, while visual dialogue thoroughly assesses the completeness of multi-round context and the quality of generated responses in language [13][16].

In 2023, the OpenFlamingo series advanced further in supporting text-image interactive dialogue, including multi-round turn-taking and context grounding. When combined with the GPT-4V model, it can manage complex flows of text and image information [3].

5.4. Robotics and Embodied Intelligence

The use of multimodal large models in embodied AI tasks has become a cutting-edge area of research in both academia and industry. Common tasks include object grasping, path planning, and human-computer interaction, where models must generate action strategies or language instructions based on environmental perception (e.g., camera images, sensor data).

For example, PaLM-E can produce natural language descriptions or action commands to perform robotic arm tasks after receiving image and status information [2]. RT-2 further proposes a three-model approach for Vision-Language-Action, enabling the transfer of knowledge from web pages to real-world applications operations [9]. This type of task assesses the ability to model temporal perception and understand language commands, requiring the integration of perception, planning, and action.

6. Challenges and development bottlenecks

6.1. Multimodal illusions and semantic consistency issues

Multimodal hallucination occurs when a model generates information that is unrelated or even incorrect regarding the visual input while producing text or answers, leading to issues like "inconsistent text and images" or "irrelevant answers”.

Liu et al. pointed out that although current visual language models can produce seemingly fluent answers, they often exhibit hallucination phenomena in complex tasks. The causes include limited image understanding depth, inaccurate semantic association modeling, and training data distribution bias [15]. This issue has become a key bottleneck affecting reliability, especially in professional tasks like medical image analysis and geographical scene question answering. Future solutions might include adding explicit text-image consistency modules, combining image segmentation or object detection priors, improving chain reasoning, and implementing intermediate inference verification mechanisms.

6.2. Data quality and cross-modal pre-training challenges

The key foundation for the performance of multimodal models is high-quality text-image, image-audio, video-language, and other aligned data. However, currently available multimodal datasets have the following issues: single scene distribution and lack of cross-modal generalization capabilities;

vague or redundant text descriptions that affect training signals; and a high proportion of synthetic data, which weakens actual semantic associations.

Hu et al. highlighted in their review of multimodal pre-training that current large models mostly depend on web crawl data (such as L AION-5B and YFCC100M), but their "weakly labelled" features often impact semantic quality, leading models to lack the ability to perceive details and evaluate linguistic facts [12][15].

6.3. Modal alignment mechanism design complexity

The current mainstream methods for modal fusion include shared embedding space (e.g., CLIP), cross-attention (e.g., Flamingo), query mechanisms (e.g., BLIP-2), and unified decoding (e.g., PaLM-E). Different mechanisms perform differently across various tasks, and a systematic, general-purpose modeling framework is lacking.

For example, in lengthy text input, the query mechanism can maintain visual features but results in low efficiency; meanwhile, fully fusing tokens increases computational load and amplifies modal noise [7].

Additionally, Chen et al. noted that the standardization of input formats for multimodal models is still incomplete. In most cases, there is incompatibility between the token encoding methods of different modalities, which impacts the robustness of multi-source information fusion [4].

6.4. Generalisation ability and low resource migration challenges

Although many MLLMs demonstrate strong transfer performance on large-scale datasets, they still encounter issues with limited generalization in small sample sizes, cross-domain, and low-resource scenarios, especially in the following three types of tasks: multilingual and multimodal scenarios, specialized fields (such as medicine, geography, and manufacturing), and multi-round text-image dialogue tasks.

Zhang et al. emphasize that multimodal models are much less adaptable to "task switching" than pure language models, especially when domain-specific fine-tuning data is lacking, leading to highly unstable performance [16].

6.5. Model training cost and deployment efficiency issues

The parameters of current mainstream multimodal models often exceed billions (e.g., Flamingo 80B and GPT-4V are estimated to be over 10 billion), requiring massive computing resources for training. The inference stage also imposes high demands on latency and memory. This significantly complicates the deployment of such models on edge devices, mobile devices, and dedicated terminals.

OpenAI noted in its GPT-4V report that the computational complexity of reasoning with text-image fusion inputs is roughly 1.5 to 2 times that of pure text, with this effect being especially noticeable in situations involving high-resolution image inputs and long-context conversations [3].

7. Future trends and research directions

7.1. Stronger contextual understanding and reasoning abilities

Existing multimodal models mainly focus on direct information extraction and generation when processing modalities like images and text. However, they still face difficulties with more complex reasoning tasks (e.g., identifying causal relationships in images or understanding characters' motivations). Future models will not just "read images" but will be capable of combining context and performing deeper semantic reasoning. This might involve integrating chain-of-thought (CoT) with more varied intermediate representation methods.

7.2. Model lightweighting and deployment optimisation

Currently, mainstream multimodal models typically have a large number of parameters, leading to high inference costs that limit their practical deployment. For example, when used on mobile devices, edge terminals, or in fields like education and healthcare, they encounter certain limitations. Future research will need to focus on reducing model complexity while maintaining performance. This can be achieved through more efficient methods such as parameter sharing, model compression, and small model distillation.

