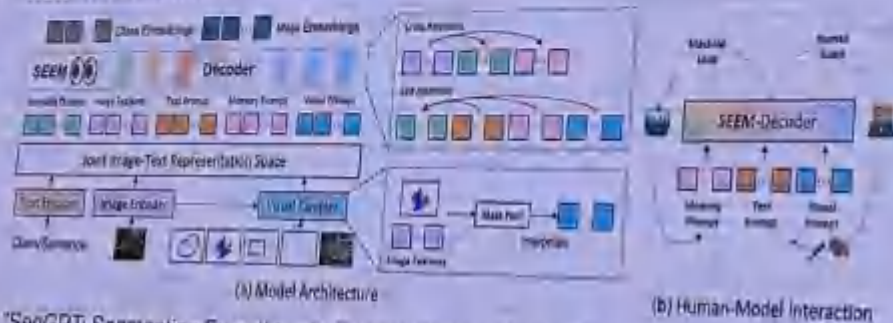


# Image Segmentation: More Explorations



- SegGPT:
  - Formulate segmentation as in-context coloring
  - Inherits from Painter, while focuses on segmentation

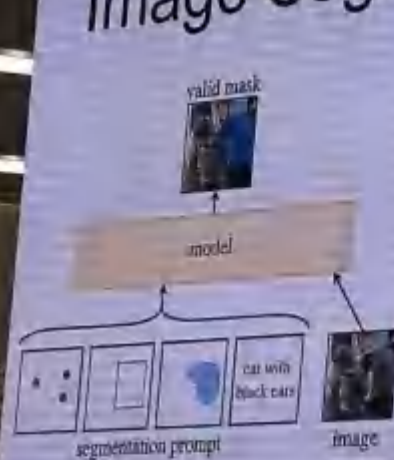


- SEEM:
  - More prompt formats, e.g., referred regions from other images
  - Project prompts into a joint visual-semantic space

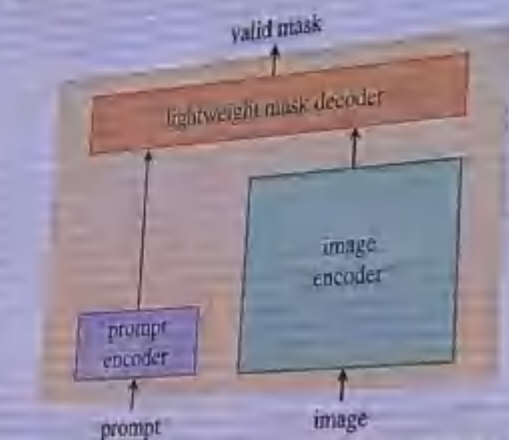
"SegGPT: Segmenting Everything In Context", Wang et al. ICCV 2023

"Segment Everything Everywhere All at Once", Zou et al. NeurIPS 2023

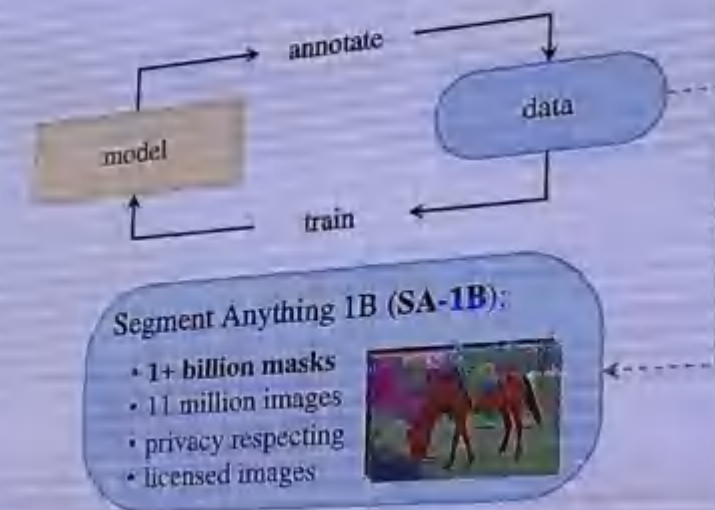
# Image Segmentation: SAM



(a) Task: promptable segmentation



(b) Model: Segment Anything Model (SAM)



(c) Data: data engine (top) & dataset (bottom)

## • Previous:

- Single task, single domain, closed set, single type of prompt, etc.

## • SAM:

- Unified and generic interface for segmentation: mask, points, box, text
- Data engine scaling law: 1+ billion masks, 11 million images
- Strong generalization ability: even aerial images, synthetic images, and medical images

