

Open in app ↗

Medium

Search

Write

★ Get unlimited access to the best of Medium for less than \$1/week. [Become a member](#)

Review of a Survey of Contemporary Multi-modal A.I Models



Alexis Ambriz

5 min read · Just now



What makes them special? Let's find out.

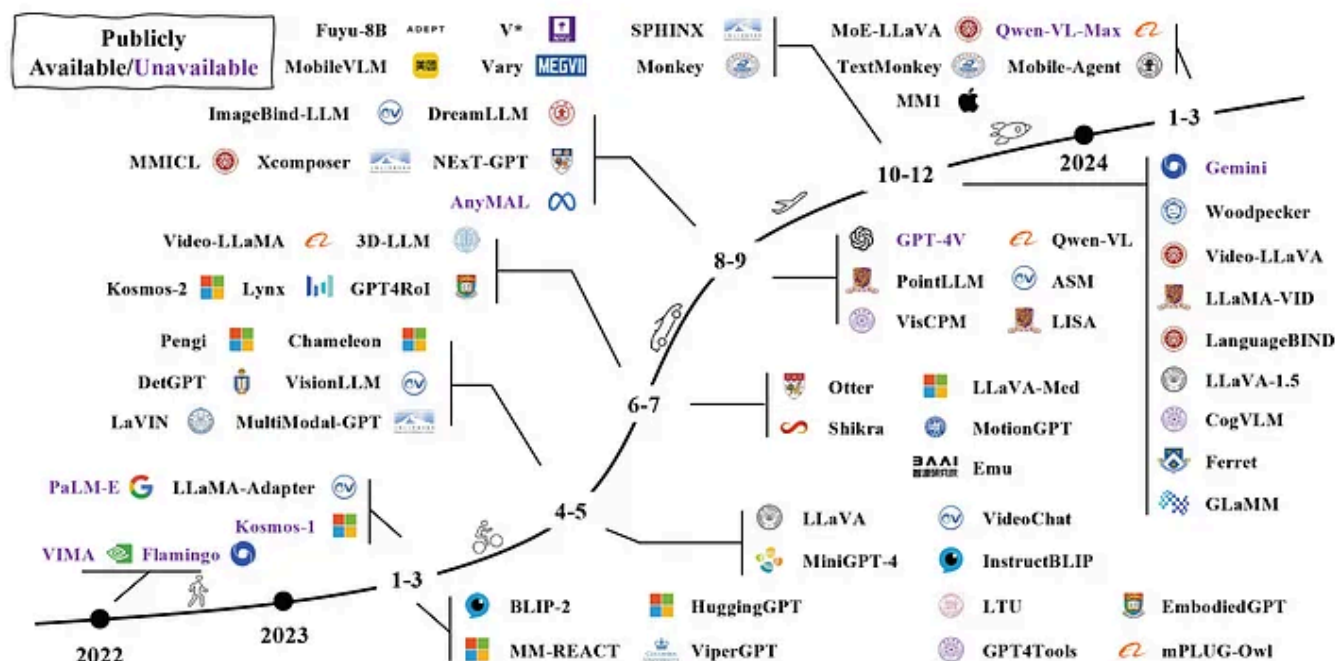


Fig. 1: A timeline of representative MLLMs. We are witnessing rapid growth in this field. More works can be found in our released GitHub page, which is updated daily.

A image taken from the original survey research paper

In the paper '[A Survey on Multimodal Large Language Models](#)' by [Shukang Yin et.al](#), the authors survey many current research papers covering the development of multi-modal large language models, also known as MLLM's. They mention that these types of models are fairly recent innovations, so looking at the state of the current best practices is a good idea if you are interested in understanding current gaps in research. In summary, the authors decided to cover 5 topics and primarily focused on three of these. The 5 topics that were covered include architecture, data & training strategies, ideological background topics, extensions in applications, and additional MLLM specific techniques.

In this short article, I'll be primarily covering the core of these — to avoid going too deep into the weeds! I'll be briefly covering the MLLM architecture & data & training strategy. I'll be going a little more in-depth in the ideological underpinnings, as well as on additional MLLM-specific techniques. I believe that the last two topics might be a bit more informative for people who have not built their own multi-modal LLM model from scratch and don't yet understand all the technical aspects.

Ideological Background Research Topics

Interface Frameworks

According to the authors, most of the research included a common interface: a LLM that can recognize whether a user's input query contains additional modalities — and then serves as a 'pivot' to handle mixed input queries and shuttle them off to alternative pre-processing steps. This interface could thus be extended to incorporate many model 'agents' that can pre-process many inputs and accomplish different tasks, while the core

LLM model serves as a synthesizer and ultimately responds to the initial query.

Mixed-Modality Input & Output

For example, the authors discussed how a majority of the works use a ‘diffusion module’ that processes the mixed-modality input and converts them into tokens that could then be understood when passed to the base LLM. The diffuser module is in a way its’ own model and contains parameters that have learned to understand a variety of data on its’ own.

