The Revolution of Multimodal Large Language Models: A Survey

Caffagni D.*[1], Cocchi F.*[1,2], Barsellotti L.*[1], Moratelli N.*[1],
Sarto S.*[1], Baraldi L.*[2], Baraldi L.[1], Cornia M.[1], Cucchiara R.[1]
University of Modena and Reggio Emilia, Italy
[1]{name.surname}@unimore.it [2]{name.surname}@phd.unipi.it

ACL 2024 Bangkok, Thailand

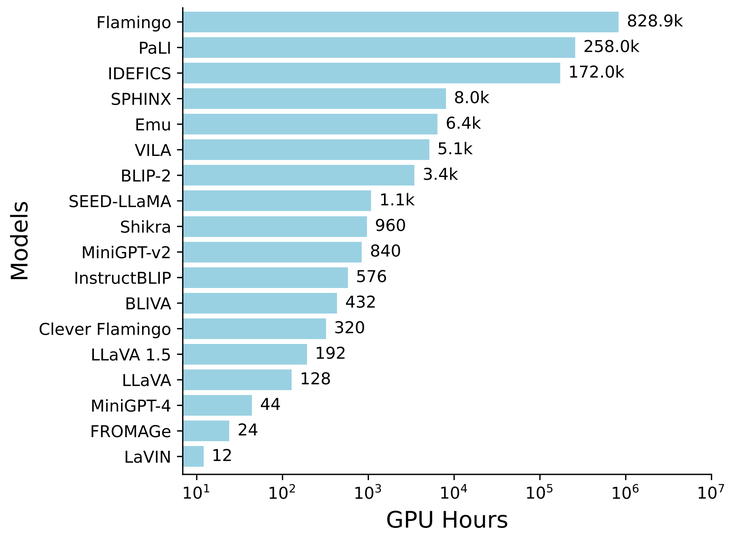
Motivation for the Survey

- **Unified Understanding of Multiple Modalities.** We provide a comprehensive understanding of how LLMs have been adapted to perform tasks that require the interaction of multiple modalities.
- **Benchmarking and Evaluation.** The rapid development and release of MLLMs necessitate a standardized evaluation approach. Hence, we provide insights on the training data, evaluation datasets, and performance metrics.
- **Guiding Future Research.** We identify the current state and the key challenges in the research field of MLLMs, to inspire future research endeavors.

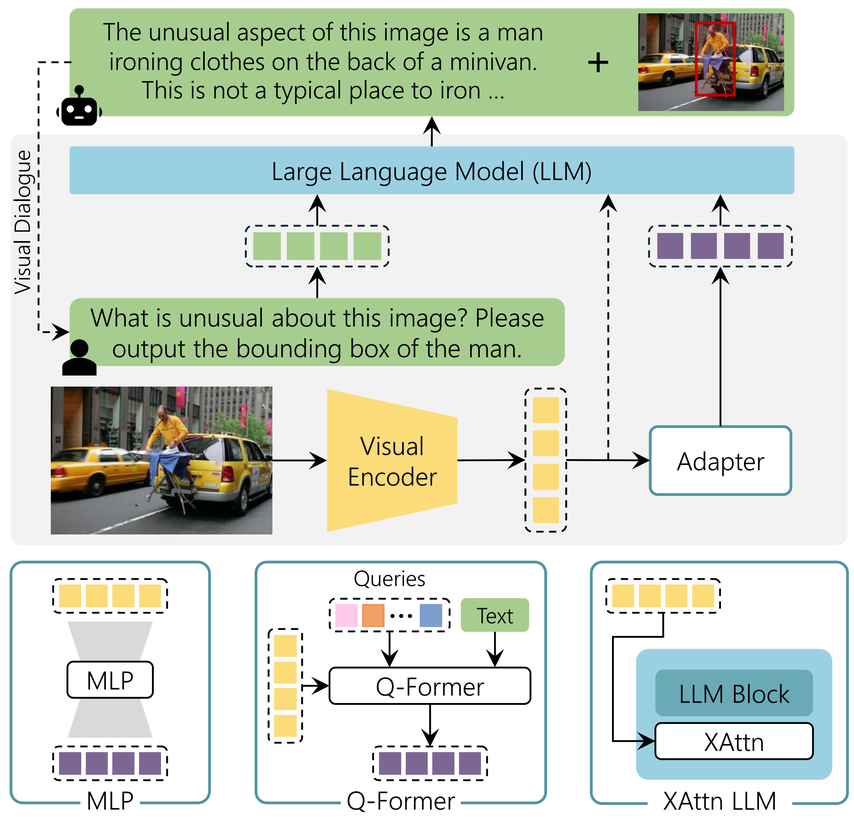
Training Strategies of MLLMs and Requirements

Training usually consists of a single-stage or two-stage process, leveraging cross-entropy loss to predict the next token and align visual and textual information.