"Segment Anything". Kirillov et al. ICCV 2023



# Representation Learning: DINOv2



| Model             | Arch.                   | Data     | Self-sup. | ViVi |      | Image |      |
|-------------------|-------------------------|----------|-----------|------|------|-------|------|
|                   |                         |          |           | val  | test | ViVi  | ViT  |
| Weakly supervised |                         |          |           |      |      |       |      |
| CLIP              | ViT-L/14                | ViT-400M | ✓         | 70.8 | 68.1 | 68.1  | 75.1 |
| CLIP              | ViT-L/14 <sub>msc</sub> | ViT-400M | ✓         | 60.5 | 60.0 | 60.0  | 70.0 |
| OpenCLIP          | ViT-M/14                | LAION-2B | ✓         | 63.6 | 60.7 | 60.7  | 72.0 |
| OpenCLIP          | ViT-L/14                | LAION-2B | ✓         | 66.7 | 64.1 | 64.1  | 75.5 |
| OpenCLIP          | ViT-L/14                | LAION-2B | ✓         | 68.2 | 66.0 | 66.0  | 77.2 |
| OpenCLIP          | ViT-L/14                | custom*  | ✓         | 65.5 | 65.4 | 66.7  | 77.6 |

| Self-supervised |          |        |   |      |      |      |      |
|-----------------|----------|--------|---|------|------|------|------|
| ALL             | ViT-L/14 | OpenAI | X | 80.4 | 76.0 | 83.8 | 88.8 |
| CLIP            | ViT-L/14 | OpenAI | X | 78.8 | 74.7 | 83.3 | 88.2 |
| OpenAI          | ViT-L/14 | OpenAI | X | 78.8 | 74.7 | 83.3 | 88.2 |
| OpenAI          | ViT-L/14 | OpenAI | X | 78.8 | 74.7 | 83.3 | 88.2 |
| OpenAI          | ViT-L/14 | OpenAI | X | 78.8 | 74.7 | 83.3 | 88.2 |
| OpenAI          | ViT-L/14 | OpenAI | X | 78.8 | 74.7 | 83.3 | 88.2 |
| OpenAI          | ViT-L/14 | OpenAI | X | 78.8 | 74.7 | 83.3 | 88.2 |
| OpenAI          | ViT-L/14 | OpenAI | X | 78.8 | 74.7 | 83.3 | 88.2 |
| OpenAI          | ViT-L/14 | OpenAI | X | 78.8 | 74.7 | 83.3 | 88.2 |
| OpenAI          | ViT-L/14 | OpenAI | X | 78.8 | 74.7 | 83.3 | 88.2 |

- Data engine: Image deduplication and retrieval, resulting in high-quality LVD-142M dataset

- Performance: Powerful and generalizable ever feature extractor, even comparable to vision language pre-trained models

"DINOv2: Learning Robust Visual Features without Supervision", Oquab et al. TMLR 2024

# Representation Learning: DINOv1 vs. DINOv2



DINOv1 [Caron et al. ICCV 2021]

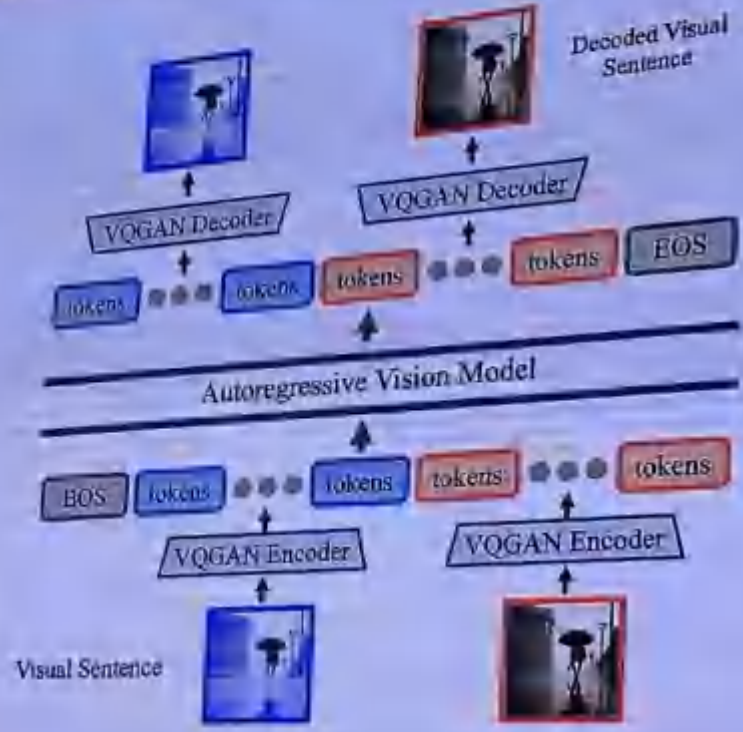


DINOv2 [Oquab et al. TMLR 2024]



# Vision Paradigm: LVM

## Visual Sentences



- Visual sentences: Novel sequential modeling without linguistic data
- Data engine: Unified Vision Dataset v1 (UVDv1) with 1.64 billion images, including unlabelled/labelled images/videos, and 3D synthetic objects

"Sequential Modeling Enables Scalable Learning for Large Vision Models", Bai et al. CVPR 2024

