Samsung Innovation Campus

Chapter 9. Deep Learning Module

Multimodal Large Language Models (MLLMs)

Large Language Models (LLMs)

What is a LLM?

A category of **foundation** models trained on extremely vast datasets in order to make them capable of understanding and generating natural language content.

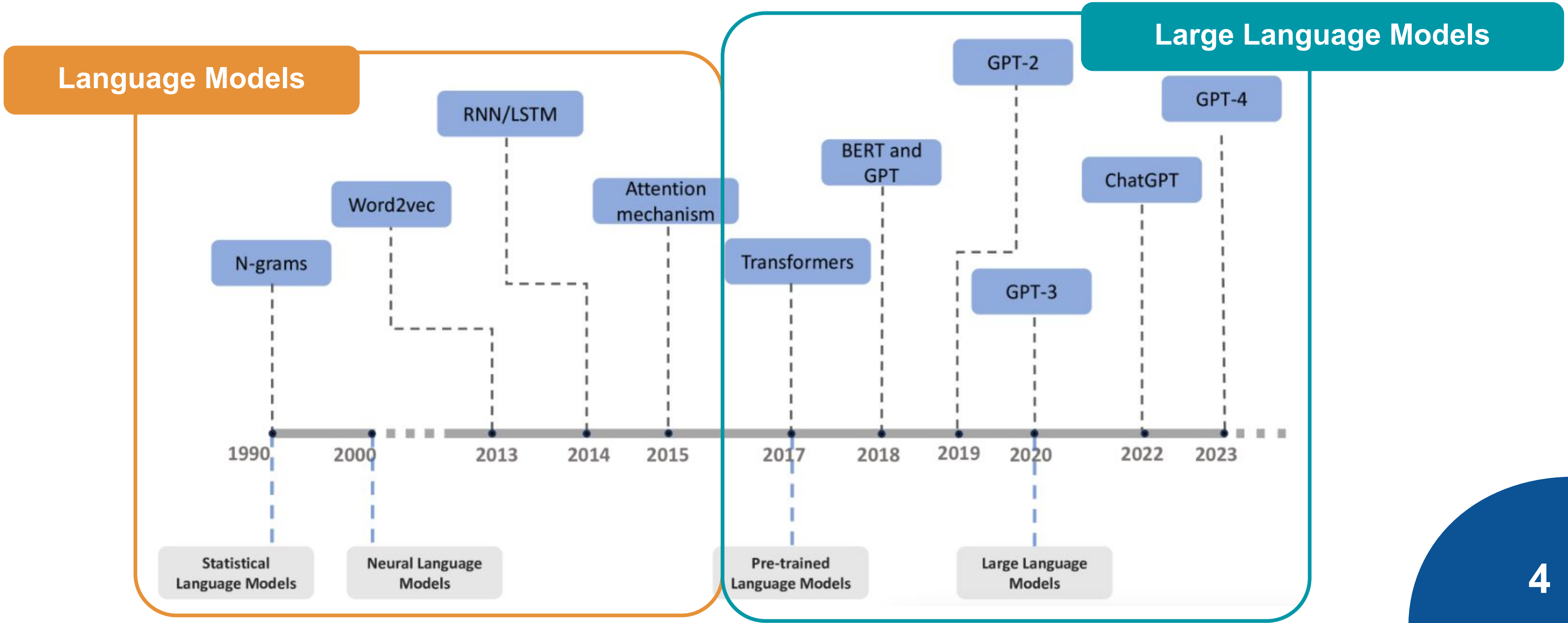
LLMs can generate human-like responses based on context

Question-Answering	Translation	Text Summarization
		

1 - History, Development, and Principles of Large Language Models

Language Modeling: The Concept

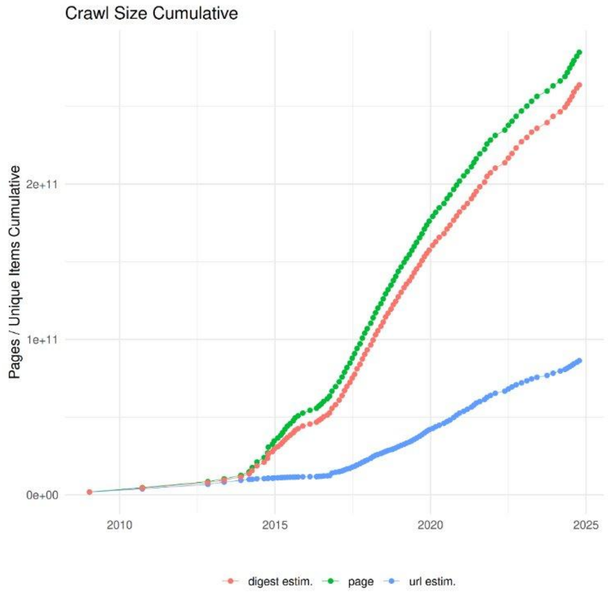
Definition: A probabilistic approach that involves predicting the next word in a sentence or sequence of words based on the context and previous words.



1 - History, Development, and Principles of Large Language Models

What made these models Large Language Models?

More Data!



Computer Power

Requires massive GPUs/TPUs and distributed computing

More Parameters

Fom millions (BERT) to trillions (GPT-4) of parameters

Architecture

Transformers & Attention Mechanisms

1 - History, Development, and Principles of Large Language Models; 2 - Common Crawl

What makes these models **Large** Language Models?

We are witnessing an **exponential growth** in model size!

More parameters

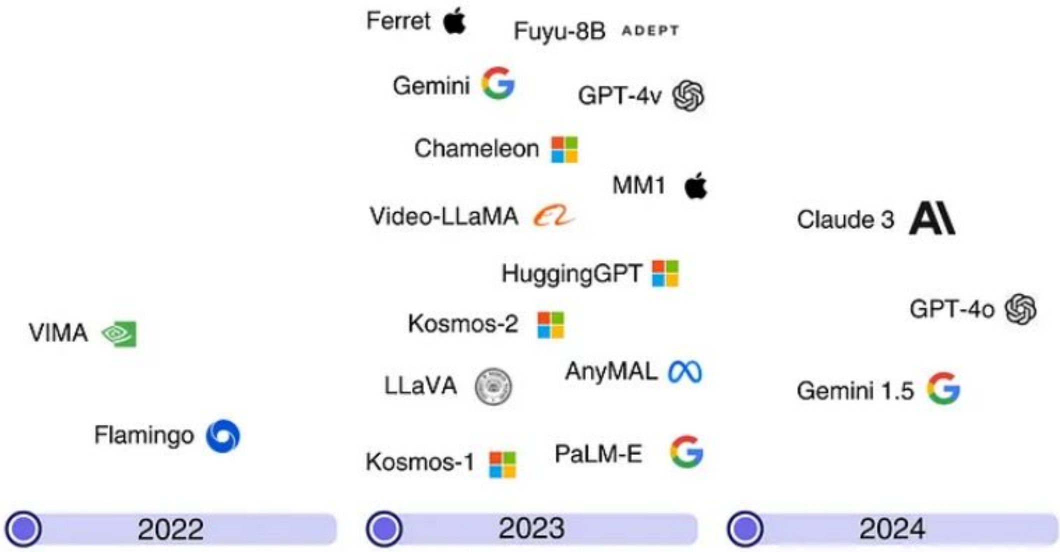
- ⬆ Fluency, reasoning, and generalization
- ⬆ Computational cost

