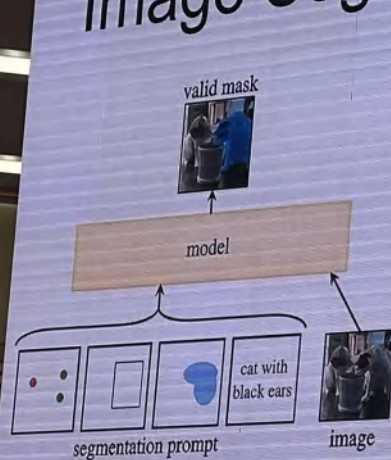
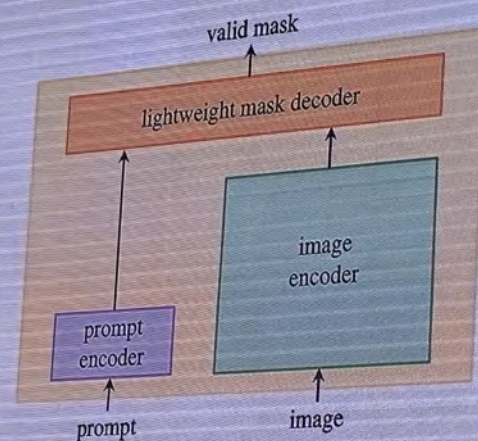


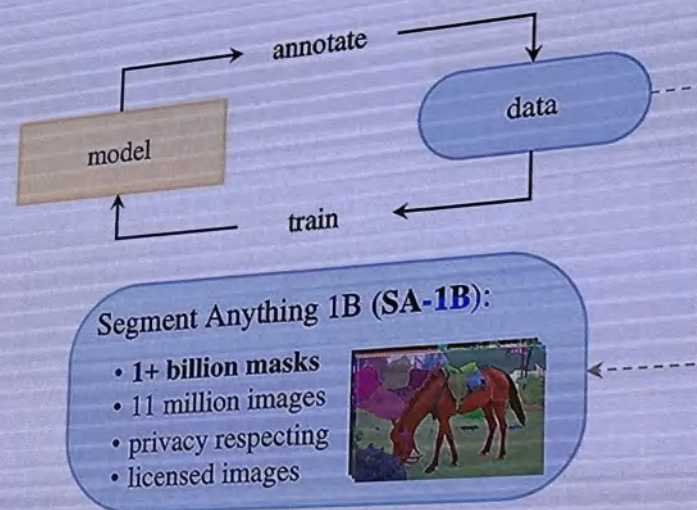
Image Segmentation: SAM



(a) Task: promptable segmentation



(b) Model: Segment Anything Model (SAM)



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

- Previous:

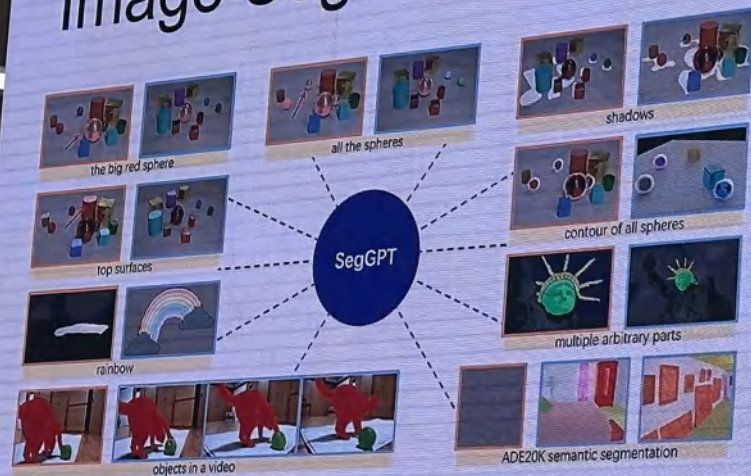
- Single task, single domain, closed set, single type of prompt, etc.

- SAM:

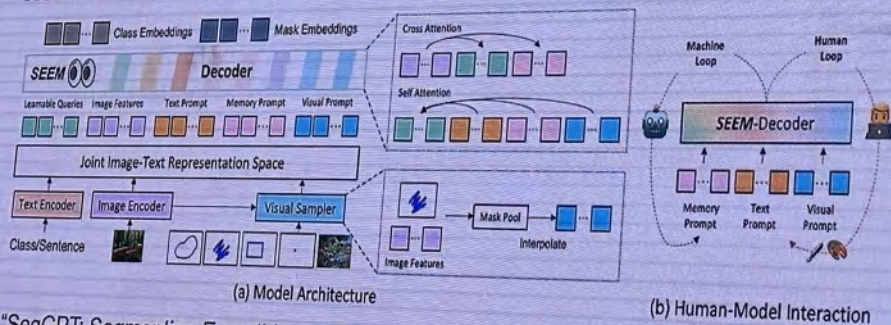
- Unified and generic interface for segmentation: mask, points, box, text
- Data engine scaling law: 1+ billion masks, 11 million images
- Strong generalization ability: even aerial images, synthetic images, and medical images