# LVM: Large Vision Models

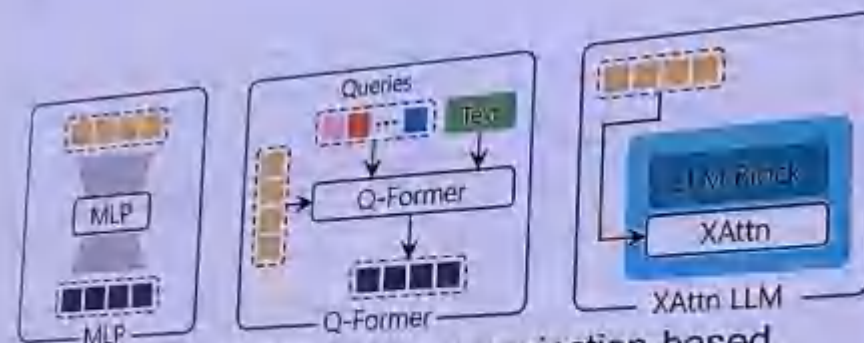
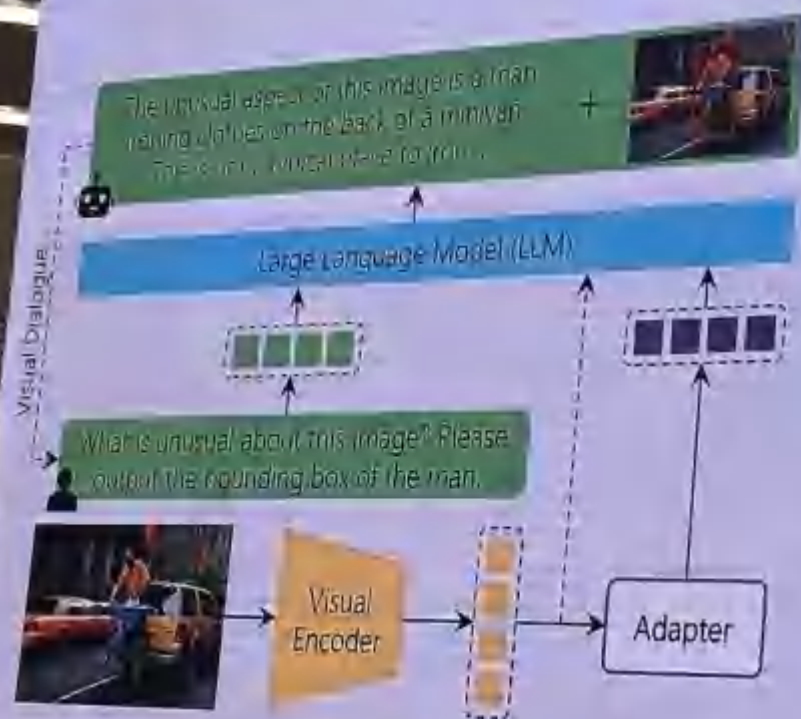
Prompts

Generated





# LVLM: General Architecture



Three types of connectors: projection-based, query-based, and fusion-based connectors

TABLE 1: A summary of commonly used image encoders.

| Variants                  | Pretraining Corpus | Resolution | Samples (B) | Parameter Size (M) |
|---------------------------|--------------------|------------|-------------|--------------------|
| DynaCLIP-CareNetFL (40)   | LAION-2B           | 320        | 29          | 197.4              |
| CLIP-ViT-H/14 (13)        | OpenAI's ViT       | 224/336    | 15          | 304.0              |
| BVA-CLIP-ViT-G/14 (40)    | LAION-2B/COYO-200M | 224        | 11          | 1000.0             |
| OpenCLIP-ViT-G/14 (40)    | LAION-2B           | 224        | 34          | 1012.7             |
| OpenCLIP-ViT-bigG/14 (40) | LAION-2B           | 224        | 34          | 1844.9             |

Commonly used image encoders

TABLE 2: A summary of commonly used open-sourced LLMs. en, zh, fr, and de stand for English, Chinese, French, and German, respectively.

| Model                   | Release Date | Pretrain Data Scale | Parameter Size (B) | Language Support | Architecture    |
|-------------------------|--------------|---------------------|--------------------|------------------|-----------------|
| MiniGPT-4 (100,000,000) | Oct-2023     | -                   | 3/ 11              | en, fr, de       | Encoder-Decoder |
| LLaMA (3)               | Feb-2023     | 1.4T tokens         | 7/ 13/ 33/ 65      | en               | Causal Decoder  |
| Yi (6)                  | Mar-2023     | 1.4T tokens         | 7/ 13/ 33          | en               | Causal Decoder  |
| LLaMA-2 (4)             | Jul-2023     | 2T tokens           | 7/ 13/ 70          | en               | Causal Decoder  |
| Qwen (5)                | Sep-2023     | 3T tokens           | 1.8 / 7 / 14 / 72  | en, zh           | Causal Decoder  |

Commonly used open-sourced LLMs

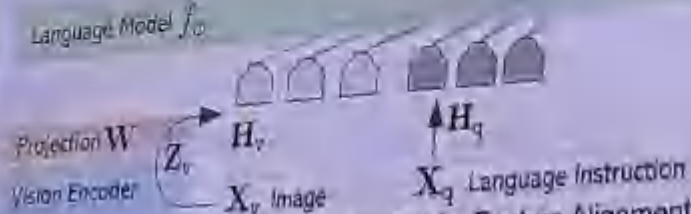
"The (R)Evolution of Multimodal Large Language Models: A Survey", Caffagni et al. arXiv:2312.00785