Beyond Text: The next frontier is Multimodal LLMs

The Rise of Multimodal LLMs

MLLMs are models that process and integrate multiple types of **data modalities.**

Text	Images	Audio	Sensory Data



1 - MLLMs and their impact in CV

MLLMs: Why should we care?

These models enable "more natural" interactions and allow for advanced problem-solving capabilities that combine **different types of information**!

Document Analysis	Extracting both textual and visual information from documents (e.g., CV analysis)
Healthcare	Combining medical imaging with patient records to improve diagnostics
Autonomous Systems	Assists in navigation by integrating visual, auditory and textual cues (e.g., autonomous driving systems)
Content Creation & Digital Art	Generating or editing images based on textual inputs

1 - MLLMs and their impact in CV

The Architecture of MLLMs: Components

1. **Modality Encoder**

Text Encoder | Vision Encoder | Audio Encoder | Video Encoder

2. **Modality Integration**

Adapters/Connectors, Integration with a LLM

3. **Decoder**

Generates the final output. Often a pre-trained LLM!

1 - The Revolution of Multimodal Large Language Models: A Survey

The Architecture of MLLMs: Modality Encoders

Text Encoder

Converts textual input to dense vector representations

Examples:

- GPT
- LLaMA
- PaLM

Vision Encoder

Converts images into feature representations compatible with text

Examples:

- ViT
- Swin Transformer
- CLIP's vision encoder

MLLMs are often built on top of a **pre-trained LLM**:
the text encoder is the one already present in the LLM

1 - The Revolution of Multimodal Large Language Models: A Survey

The Architecture of MLLMs: Modality Encoders

Audio Encoder

Converts raw audio into text embeddings or direct speech representations.

Examples:

- Whisper
- Wav2Vec
- Spectrogram Transformer

Video Encoder

Captures both **spatial** and **temporal** information

Examples:

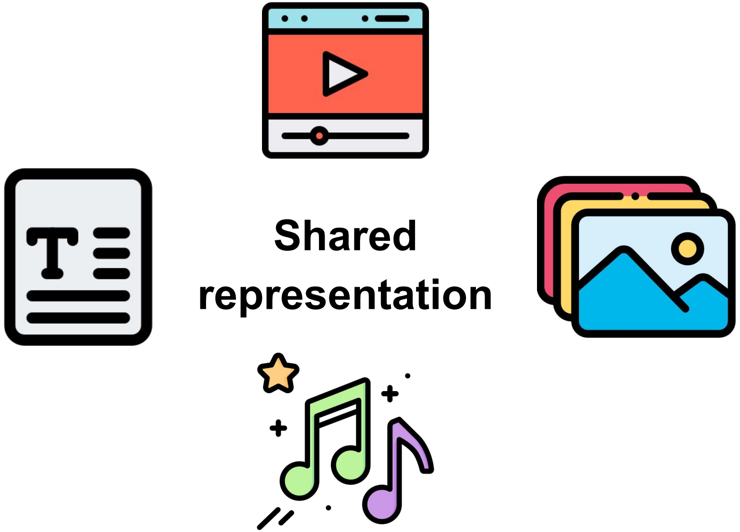
- TimeSformer
- VideoPrism
- VideoMAE

1 - VideoMAE; 2 - Sparks of Large Audio Models

The Architecture of MLLMs: Bridging the Modality Gap

How can we integrate these modalities with a pre-trained LLM?

LLMs only process text so we need to find ways to connect them to all other modalities - **We need to bridge the modality gap!**



Learnable Connectors

Expert Models

Bridging the Modality Gap: Learnable Connectors

Goal: Adapt **non-text** modalities by projecting information into a space the LLM can understand

Based on **how** multimodal information is **fused**, connectors can be implemented at:

Token-Level	**Feature-Level**
Encoder outputs are transformed into tokens and **concatenated** with text tokens before being sent into LLMs	Inserts **extra modules** that enable deep interaction between modalities

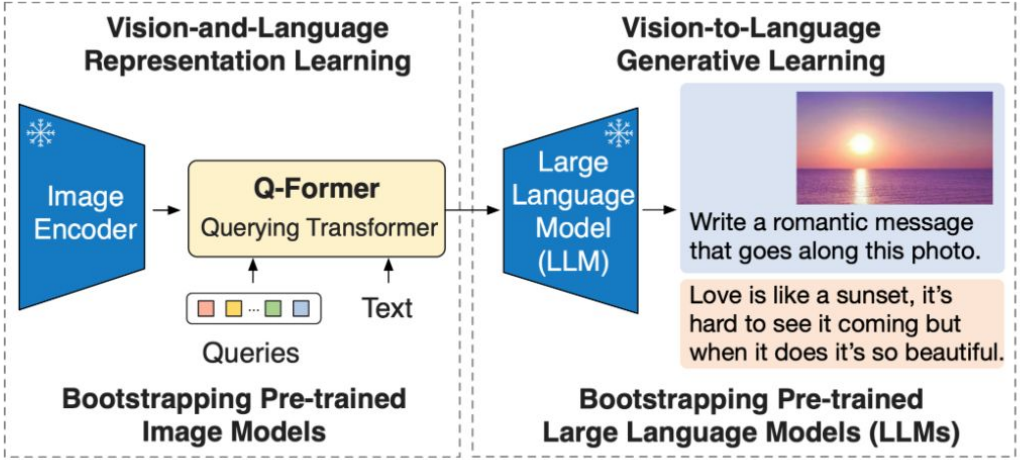
1 - A survey on multimodal large language models