"Segment Anything", Kirillov et al. ICCV 2023

Image Segmentation: More Explorations



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

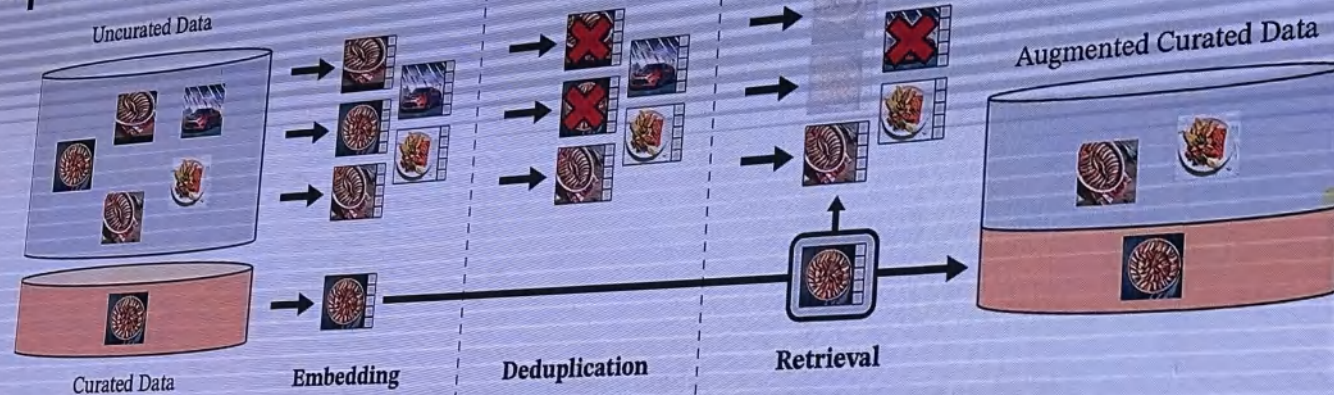


- SEEM:
 - More prompt formats, e.g., referred regions from other images
 - Project prompts into a joint visual-semantic space