"A Survey on Multimodal Large Language Models", Yin et al. arXiv:2306.13549

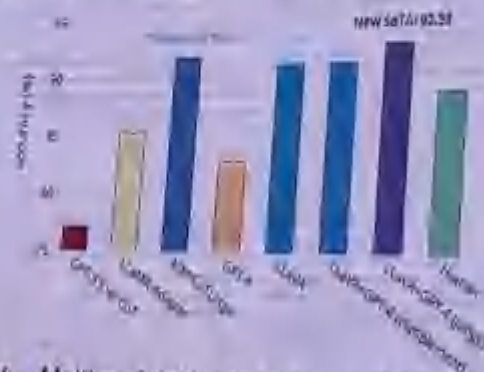
# LLaVA: Large Language and Vision Assistant

Language Response  $X_p$

Language Model  $f_{\phi}$



Two-stage instruction-tuning: 1. Pre-training for Feature Alignment: Only the projection matrix is updated, based on a subset of CC3M 2. Fine-tuning End-to-End: Both the projection matrix and LLM are updated



Two evaluation benchmarks: Multimodal Chatbot, ScienceQA

"Visual Instruction Tuning", Liu et al. NeurIPS 2023

Visual input example, Extreme ironings



User: LLaVA

What is unusual about this image?  
The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.

User: LLaVA

(Start a new conversation, and clear the history)  
What's happening in the scene?  
The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention due to his unconventional choice of ironing his clothes on top of a moving car. The city street around him is bustling with activity, adding to the unique nature of the scene.

User: GPT-4 (30)

What is unusual about this image?  
The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

User: BLIP-2

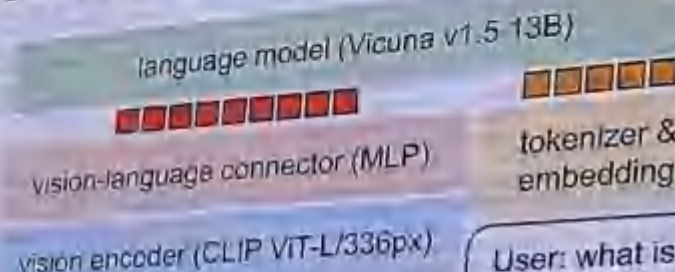
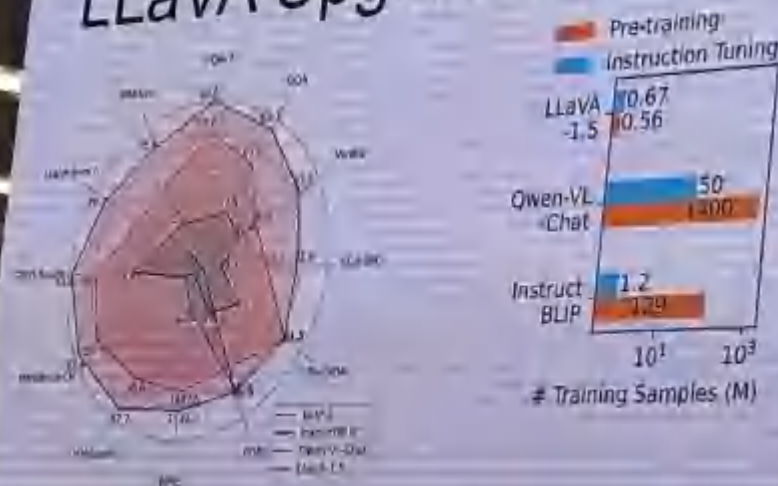
What is unusual about this image?  
a man is sitting on the back of a yellow cab

User: OpenFlamingo

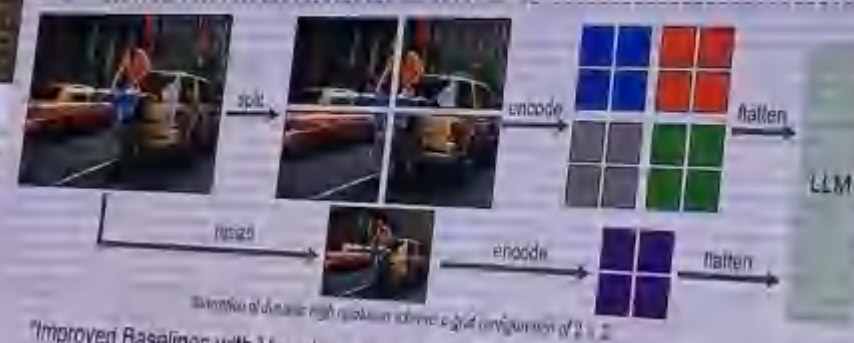
What is unusual about this image?  
The man is drying his clothes on the hood of his car.



# LLaVA Upgrade: LLaVA-1.5 & LLaVA-1.6



User: what is unusual about this image?