Learnable Connectors: Token-Level

Q-Former: Contains a set of **learnable query embeddings** that attend to the images features through cross-attention

- Each query learns to extract specific types of visual information

- The processed queries become a **smaller**, more focused set of visual tokens

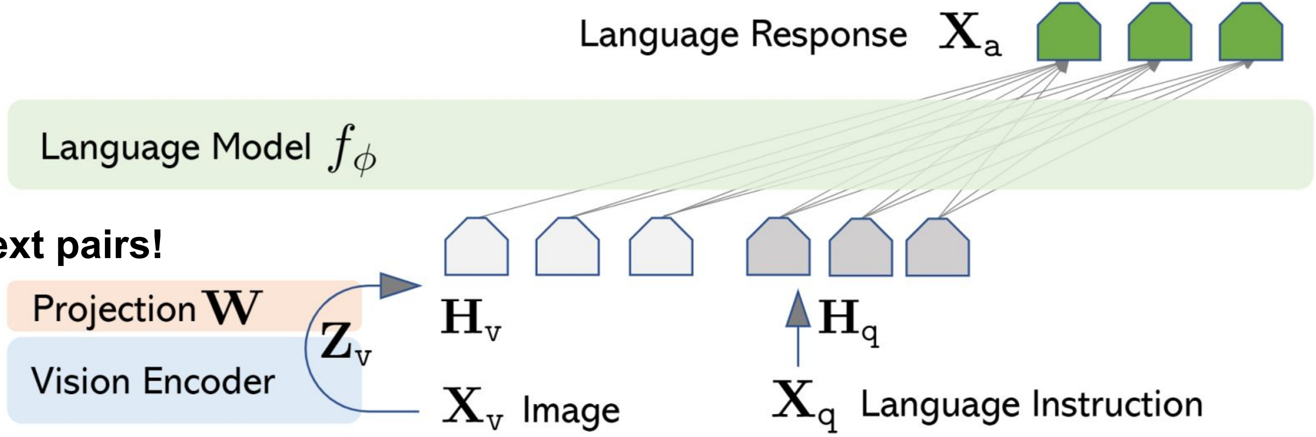


1 - [A survey on multimodal large language models]; 2 - [BLIP2 Paper]

Learnable Connectors: Token-Level

MLP Adapters: Uses a MLP module to **project** the visual features into the same embedding space as text.

Trained on image-text pairs!

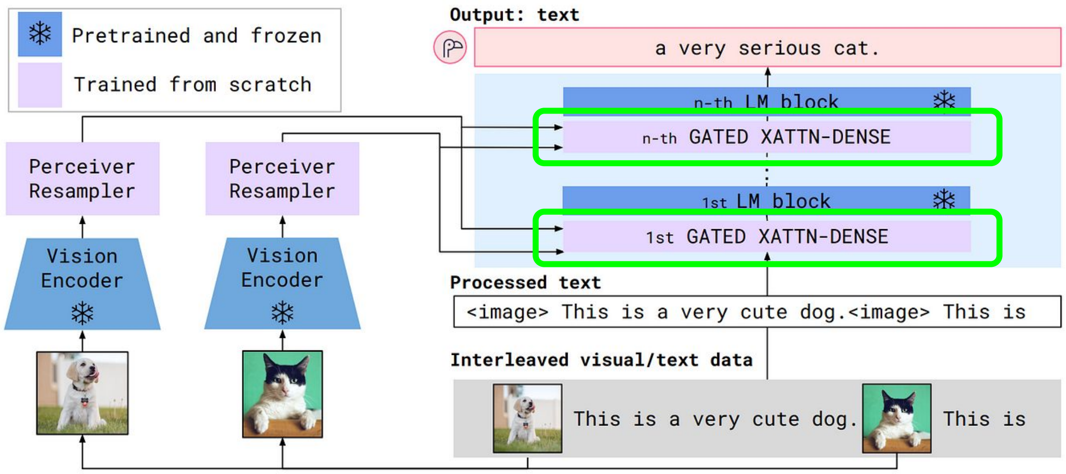


$$H_v = W \cdot Z_v + b$$

1 - A survey on multimodal large language models; 2 - Llava Models

Learnable Connectors: Feature-Level

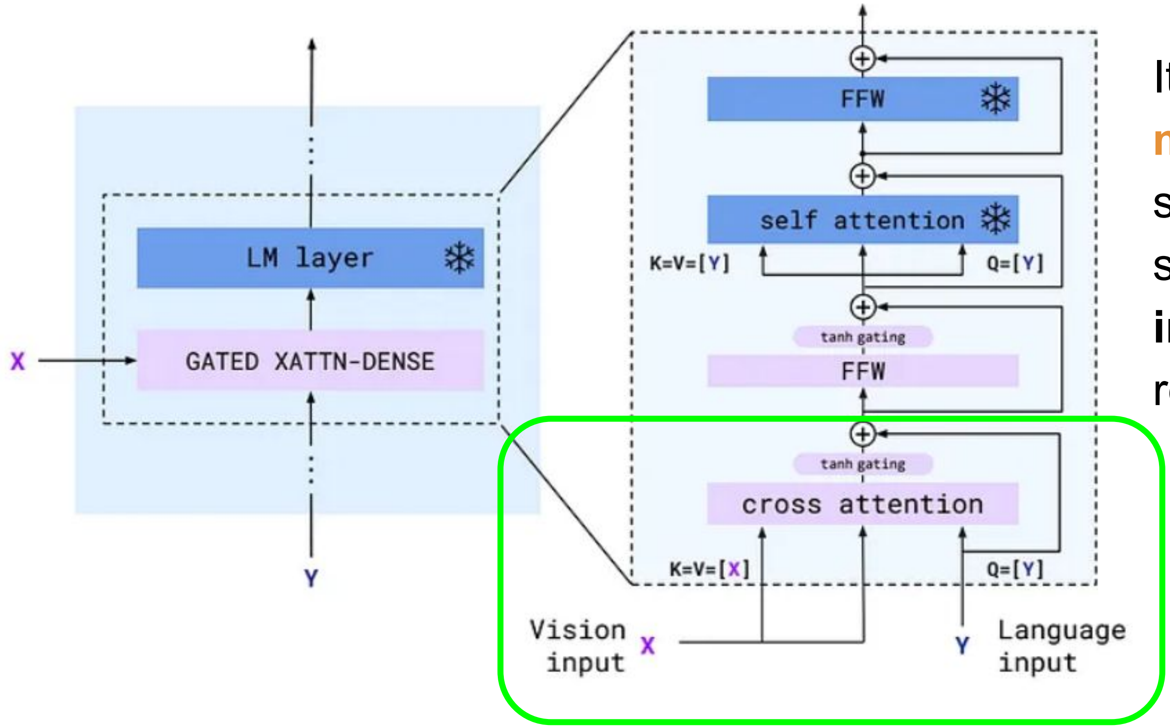
Cross-Attention: Enables deep interaction between modalities by inserting **extra layers** between the frozen LLM and the vision inputs.