7.3. Multilingual and multimodal fusion modelling

Currently, multimodal models mostly rely on English data, and their ability to handle multilingual tasks is still limited. Future multimodal models should be capable of adapting to various cultural contexts and real-world user needs, such as Chinese text-image question-answering and the creation of Arabic video subtitles, along with other language and modality combinations. This requires developing more generalized language-modality alignment mechanisms based on multilingual data pre-training.

7.4. Deep customization of application scenarios

Future research should focus not only on developing the technology itself but also on addressing practical application issues. For example, in fields such as medical image analysis, forensic evidence modeling, and educational tools, multi-modal models (Multi-Style Model) can be combined with domain knowledge to optimize tasks. An important area for future research is how to ensure the generalizability of models while quickly adapting to the specific needs of different industries.

8. Conclusion

Multimodal large language models (MLMs) have become a popular topic in current research as artificial intelligence continues to advance. This paper systematically reviews the development background, structural classification, key models, typical applications, existing challenges, and future trends of multimodal large models, aiming to fully present their technological evolution and current application status.

From a model architecture perspective, representative models like CLIP, BLIP, PaLM-E, and GPT-4V have progressed from early methods of input concatenation and contrastive learning to modern approaches that unify decoding and pattern-sharing spatial modeling. This evolution showcases different ways of pattern integration and highlights a clear trend toward unified architecture design.

At the application level, MLMS has been widely applied in scenarios such as image and text generation, cross-modal retrieval, visual question answering, multimodal dialogue, and body intelligence, and has achieved leading performance in multiple tasks. However, the model still faces many challenges, including modal hallucination phenomena, high training costs, lack of reasoning capabilities, and deployment efficiency issues, all of which require the model to address.

Future research on multimodal large models should continue to expand the generalization capabilities of reasoning tasks involving context understanding and resource allocation under resource-limited conditions. It should also focus more on integrating and applying these models to real-world situations. Additionally, this will become a key direction for ongoing progress in this field, aiming to develop a unified model that combines multilingual, multi-task, and multimodal capabilities.

Overall, multimodal large models not only give artificial intelligence a more "human-like' way of processing, but also provide vital support for building artificial general intelligence (AGI). As technology advances in the future, MLMs will play an increasingly significant role in academic research and industry practice.

References

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

[2] Driess, D., Xia, F., Chowdhery, A., et al.: 'PaLM-E: An Embodied Multimodal Language Model', arXiv Prepr., 2023, arXiv:2303.03378

[3] OpenAI: 'GPT-4 Technical Report', Tech. Rep., OpenAI, 2023

[4] Chen, M., Zhang, Y., Fang, Z., et al.: 'Multimodal Pretraining: A Survey', arXiv Prepr., 2023, arXiv:2302.00487

[5] Li, J., Selvaraju, R.R., Gotmare, A., et al.: 'ALBEF: Cross-modal Contrastive Learning for Vision-Language Pre-training', arXiv Prepr., 2021, arXiv:2107.07651

[6] Li, J., Li, D., Xiong, C., Hoi, S.: 'BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models', arXiv Prepr., 2023, arXiv:2301.12597

[7] Alayrac, J.B., Donahue, J., Luc, P., et al.: 'Flamingo: A Visual Language Model for Few-Shot Learning', arXiv Prepr., 2022, arXiv:2204.14198

[8] Yu, Z., Li, D., Liu, S., et al.: 'Instruction Tuning for Multimodal LLMs: Methods and Evaluations', arXiv Prepr., 2023, arXiv:2309.13036

[9] Brohan, A., Chebotar, Y., Driess, D., et al.: 'RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control', arXiv Prepr., 2023, arXiv:2307.15818

[10] OpenAI: 'GPT-4o: Introducing GPT-4 Omni', Tech. Rep., OpenAI, 2024

[11] Jiang, W., Zhang, J., Lin, X.: 'Multimodal Large Language Models: A Survey', arXiv Prepr., 2023, arXiv:2305.06541

[12] Hu, Y., Shen, J., Liu, Z., et al.: 'Survey of Vision-Language Pretrained Models', Sci. China Inf. Sci., 2023, 66, (11), 182101

[13] Zhu, D., Hu, Z., Liu, Y., et al.: 'MiniGPT-4: Enhancing Vision-Language Understanding with GPT-4', arXiv Prepr., 2023, arXiv:2304.10592

[14] Jia, C., Yang, Y., Xia, Y., et al.: 'Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision', arXiv Prepr., 2021, arXiv:2102.05918

[15] Liu, H., Li, J., Lin, J., et al.: 'Visual Instruction Tuning', arXiv Prepr., 2023, arXiv:2304.08485

[16] Zhang, Y., Li, T., Feng, Y., et al.: 'Multimodal Large Language Models: Recent Advances and Future Trends', arXiv Prepr., 2024, arXiv:2401.03461