- **Single-Stage Training.** Methods like LLaMA-Adapter [9] and Kosmos-1 [1] train visual and textual components simultaneously, often with joint training using image-text pairs and instructions.
- **Two-Stage Training.** Approaches such as LLaVA [3] and MiniGPT-4 [10] initially align image features with text embeddings prompting the pipeline to caption images. Following this, a second stage is performed to enhance multimodal conversational capabilities.



The Recipe for MLLMs



- **LLM.** A large model [7] designed to handle textual information utilizing in-context learning and instruction-tuning to align its behavior with human expectations.
- **Visual Encoder.** It commonly employs Vision Transformers (ViT) trained with contrastive learning to align visual and textual embeddings, with popular choices being CLIP [4] and EVA-CLIP [6] for providing visual features.
- **Vision-to-Language Adapters.** These modules facilitate interoperability between visual and textual domains.
  - **Linear and MLP Projections**: Simple linear layers or MLPs translate visual inputs into textual embeddings effectively.
  - **Q-Former**: A Transformer-based model [2] with learnable queries and shared self-attention layers for aligning visual and textual representations.
  - **Cross-Attention Layers**: Added to LLMs to integrate visual information, often paired with mechanisms like Perceiver to reduce computational complexity.

Visual Grounding

Visual grounding in MLLMs enables interactive dialogues that involve positioning content, requiring the ability to both understand and localize regions based on textual descriptions. Key approaches to achieving visual grounding include:

- **Region-as-Text**: Converts visual regions into text by representing bounding boxes or polygons as coordinates within the generated text.
- **Embedding-as-Region**: Employs region encoders to process input regions and outputs embeddings extracted from the last layer of the MLLM to a decoder, which produces the output regions.
- **Text-to-Grounding**: Uses open-vocabulary models that can interpret textual categories and link them to visual regions, enabling grounding of text descriptions in images.

Image Generation and Editing

This advancement is realized through integrating MLLMs with image generation pipelines.

- **Integration with frozen Diffusion Models.** This approach maps the output of a frozen LLM to a frozen diffusion model using a trained mapper, avoiding the need to fine-tune both components.
- **Fine-Tuning Strategies.** The LLM is fine-tuned to improve multimodal generation capabilities, using reconstruction loss to enforce alignment with diffusion models. Other alternatives refine discrete and continuous visual tokens.
- **End-to-End Training.** This method entails directly fine-tuning diffusion models with embeddings generated by the LLM or integrating LLMs with discrete image encoders like VQ-GAN for joint image and text token prediction.

Other Multimodal Applications

- **Any-Modality Models.** Designed to handle multiple modalities simultaneously, such as images, videos, audios, 3D data, and IMU sensor signals. Notable approaches include Transformer-based alignments like Q-Former and Perceiver, as well as models like NExT-GPT [8] and Unified-IO 2 [5], which can also generate outputs across different modalities.
- **Domain-Specific MLLMs.**These include models for document analysis, embodied AI and robotics, and specialized fields such as medicine and autonomous driving. These models are either trained from scratch or fine-tuned from existing MLLMs to address unique domain-specific needs.

Prominent Research Directions

- **Multimodal Retrieval-Augmented Generation**: While retrieval-augmented generation RAG is a consolidated technique in LLMs, its application in MLLMs is still under-explored.
- **Correction of Hallucinations**: MLLMs tend to exhibit high hallucination rates, especially when generating longer captions. Understanding and correcting the underlying causes of hallucinations remains an important challenge.
- **Prevent Harmful and Biased Generation**: Ensuring the safety and fairness of large-scale models is of fundamental interest in the community. Models trained on web-crawled data are prone to generating inappropriate and biased content.

[1] Huang et al. *Language is not all you need: Aligning perception with language models*. NeurIPS, 2023.
[2] Li et al. *BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models*. ICML, 2023.
[3] Liu et al. *Visual Instruction Tuning*. NeurIPS, 2023.
[4] Radford et al. *Learning transferable visual models from natural language supervision*. ICML, 2021.
[5] Lu et al. *Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision Language Audio and Action*. CVPR, 2024.
[6] Sun et al. *EVA-CLIP: Improved Training Techniques for CLIP at Scale*. arXiV, 2023.
[7] Touvron et al. *LLaMA: Open and Efficient Foundation Language Models*. arXiv, 2023.
[8] Wu et al. *NExT-GPT: Any-to-Any Multimodal LLM*. ICML, 2024.
[9] Zhang et al. *LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention*. ICLR, 2024.
[10] Zhu et al. *MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models*. arxiv, 2023.