Instead of **adapting** the vision/audio features to fit the LLM, cross-attention layers allow the LLM to process them as **contextual information**

1 - A survey on multimodal large language models; 2 - Flamingo Paper

Learnable Connectors: Feature-Level



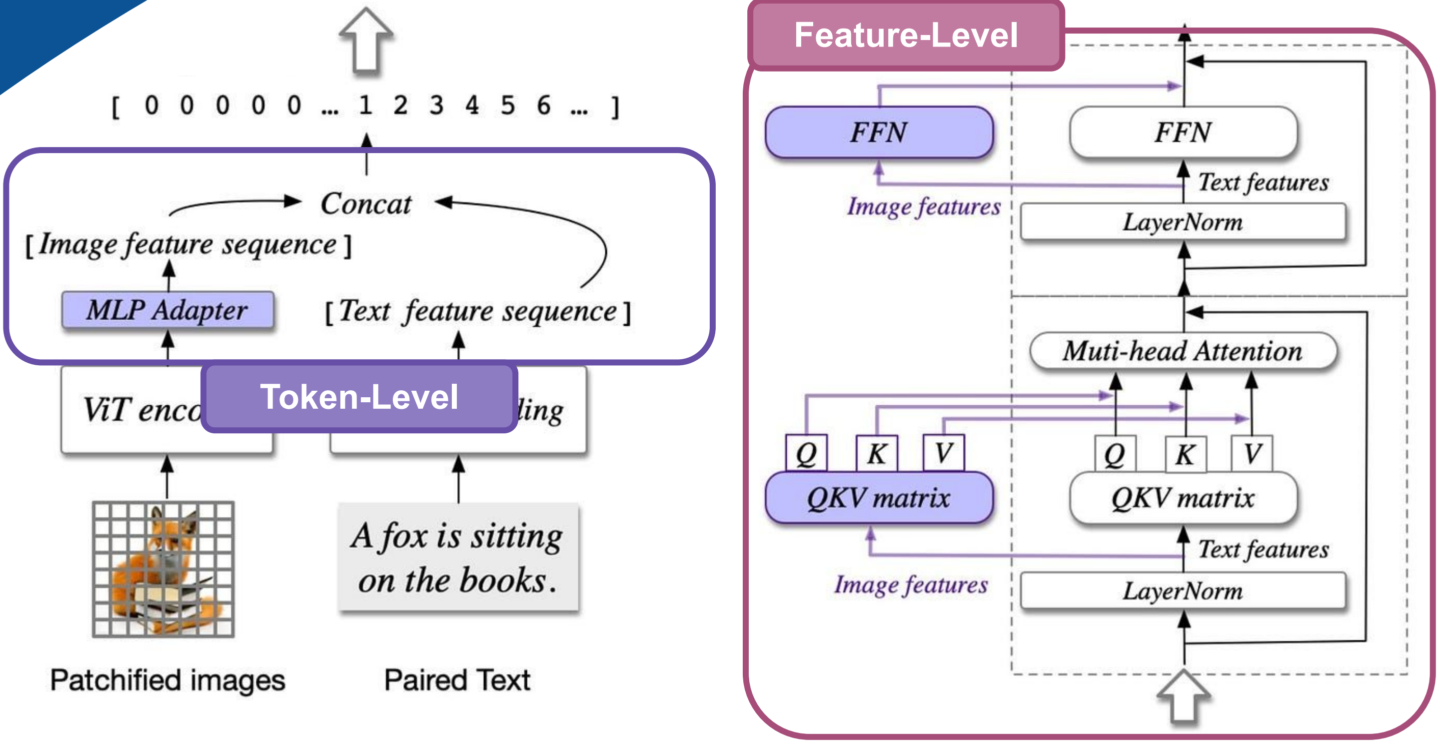
It leverages a **gated mechanism**: it allows to selectively control how much should the visual content **influence** the text representation

1 - A survey on multimodal large language models; 2 - Flamingo Paper

Learnable Connectors: Feature-Level

Visual Expert Module: Trainable visual processing unit directly inserted inside the transformer blocks of the LLM. Allows the model to process the **multimodal context** across all transformer layers.

1 - A survey on multimodal large language models; 2 - CogVLM Paper



(a) The input of visual language model

(b) The visual expert built on the language model

Image features are processed through a **separate set** of QKV matrices, specifically dedicated to visual tokens

Figure 3: The architecture of CogVLM. (a) The illustration about the input, where an image is processed by a pretrained ViT and mapped into the same space as the text features. (b) The Transformer block in the language model. The image features have a different QKV matrix and FFN. Only the purple parts are trainable.

1 - A survey on multimodal large language models; 2 - CogVLM Paper

Bridging the Modality Gap: Expert Modules

Goal: Convert multimodal inputs to language **without the need for training**

So **expert modules** are pre-trained, task-specific models that will process a specific modality to convert it into a textual representation before passing it to the LLM

Works well for applications like **video-to-text** or **speech-to-text**

Information Loss

Text cannot fully represent spatial or temporal relationships

1 - A survey on multimodal large language models;

Expert Modules

The VideoChat architecture attempts to bypass the information loss limitation of expert modules by also including video embeddings as input in order to improve **spatial-temporal reasoning**