## LLaVA-1.5:

- Larger resolution vision encoder with more details
- MLP vision-language connector: change linear projection to two-layer MLP
- Academic task oriented data: include open-knowledge VQA (OKVQA, A-OKVQA) and OCR (OCR-VQA, TextCaps), and region-level VQA (Visual Genome, RefCOCO)
- Response formatting prompts: adjust the output format according to the user's instructions

## LLaVA-1.6:

- Dynamic high resolution input:  $\{2 \times 2, 1 \times \{2, 3, 4\}, \{2, 3, 4\} \times 1\}$
- Data Mixture: high-quality user instruct data (LAION-GPT-V, ShareGPT-4V), multimodal document/chart data (DocVQA, SynDog-EN, ChartQA, DVQA, AI2D)
- Scaling LLM backbone: up to 34B

"Improved Baselines with Visual Instruction Tuning", Liu et al. arXiv:2310.03744  
<https://llava-vl.github.io/blog/2024-01-30-llava-next/>, Liu et al. Blog 2024

# Generalist LVLM Configurations

| Model                               | LLM              | Visual Encoder | V2L Adapter     | VInstr. Tuning | Main Tasks & Capabilities                                      |
|-------------------------------------|------------------|----------------|-----------------|----------------|--|
| BLIP-2 (Li et al., 2023)            | FlanT5-XL (11B)* | EVA ViT-g      | Q-Former        | ✓              | Visual Dialogue, VQA, Captioning, Retrieval                    |
| PRIDEv2 (Koh et al., 2023b)         | GPT-4 7B*        | CLIP ViT-L     | Linear          | ✓              | Visual Dialogue, Captioning, Retrieval                         |
| Kosmos-2 (Wang et al., 2023a)       | Magicoder-7.5B   | CLIP ViT-L     | Q-Former*       | ✓              | Visual Dialogue, VQA, Captioning                               |
| LLaMA-Adapter V2 (Gao et al., 2023) | LLaMA-7B*        | CLIP ViT-L     | Linear          | ✓              | VQA, Captioning  |
| OpenFlamingo (Ayoub et al., 2023)   | LLaMA-7B*        | CLIP ViT-L     | XAim LLM        | ✓              | VQA, Captioning  |
| FlanT5 (Joshi et al., 2022)         | LLaMA-7B*        | CLIP ViT-L     | XAim LLM        | ✓              | Visual Dialogue, VQA, Captioning                               |
| FLAN (Joshi et al., 2022)           | LLaMA-7B*        | CLIP ViT-L     | XAim LLM        | ✓              | Multilingual, VQA, Captioning, Retrieval                       |
| FLAN (Joshi et al., 2022)           | LLaMA-7B*        | CLIP ViT-L     | XAim LLM        | ✓              | Multilingual, VQA, Captioning                                  |
| LLaMA (Tou et al., 2023a)           | LLaMA-7B*        | CLIP ViT-L     | Linear          | ✓              | Visual Dialogue, VQA, Captioning                               |
| Mistral-7B (Jiang et al., 2023a)    | LLaMA-7B*        | EVA ViT-g      | Linear          | ✓              | VQA, Captioning  |
| OpenFLAN (Joshi et al., 2023)       | LLaMA-7B*        | CLIP ViT-L     | Q-Former*       | ✓              | Visual Dialogue, VQA   |
| Mistral-7B (Jiang et al., 2023)     | LLaMA-7B*        | EVA ViT-g      | Q-Former*       | ✓              | Visual Dialogue, VQA, Captioning                               |
| Mistral-7B (Jiang et al., 2023)     | LLaMA-7B*        | CLIP ViT-L     | XAim LLM        | ✓              | Visual Dialogue, VQA, Captioning                               |
| LLaMA (Tou et al., 2023)            | LLaMA-7B*        | CLIP ViT-L     | MLP             | ✓              | Visual Dialogue, VQA, Captioning                               |
| LLaMA (Tou et al., 2023)            | LLaMA-7B*        | CLIP ViT-L     | XAim LLM        | ✓              | VQA, Captioning  |
| LLaMA (Tou et al., 2023)            | LLaMA-7B*        | CLIP ViT-L     | Q-Former*       | ✓              | Visual Dialogue, VQA, Captioning, Referring, REC               |
| LLaMA (Tou et al., 2023)            | LLaMA-7B*        | CLIP ViT-L     | Linear          | ✓              | Visual Dialogue, VQA, Captioning, Referring, REC, GroundCap    |
| LLaMA (Tou et al., 2023)            | LLaMA-7B*        | CLIP ViT-L     | XAim LLM        | ✓              | Visual Dialogue, VQA, Captioning                               |
| LLaMA (Tou et al., 2023)            | LLaMA-7B*        | CLIP ViT-L     | MLP             | ✓              | Visual Dialogue, VQA, Captioning                               |
| LLaMA (Tou et al., 2023)            | LLaMA-7B*        | EVA ViT-g      | Q-Former+Linear | ✓              | Visual Dialogue, VQA, Captioning                               |
| LLaMA (Tou et al., 2023)            | LLaMA-7B*        | OpenCLIP ViT-H | XAim LLM        | ✓              | Visual Dialogue, VQA, Captioning                               |
| LLaMA (Tou et al., 2023)            | LLaMA-7B*        | OpenCLIP ViT-H | Q-Former*       | ✓              | Visual Dialogue, Multilingual, VQA, Captioning, REC            |
| LLaMA (Tou et al., 2023)            | LLaMA-7B*        | CLIP ViT-L     | Linear          | ✓              | Visual Dialogue, VQA, Captioning                               |
| LLaMA (Tou et al., 2023)            | LLaMA-7B*        | CLIP ViT-L     | Linear          | ✓              | Visual Dialogue, Captioning, Referring, REC, GroundCap         |
| LLaMA (Tou et al., 2023)            | LLaMA-7B*        | CLIP ViT-L     | MLP             | ✓              | Visual Dialogue, VQA, Captioning                               |
| LLaMA (Tou et al., 2023)            | LLaMA-7B*        | EVA ViT-g      | Linear          | ✓              | Visual Dialogue, VQA, Captioning, Referring, REC, GroundCap    |
| LLaMA (Tou et al., 2023)            | LLaMA-7B*        | CLIP ViT-L     | Linear          | ✓              | Visual Dialogue, VQA, Captioning, Referring, REC, GroundCap    |
| LLaMA (Tou et al., 2023)            | LLaMA-7B*        | EVA ViT-g      | MLP             | ✓              | Visual Dialogue, VQA, Captioning, REC                          |
| LLaMA (Tou et al., 2023)            | LLaMA-7B*        | EVA ViT-g      | Linear          | ✓              | Visual Dialogue, VQA, Captioning                               |
| LLaMA (Tou et al., 2023)            | LLaMA-7B*        | EVA ViT-g      | Q-Former+MLP    | ✓              | Visual Dialogue, VQA, Captioning, REC                          |
| LLaMA (Tou et al., 2023)            | LLaMA-7B*        | CLIP ViT-L     | Q-Former*       | ✓              | Visual Dialogue, VQA, Captioning                               |
| LLaMA (Tou et al., 2023)            | LLaMA-7B*        | Mistral        | Linear          | ✓              | Visual Dialogue, VQA, Captioning                               |
| LLaMA (Tou et al., 2023)            | LLaMA-7B*        | CLIP ViT-L     | ResNet blocks   | ✓              | Visual Dialogue, VQA, Captioning, Referring, REC, GroundCap    |
| LLaMA (Tou et al., 2023)            | LLaMA-7B*        | CLIP ViT-L     | Linear          | ✓              | Visual Dialogue, VQA, Captioning                               |
| LLaMA (Tou et al., 2023)            | LLaMA-7B*        | Mistral        | Linear          | ✓              | Visual Dialogue, Multilingual, VQA, Captioning, Referring, REC |