Key Technical Topics

The Architecture

Most of the research that the authors looked at featured multi-modal models that utilized a modality encoder and a modality ‘interface’ along with the core LLM model.

The Data and Training Strategy

The authors note that many multi-modal models recognize the necessity to incorporate a variety of data that is not only multi-modal, but recognizes the variance in granularity, multi-language support, and domain-specific knowledge to avoid overfitting and/or training a model that is only generalizable to specific contexts. For example, it is possible to improve a MLLM models’ response by improving a user’s ability to interact with their multi-modal inputs as they do with text — in a more granular way. In a text chain, it is common to refer back to a specific detail in a previous message to have the model recognize a previous piece of information. In image modalities, it is possible to provide a user with a segmented image and allow for users to highlight specific segments within the image they would like the

model to refresh within their memory context (i.e using a natural language pointer to a bounding box within the image).

They also note that many common dataset's that are used to train multi-modal models are lacking in multi-lingual support as they are sourced by scraping the internet for images or videos that contain text captions. Given that the majority of the datasets were generated by scraping primarily english content, there is a lack of datasets to train multi-modal models in other languages in a way that retains cultural specific references while reducing hallucinations.

Additional MLLM Techniques

M-ICL (Multi-Modal In-Context Learning)

The authors claim that this is one of the most important capabilities in training a MLLM as it doesn't rely on re-training using an abundance of multi-modal training data. As such, a multi-modal ICL template can be improved and applied quickly and significantly impact the output of the model. It differs from traditional ICL techniques as-in single-modality models as they incorporate a variety of modalities and exemplars within the inference prompt system template.

M-COT (Multi-Modal Chain-of-Thought)

This technique is similar to non-multimodal chain-of-thought techniques as they both rely on decomposing a user input question into multiple queries. The model will then be able to respond to the original question by responding to multiple smaller queries before combining the reasoning from all responses and inferring one final response to the original question. This technique is also usually only applied for complex datasets where providing more scratch-pad space when answering a question is needed.

Specific to MLLM architectures, however, the authors found that most M-COT implementations in the research they reviewed use a few or zero-shot training paradigm — thus making the pre-training stage one of the most simple to implement. M-COT also usually implements single-chain chains of thoughts as opposed to the more advanced Markov Decision Tree structure chains — we may assume they may be harder to implement in a multi-modal context. The authors do mention however that they found a paper that proposed ‘DD-COT’ or ‘Duty-distinct Chain of Thought’ that utilizes a tree chain data structure given multi-modal input.

LAVR (LLM-Aided Visual Reasoning)

Finally, in the context of combining text with images and video, the authors found that the LAVR technique is able to apply strong generalization abilities, other emergent capabilities such as finding hidden meaning within images (such as when given a meme as input), as well as providing improved interactivity and control to users. LLM-Aided visual reasoning is in essence a multi-agent system that contains models that are trained specifically for one modality. The base LLM is provided with helper LLM’s to reason about the multi-modal inputs’ that it doesn’t recognize itself. Separating the modalities as opposed to having a single multi-modal model allows for the base LLM model to serve as the task delegator, decision maker, and ‘semantics refiner’(i.e to improve language understanding).

[Data Science](#)[Software Engineering](#)



Written by Alexis Ambriz


Edit profile

0 Followers

SWE & Data Analyst

Recommended from Medium



 Abdul Hanan

The Art of Goal Setting: Turning Your Dreams into Actionable Steps

Goal setting is much more of an art than a science—a guide to the life that one wants,...

★ 1d ago 🖱 580 💬 16  



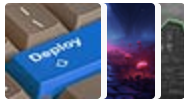
 Jessica Stillman

Jeff Bezos Says the 1-Hour Rule Makes Him Smarter. New...

Jeff Bezos's morning routine has long included the one-hour rule. New...

★ Oct 30 🖱 8.8K 💬 182  

Lists



Predictive Modeling w/ Python

20 stories · 1676 saves



General Coding Knowledge

20 stories · 1752 saves



Stories to Help You Grow as a Software Developer

19 stories · 1476 saves



Practical Guides to Machine Learning

10 stories · 2031 saves



Abdur Rahman in Stackademic

Python is No More The King of Data Science

5 Reasons Why Python is Losing Its Crown



Oct 22



6.8K



29



Tazeem Muqaddas

“Think Social Media Is Harmless? The Truth Will Shock You”

Social media has ingrained itself into our daily lives in the modern world. The majority of us...



1d ago



96



14



Medium Staff in The Medium Blog

It happened on Medium: October 2024 roundup



Harendra

How I Am Using a Lifetime 100% Free Server

Much-highlighted insight, notable new publications, and evergreen lists

4d ago  5.7K  82



Get a server with 24 GB RAM + 4 CPU + 200 GB Storage + Always Free

 Oct 25  5K  67



See more recommendations