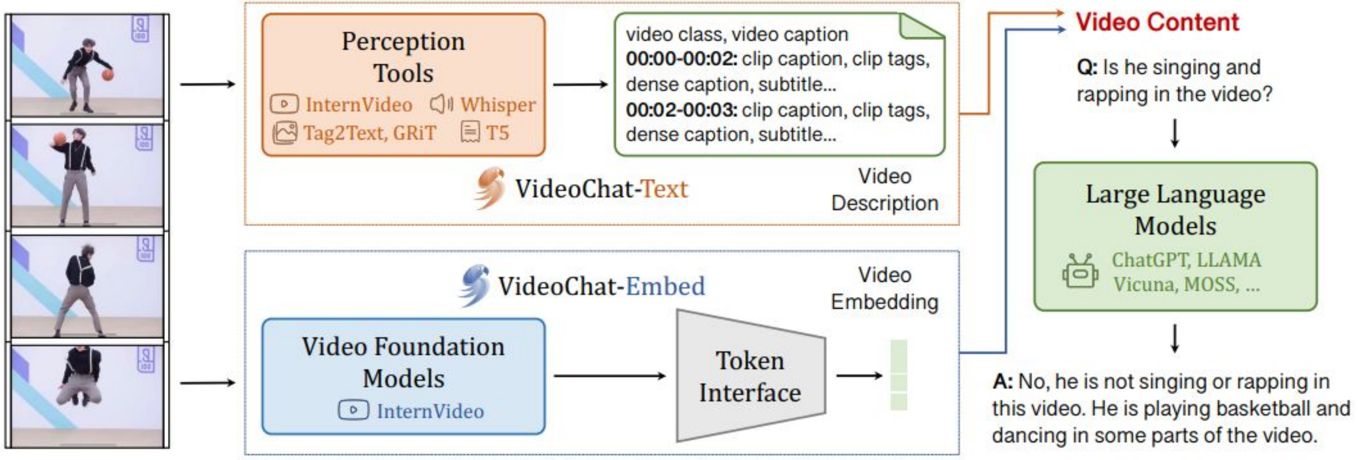


Figure 1: **Framework. VideoChat-Text** textualizes videos in stream. **VideoChat-Embed** encodes videos as embeddings. Both video content can be input in LLMs for multimodal understanding.

1 - A survey on multimodal large language models; 2 - VideoChat Paper

The Architecture of MLLMs: Pre-trained LLM

Model	Release date	Pre-train data scale	Parameter size (B)	Language support	Architecture
Flan-T5-XL/XXL [44]	Oct. 2022	–	3/11	En, Fr, De	Encoder decoder
LLaMA [45]	Feb. 2023	1.4T tokens	7/13/33/65	En	Causal decoder
Vicuna [46]	March 2023	1.4T tokens	7/13/33	En	Causal decoder
LLaMA-2 [47]	July 2023	2T tokens	7/13/70	En	Causal decoder
Qwen [48]	Sept. 2023	3T tokens	1.8/7/14/72	En, Zh	Causal decoder
LLaMA-3 [49]	April 2024	15T tokens	8/70/405	En, Fr, De, etc.	Causal decoder

Scaling up in **parameters** usually leads to better performance

Autoregressive generation

1 - A survey on multimodal large language models;

MLLMs: Training Strategy & Data

A Multimodal Large Language Model undergoes **three stages** of training, with each phase requiring different types of data and fulfilling different objectives.

1 Pre-training

Aims to align different modalities and learn multimodal world knowledge

Data: Large-scale dataset of image-text pairs

2 Instruction Tuning

Aims to teach models to better understand the instructions from users and fulfill the demanded tasks.
Integration of safety

Data: Task-specific datasets

3 Alignment Tuning

Aims to align the model with human preferences through techniques like RLHF and DPO

Data: Feedback for model responses

1 - A survey on multimodal large language models;

MLLMs: Evaluation Methods

After completing training, we need to ensure the model's real-world applicability across multiple tasks.

Evaluation Metrics

- Text Generation
- Vision-Language Understanding
- Audio Understanding
- Multimodal Coherence

Benchmark Datasets

- Text & Vision Tasks
- Video & Multimodal Tasks
- Audio & Speech

1 - A survey on multimodal large language models; 2 - A Survey on Evaluation of Multimodal Large Language Models

MLLMs: Evaluation Methods

Another thing to consider while evaluating MLLMs is how we are going to evaluate the model in **general tasks** and **specific tasks**

General Tasks

Multimodal Recognition, Perception and Reasoning

Trustworthiness: Hallucinations, Bias, Safety, Ethics

Specific Tasks

Socioeconomic: Cultural Analysis

Natural Science & Engineering: Mathematics, Biology, Code

Medical Tasks

1 - A Survey on Evaluation of Multimodal Large Language Models

General Tasks

RNAL OF LATEX CLASS FILES,AUGUST 2024 4

TABLE 1: Summary of the general evaluation tasks.

Tasks	Task Description	Related Benchmarks
Multi-modal Recognition		
Concept recognition	Recognizing visual concepts, e.g., objects, instances and scenes.	MMBench [21], MM-Vet [22], Seed-Bench [1], MME [23], MMStar [24], LLaVA-Bench [25], Open-VQA [26], MDVP-Bench [27], P²GB [28], EQBEN [29], MUIRBENCH [30], TouchStone [31], mPlug-Owl [32], MMIU [33], LogicVista [34], CODIS [35]
Attribute recognition	Recognizing visual subject's attributes e.g., style, quality, mood, quantity, material, and human's profession.	MMBench [21], MM-Vet [22], Seed-Bench [1], V*Bench [36], MMVP [37], CV-Bench [38], Visual CoT [39], EQBEN [29], SPEC [40], VL-Checklist [41], ARO [42], MUIRBENCH [30], COMPBENCH [43], MME [23], Open-VQA [26], TouchStone [31], ImplicitAVE [44],

Tasks	Task Description	Related Benchmarks
Trustworthiness		
Robustness	The capability of MLLMs to maintain performance under various conditions, including adversarial inputs or noisy environments.	CHEF [87], MAD-Bench [88], MMR [89], MM-SpuBench [90], BenchLMM [91], Multi-Trust [92]
Hallucination	The tendency of MLLMs to generate information that is incorrect, irrelevant, or fabricated.	POPE [93], UNIHD [94], VideoHallucer [95], CAP2QA [96], CHEF [87], GAVIE [97], HaELM [98], M-HalDetect [99], Bingo [100], Hallusion-Bench [101], AMBER [102], MM-SAP [103], VHTest [104], CorrelationQA [105],
Ethic	The adherence of MLLMs to ethical guidelines, ensuring outputs align with moral and societal values.	Multi-Trust [92]
Bias	The presence and extent of unfair biases in the MLLM's predictions, which could lead to discrimination or skewed results.	Multi-Trust [92], RTVLM [106]
Safety	The potential risks posed by the MLLM, such as generating harmful content, promoting dangerous behavior, or being misused.	MM-SafetyBench [107], MMUBench [108], Jailbreakv-28k [109], Shield [110], RTVLM [106], Multi-Trust [92],