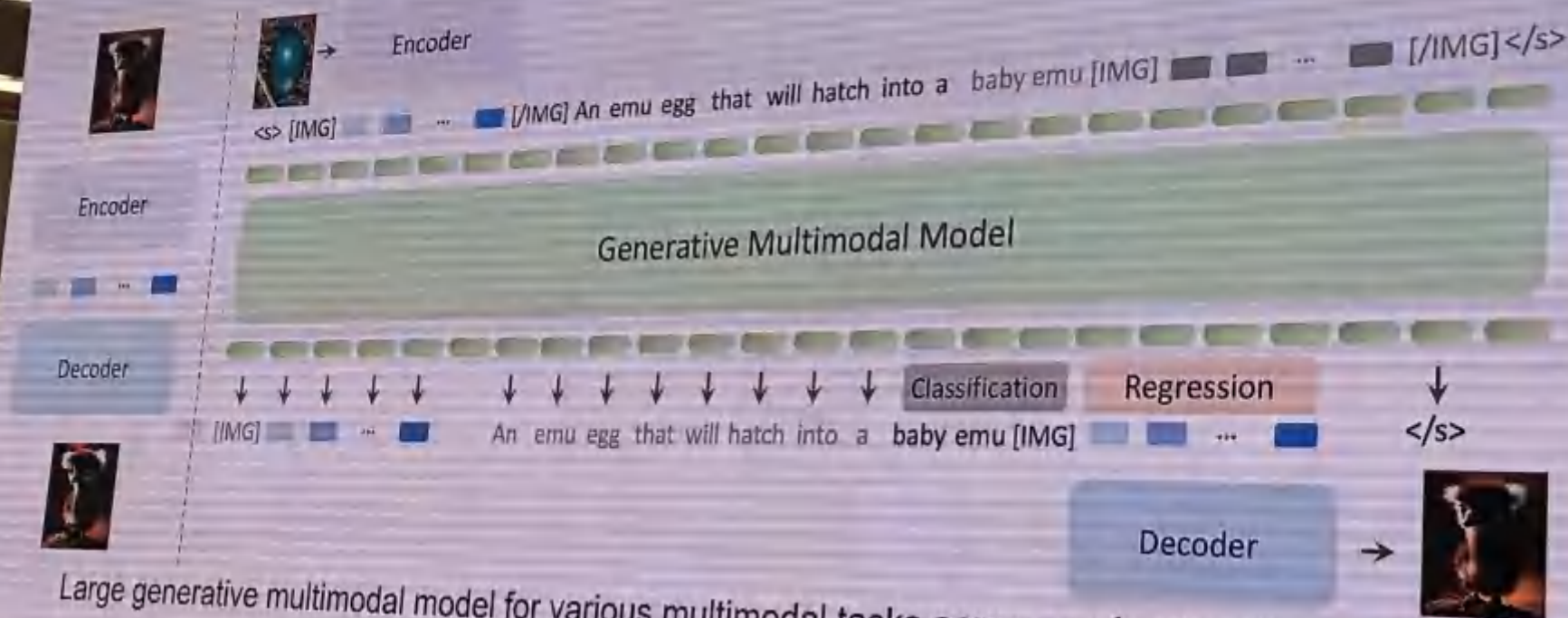
The (R)Evolution of Multimodal Large Language Models: A Survey". Caffagni et al. arXiv:2312.00785