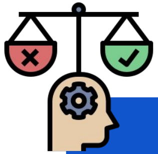
Tasks	Task Description	Related Benchmarks
		tionQA [105],
Ethic	The adherence of MLLMs to ethical guidelines, ensuring outputs align with moral and societal values.	Multi-Trust [92]
Bias	The presence and extent of unfair biases in the MLLM's predictions, which could lead to discrimination or skewed results.	Multi-Trust [92], RTVLM [106]
Safety	The potential risks posed by the MLLM, such as generating harmful content, promoting dangerous behavior, or being misused.	MM-SafetyBench [107], MMUBench [108], Jailbreakv-28k [109], Shield [110], RTVLM [106], Multi-Trust [92],

1 - A Survey on Evaluation of Multimodal Large Language Models

Trustworthiness: Hallucinations, Bias & Safety

Hallucinations

Generating false or misleading information



Bias

Reinforcing stereotypes and/or having skewed representation of cultures/demographics

Safety

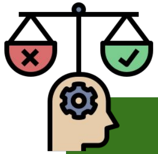
Generating harmful/unethical content, exposing private data, vulnerability to adversarial attacks

1 - A Survey on Evaluation of Multimodal Large Language Models

Trustworthiness: Hallucinations, Bias & Safety

Hallucinations

Confidence scoring and fact-checking with retrieval based methods



Bias

Diverse and balanced training data and implementation of bias-sensitive loss functions

Safety

Adversarial training and implementation of NSFW filters

1 - A Survey on Evaluation of Multimodal Large Language Models; 2 - Hallucination of Multimodal Large Language Models: A Survey

Tasks	Tasks Description	Related Benchmarks
Socioeconomic		
Cultural Analysis	The capability of MLLMs in understanding cultural norms, expressions, and practices across different societies.	CVQA [111]
Societal Analysis	The capability of MLLMs to comprehend and analyze societal issues, trends, and dynamics	VizWiz [112], MM-Soc [113], TransportationGames [114]

Other Applications		
3D Point Cloud	Interpret and process 3D spatial data for applications like robotics or autonomous driving.	ScanQA [141], LAMM [142], M3DBench [143], SpatialRGPT [62]
Video	The MLLMs' ability to understand, summarize, and reason about video content.	VideoHallucer [95], MMBench-Video [144], SOK-Bench [145], MVBench [146]
Remote Sensing	Process and analyze satellite or aerial images for environmental monitoring, agriculture, and more.	HighDAN [147], RSGPT [148], MDAS [149]
Audio	The ability of MLLMs in understanding audio, like speech recognition, audio event detection, and sound classification.	AIRBench [150], Dynamic-superb [151], MuChoMusic [152]

Other Applications		
3D Point Cloud	Interpret and process 3D spatial data for applications like robotics or autonomous driving.	ScanQA [141], LAMM [142], M3DBench [143], SpatialRGPT [62]
Video	The MLLMs' ability to understand, summarize, and reason about video content.	VideoHallucer [95], MMBench-Video [144], SOK-Bench [145], MVBench [146]
Remote Sensing	Process and analyze satellite or aerial images for environmental monitoring, agriculture, and more.	HighDAN [147], RSGPT [148], MDAS [149]
Audio	The ability of MLLMs in understanding audio, like speech recognition, audio event detection, and sound classification.	AIRBench [150], Dynamic-superb [151], MuChoMusic [152]

1 - A Survey on Evaluation of Multimodal Large Language Models

Category	Benchmark	Llama 3 8B	Gemma 2 9B	Mistral 7B	Llama 3 70B	Mixtral 8x22B	GPT 3.5 Turbo	Llama 3 405B	Nemotron 4 340B	GPT-4 (0125)	GPT-4o	Claude 3.5 Sonnet
General	MMLU (5-shot)	69.4	**72.3**	61.1	**83.6**	76.9	70.7	87.3	82.6	85.1	89.1	**89.9**
	MMLU (0-shot, CoT)	**73.0**	72.3△	60.5	**86.0**	79.9	69.8	88.6	78.7◁	85.4	**88.7**	88.3
	MMLU-Pro (5-shot, CoT)	**48.3**	–	36.9	**66.4**	56.3	49.2	73.3	62.7	64.8	74.0	**77.0**
	IFEval	**80.4**	73.6	57.6	**87.5**	72.7	69.9	**88.6**	85.1	84.3	85.6	88.0
Code	HumanEval (0-shot)	**72.6**	54.3	40.2	**80.5**	75.6	68.0	89.0	73.2	86.6	90.2	**92.0**
	MBPP EvalPlus (0-shot)	**72.8**	71.7	49.5	**86.0**	78.6	82.0	88.6	72.8	83.6	87.8	**90.5**
Math	GSM8K (8-shot, CoT)	**84.5**	76.7	53.2	**95.1**	88.2	81.6	**96.8**	92.3◇	94.2	96.1	96.4◇
	MATH (0-shot, CoT)	**51.9**	44.3	13.0	**68.0**	54.1	43.1	73.8	41.1	64.5	**76.6**	71.1
Reasoning	ARC Challenge (0-shot)	83.4	**87.6**	74.2	**94.8**	88.7	83.7	**96.9**	94.6	96.4	96.7	96.7
	GPQA (0-shot, CoT)	32.8	–	28.8	**46.7**	33.3	30.8	51.1	–	41.4	53.6	**59.4**
Tool use	BFCL	**76.1**	–	60.4	84.8	–	**85.9**	88.5	86.5	88.3	80.5	**90.2**
	Nexus	**38.5**	30.0	24.7	**56.7**	48.5	37.2	**58.7**	–	50.3	56.1	45.7
Long context	ZeroSCROLLS/QuALITY	81.0	–	–	90.5	–	–	**95.2**	–	**95.2**	90.5	90.5
	InfiniteBench/En.MC	65.1	–	–	78.2	–	–	**83.4**	–	72.1	82.5	–
	NIH/Multi-needle	98.8	–	–	97.5	–	–	98.1	–	**100.0**	**100.0**	90.8
Multilingual	MGSM (0-shot, CoT)	**68.9**	53.2	29.9	**86.9**	71.1	51.4	**91.6**	–	85.9	90.5	**91.6**

Table 2 Performance of finetuned Llama 3 models on key benchmark evaluations.

1 - The Llama 3 Herd of Models

Exam	Llama 3 8B	Llama 3 70B	Llama 3 405B	GPT-3.5 Turbo	Nemotron 4 340B	GPT-4o	Claude 3.5 Sonnet
LSAT	53.9 ±4.9	74.2 ±4.3	**81.1 ±3.8**	54.3 ±4.9	73.7 ±4.3	77.4 ±4.1	80.0 ±3.9
SAT Reading	57.4 ±4.2	71.4 ±3.9	74.8 ±3.7	61.3 ±4.2	–	82.1 ±3.3	**85.1 ±3.1**
SAT Math	73.3 ±4.6	91.9 ±2.8	94.9 ±2.3	77.3 ±4.4	–	95.5 ±2.2	**95.8 ±2.1**
GMAT Quant.	56.0 ±19.5	84.0 ±14.4	**96.0 ±7.7**	36.0 ±18.8	76.0 ±16.7	92.0 ±10.6	92.0 ±10.6
GMAT Verbal	65.7 ±11.4	85.1 ±8.5	86.6 ±8.2	65.7 ±11.4	91.0 ±6.8	**95.5 ±5.0**	92.5 ±6.3
GRE Physics	48.0 ±11.3	74.7 ±9.8	80.0 ±9.1	50.7 ±11.3	–	89.3 ±7.0	**90.7 ±6.6**
AP Art History	75.6 ±12.6	84.4 ±10.6	**86.7 ±9.9**	68.9 ±13.5	71.1 ±13.2	80.0 ±11.7	77.8 ±12.1
AP Biology	91.7 ±11.1	**100.0 ±0.0**	**100.0 ±0.0**	91.7 ±11.1	95.8 ±8.0	**100.0 ±0.0**	**100.0 ±0.0**
AP Calculus	57.1 ±16.4	54.3 ±16.5	88.6 ±10.5	62.9 ±16.0	68.6 ±15.4	**91.4 ±9.3**	88.6 ±10.5
AP Chemistry	59.4 ±17.0	**96.9 ±6.0**	90.6 ±10.1	62.5 ±16.8	68.8 ±16.1	93.8 ±8.4	**96.9 ±6.0**
AP English Lang.	69.8 ±12.4	90.6 ±7.9	94.3 ±6.2	77.4 ±11.3	88.7 ±8.5	**98.1 ±3.7**	90.6 ±7.9
AP English Lit.	59.3 ±13.1	79.6 ±10.7	83.3 ±9.9	53.7 ±13.3	**88.9 ±8.4**	**88.9 ±8.4**	85.2 ±9.5
AP Env. Sci.	73.9 ±12.7	89.1 ±9.0	**93.5 ±7.1**	73.9 ±12.7	73.9 ±12.7	89.1 ±9.0	84.8 ±10.4
AP Macro Eco.	72.4 ±11.5	**98.3 ±3.3**	**98.3 ±3.3**	67.2 ±12.1	91.4 ±7.2	96.5 ±4.7	94.8 ±5.7
AP Micro Eco.	70.8 ±12.9	91.7 ±7.8	93.8 ±6.8	64.6 ±13.5	89.6 ±8.6	**97.9 ±4.0**	**97.9 ±4.0**
AP Physics	57.1 ±25.9	78.6 ±21.5	**92.9 ±13.5**	35.7 ±25.1	71.4 ±23.7	71.4 ±23.7	78.6 ±21.5
AP Psychology	94.8 ±4.4	**100.0 ±0.0**	**100.0 ±0.0**	94.8 ±4.4	**100.0 ±0.0**	**100.0 ±0.0**	**100.0 ±0.0**
AP Statistics	66.7 ±17.8	59.3 ±18.5	85.2 ±13.4	48.1 ±18.8	77.8 ±15.7	92.6 ±9.9	**96.3 ±7.1**
AP US Gov.	90.2 ±9.1	97.6 ±4.7	97.6 ±4.7	78.0 ±12.7	78.0 ±12.7	**100.0 ±0.0**	**100.0 ±0.0**
AP US History	78.0 ±12.7	**97.6 ±4.7**	**97.6 ±4.7**	85.4 ±10.8	70.7 ±13.9	95.1 ±6.6	95.1 ±6.6
AP World History	94.1 ±7.9	**100.0 ±0.0**	**100.0 ±0.0**	88.2 ±10.8	85.3 ±11.9	**100.0 ±0.0**	97.1 ±5.7
AP Average	74.1 ±3.4	87.9 ±2.5	**93.5 ±1.9**	70.2 ±3.5	81.3 ±3.0	93.0 ±2.0	92.2 ±2.1
GRE Quant.	152.0	158.0	162.0	155.0	161.0	**166.0**	164.0
GRE Verbal	149.0	166.0	166.0	154.0	162.0	**167.0**	**167.0**

Table 17 Performance of Llama 3 models and GPT-4o on a variety of proficiency exams including LSAT, SAT, GMAT, and AP, and GRE tests. For GRE exams, we report normalized score; for all others, we report accuracy. For the bottom two rows corresponding to GRE Quant. and GRE Verbal, we report the scaled scores out of 170.

1 - The Llama 3 Herd of Models

MLLMs: Evaluation Metrics

There are three main approaches to evaluation “metrics”:

1. **Metric-based evaluation**

   Automated metrics measure performance based on benchmarks

2. **Human-based evaluation**

   Experts or crowd workers assess model responses

3. **GPT4-based evaluation**

   Large models assess other models

1 - A Survey on Evaluation of Multimodal Large Language Models

MLLMs: Evaluation Metrics

Metric-based evaluation uses numerical, objective metrics for comparison.

Works well for classification, translation, and retrieval tasks

These include the **classification "classics"**: Accuracy, Precision, Recall, F1-Score…

Machine Translation, Text Summarization, and Image Captioning

Bilingual Evaluation Understudy

$$\mathrm{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

1 - A Survey on Evaluation of Multimodal Large Language Models

Metric-based

MLLMs: Evaluation Metrics

Metric-based evaluation uses numerical, objective metrics for comparison.