# LVLM: Further Explorations

- High resolution input (from 448px to 4K):
  - Qwen-VL [Alibaba], SPHINX [SHAILab&CUHK&ShanghaiTech], Monkey [HUST], CogAgent [THU], Ureader [ECNU&DAMO&RUC], LLaVA-Next [Wisconsin&Microsoft], InfiMM-HD [CAS&UCAS&ByteDance], Fuyu [Adept], Otter-HD [NTU], LLaVA-HR [XMU&Pengcheng], LLaVA-UHD [THU&NUS&UCAS], Griffin v2 [CAS etc.], InternLM-XComposer2-4KHD [SHAILab etc.], Mini-Gemini [CUHK&SmartMore]
- Region/Pixel-level understanding:
  - BuboGPT [ByteDance], Vision-LLM [SHAILab etc.], Shikra [SenseTime&BUAA&SJTU], GPT4-ROI [HKU&SHAILab&Alibaba], PVIT [THU&CAS], Ferret [Columbia&Apple], Kosmos2 [Microsoft], DelGPT [HKUST&HKU&SJTU], PixelLM [BJTU&USTB&ByteDance], GLaMM [MBZUAI etc.], Groma [HKU&ByteDance], Griffin&Griffin v2 [CAS etc.], LISA & LISA++ [CUHK&SmartMore], RegionGPT [HKU&NVIDIA], PSALM [HUST&Microsoft], PixelLLM [Google&UCSD], Lenna [Meituan&BUAA]
- Lightweight model (about 3B):
  - TinyGPT-V [Notre Dame&Lehigh&SJTU], MobileVLM&MobileVLM-v2 [Meituan etc.], LLaVA-Phi [Midea&ECNU], MC-LLaVA, Imp-v1 [HDU], MiniCPM-V, Moondream, TinyLLaVA-v1 [BUAA&THU], Bunny [BAAI etc.], Mipha [Midea&ECNU&SHU], VL-Mamba [Adelaide&CAS&UCAS]
- Understanding and generation:
  - SEED-X [Tencent], Emu2 [BAAI&THU&PKU], VL-GPT [XJTU&Tencent&HKU], LaVIT [PKU&Kuaishou], DreamLLM [XJTU et al.]
- More modalities:
  - Unified-IO 2 [AI2&UIUC&UW], NEXT-GPT [NUS]