Works well for classification, translation, and retrieval tasks

These include the **classification "classics"**: Accuracy, Precision, Recall, F1-Score…

Machine Translation, Text Summarization, and Image Captioning

Consensus-based Image Description Evaluation

$$\mathrm{CIDEr} = \frac{1}{m} \sum_{i=1}^{m} w_i \cdot \log \left(\frac{\mathrm{count}(i)}{\mathrm{frequency}(i)} \right)$$

1 - A Survey on Evaluation of Multimodal Large Language Models

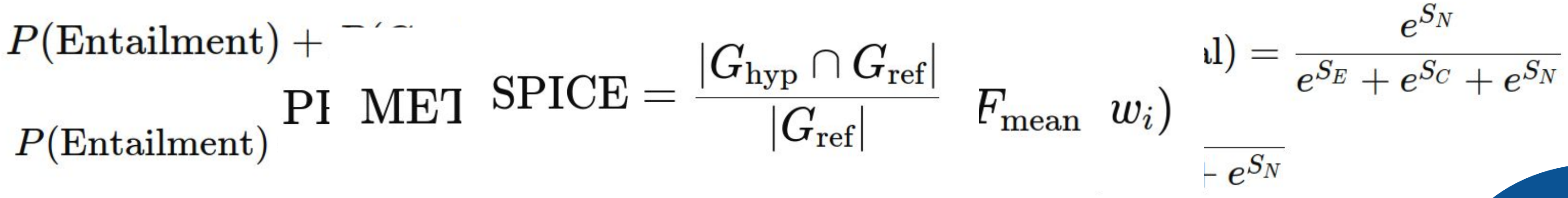
MLLMs: Evaluation Metrics

Metric-based evaluation uses numerical, objective metrics for comparison.

Works well for classification, translation, and retrieval tasks

These include the **classification "classics"**: Accuracy, Precision, Recall, F1-Score…

Machine Translation, Text Summarization, and Image Captioning



1 - A Survey on Evaluation of Multimodal Large Language Models

MLLMs: Evaluation Metrics

Human-based evaluation is used to asses models through direct human judgment.

It helps evaluate aspects that automated scores miss!

Likert Scale: Evaluating some model characteristic from 1 to 5

A/B Testing: Comparing two versions of the model

Expert Review: The model is evaluated by domain experts

Very Time Consuming! **Very Expensive!** **Potential biased evaluation**

1 - A Survey on Evaluation of Multimodal Large Language Models

MLLMs: Evaluation Metrics

GPT-based evaluation is based on using large models (like GPT-4) to assess the quality of model outputs

Good alternative for automated assessment that is consistent across a variety of tasks. It is also reference-free!

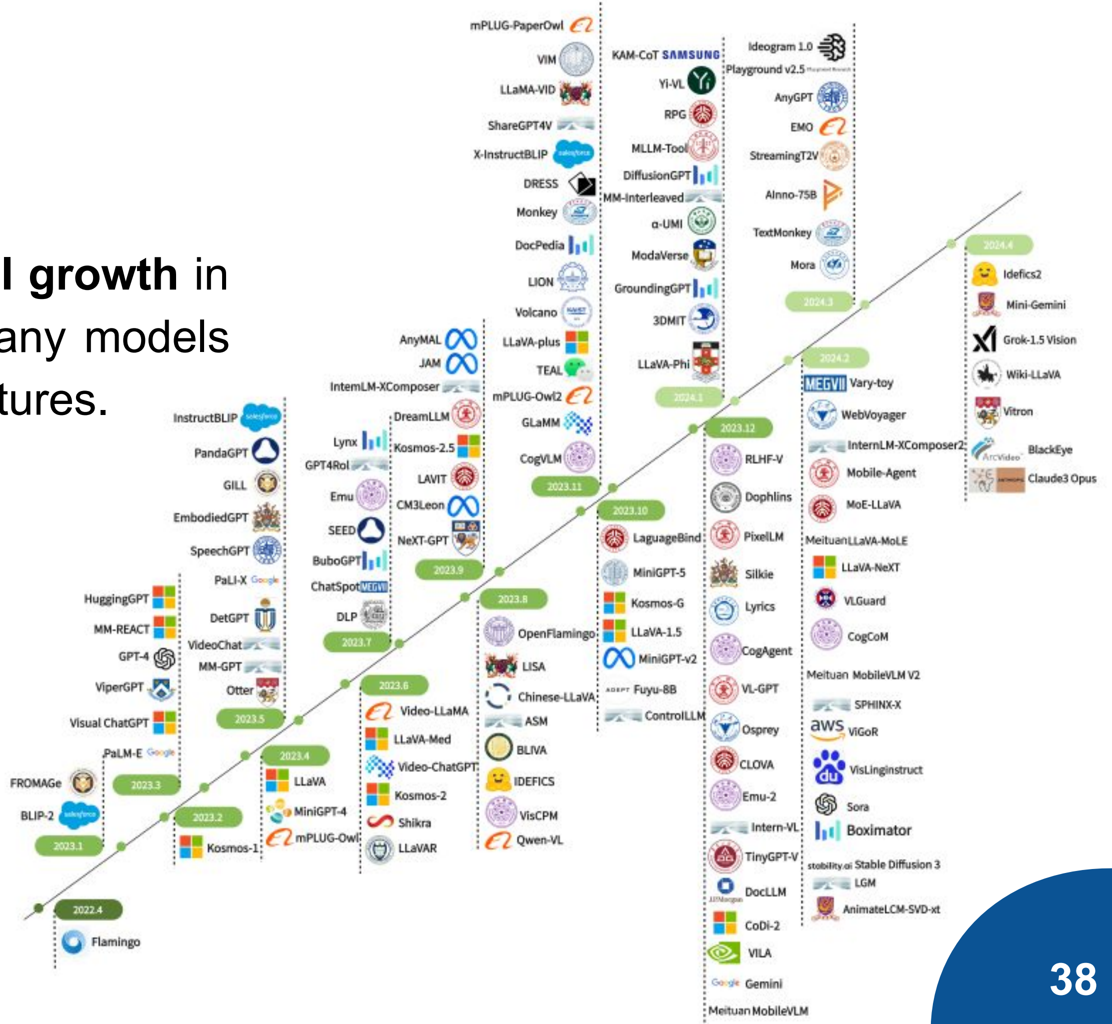
Cost-effective!	Cannot fully replace human evaluation
Quick feedback	May inherent model biases
Systematic Evaluation	Technical Content

1 - A Survey on Evaluation of Multimodal Large Language Models

A World of MLLMs

The field has seen **exponential growth** in the number of MLLMs, with many models building upon previous architectures.

From task-specific to generalized tasks



1 - A Comprehensive Review of Multimodal Large Language Models

MLLMs: Evaluation Metrics

GPT-based evaluation is based on using large models (like GPT-4) to assess the quality of model outputs

Good alternative for automated assessment that is consistent across a variety of tasks. It is also reference-free!

1 - A Survey on Evaluation of Multimodal Large Language Models

Exercise Time