# Understanding and Generation: Emu2

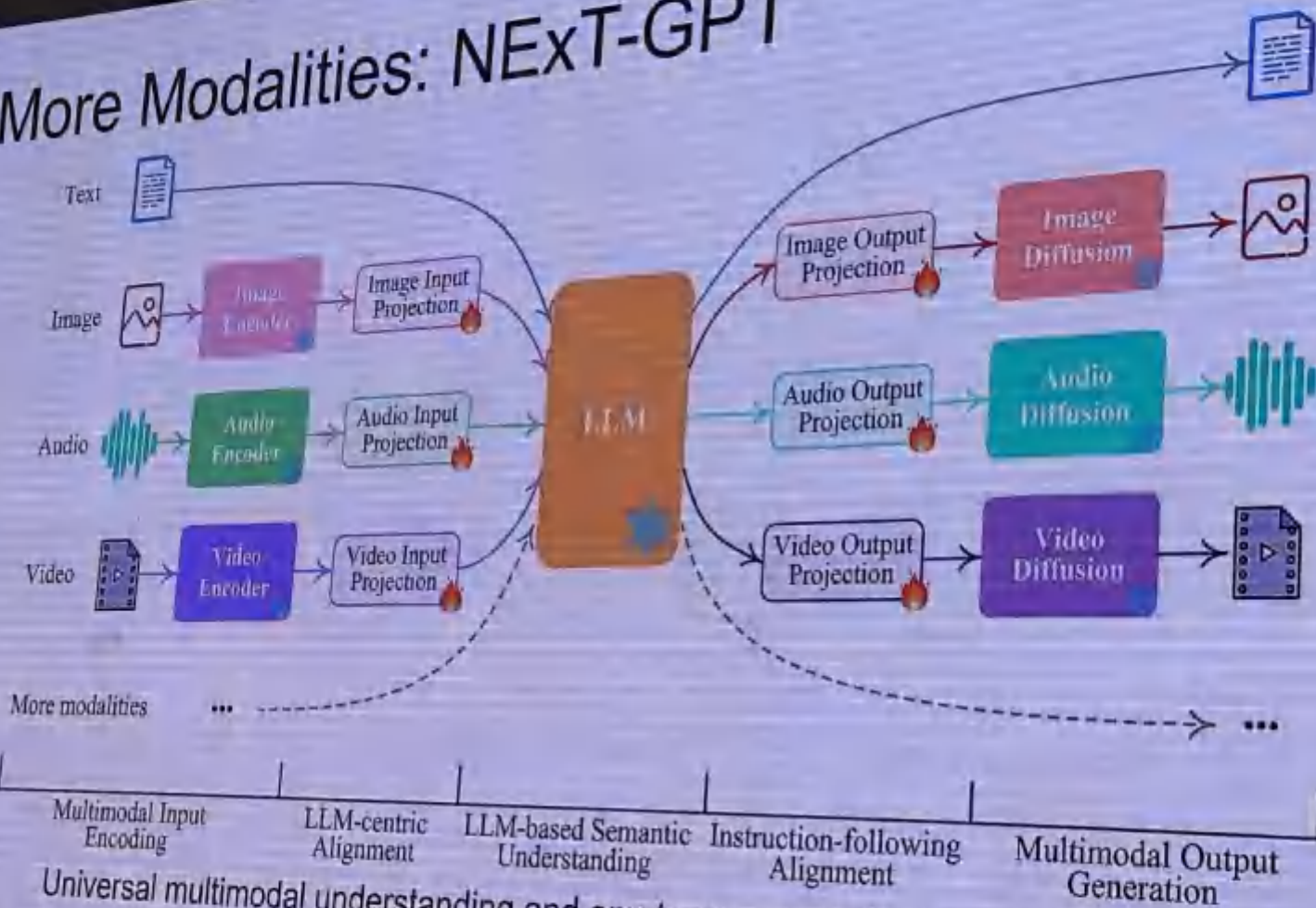


Large generative multimodal model for various multimodal tasks across understanding and generation, with remarkable in-context learning abilities

"Generative Multimodal Models are In-Context Learners", Sun et al. CVPR 2024

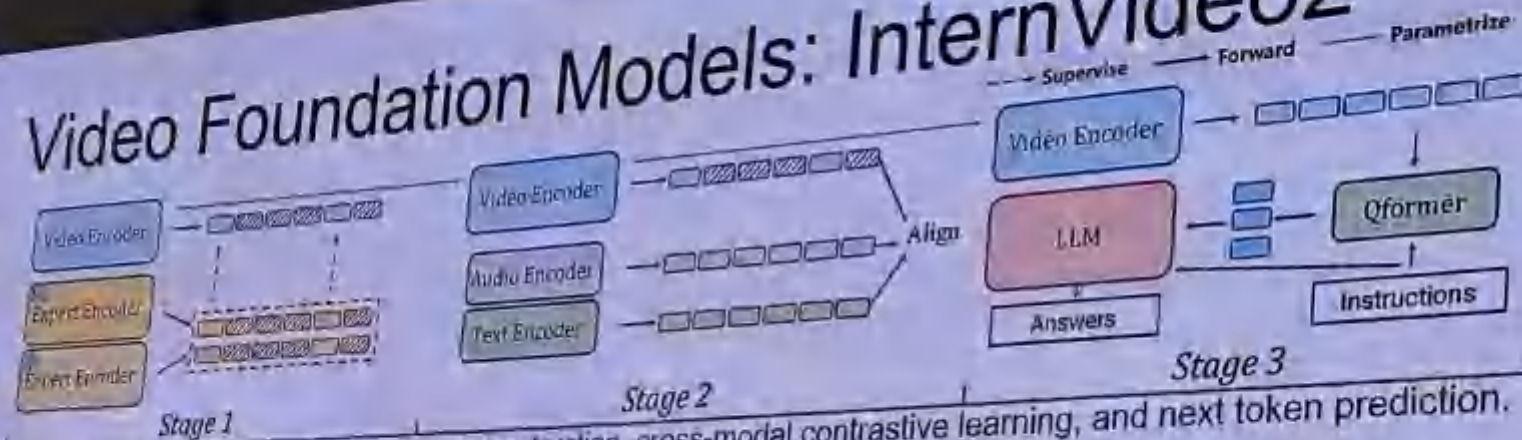


# More Modalities: NExT-GPT

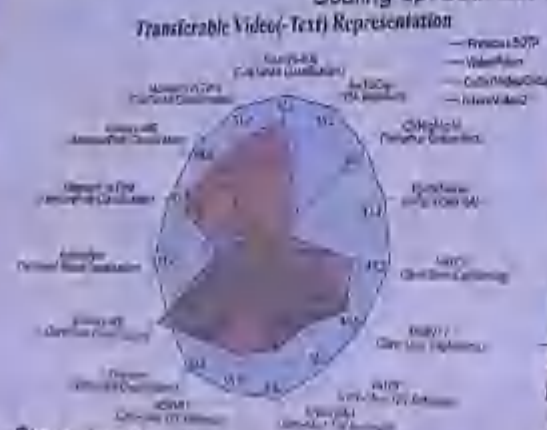


"NExT-GPT: Any-to-Any Multimodal LLM", Wu et al, arXiv:2309.05519

# Video Foundation Models: InternVideo2



Progressive training: masked video token reconstruction, cross-modal contrastive learning, and next token prediction.  
Scaling up: both data size (412M data entries) and model size (6B)  
Support for Long Videos



Strong transferable visual and visual-linguistic representations across 70 video understanding tasks, including action recognition, video-text understanding, video-centric dialogue, long-form video understanding and procedure-aware reasoning

"InternVideo2: Scaling Video Foundation Models for Multimodal Video Understanding", Wang et al. arXiv:2403.15377