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

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

Representation Learning: DINOv2



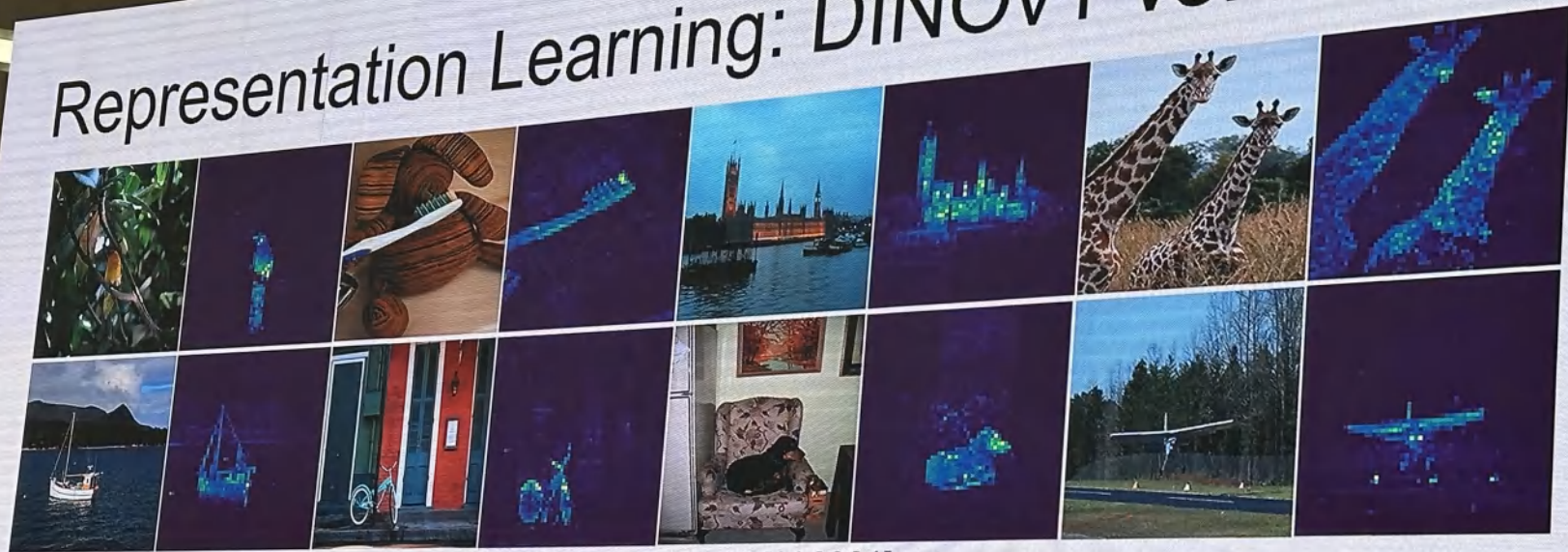
Method	Arch.	Data	Text sup.	kNN		linear	
				val	val	ResL	V2
Weakly supervised							
CLIP	ViT-L/14	WIT-400M	✓	79.8	84.3	88.1	75.3
CLIP	ViT-L/14 ₃₃₆	WIT-400M	✓	80.5	85.3	88.6	75.8
SWAG	ViT-H/14	IG3 6B	✓	82.6	85.7	83.7	77.6
OpenCLIP	ViT-H/14	LAION-2B	✓	81.7	84.4	88.4	75.5
OpenCLIP	ViT-G/14	LAION-2B	✓	83.2	86.2	89.4	77.2
EVA-CLIP	ViT-g/14	custom*	✓	83.5	86.4	89.3	77.4
Self-supervised							
MAE	ViT-H/14	INet-1k	X	49.4	76.6	83.3	64.8
DINO	ViT-S/8	INet-1k	X	78.6	79.2	85.5	68.2
SEERv2	RG10B	IG2B	X	-	79.8	-	-
MSN	ViT-L/7	INet-1k	X	79.2	80.7	86.0	69.7
ExViT	Swin-B/W=14	INet-1k	X	79.4	81.3	87.0	70.4
Mugs	ViT-L/16	INet-1k	X	80.2	82.1	86.9	70.8
IBOT	ViT-L/16	INet-22k	X	72.9	82.3	87.5	72.4
DINOv2	ViT-S/14	LVD-142M	X	79.0	81.1	86.6	70.9
	ViT-B/14	LVD-142M	X	82.1	84.5	88.3	75.1
	ViT-L/14	LVD-142M	X	83.5	86.3	89.5	78.0
	ViT-g/14	LVD-142M	X	83.5	86.5	89.6	78.4

- 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

Single images, e.g. LAION

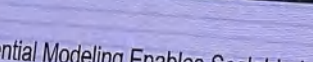


Image sequences,
e.g. videos, 3D
rotations, synthetic
viewpoints

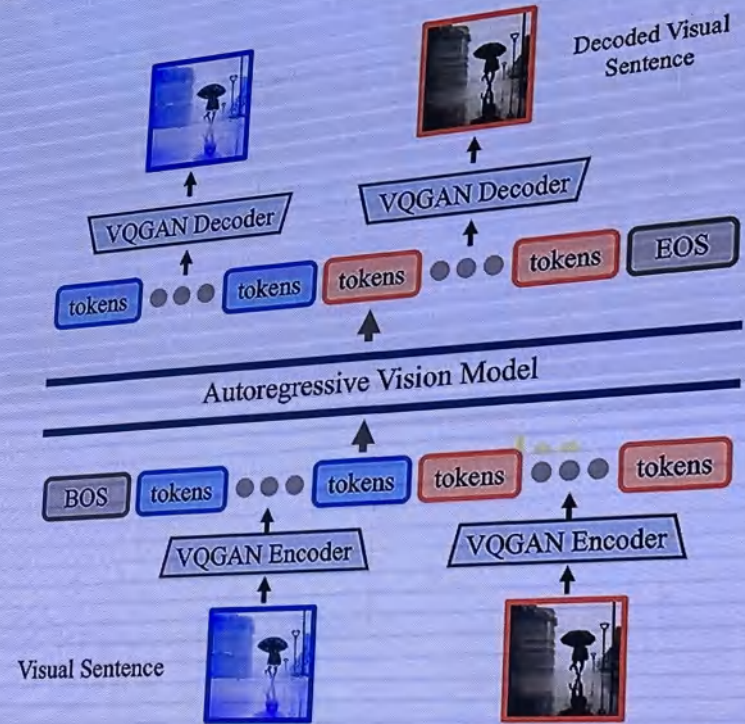
Images with annotation,
e.g. style transfer, object
detection, low light
enhancement

Images with free form
annotation,
e.g. object detection +
instance segmentation etc

Videos with
annotation,
e.g. video
segmentation

UVD-V1: Unified Vision Dataset

420B tokens,
50 Datasets.



- 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

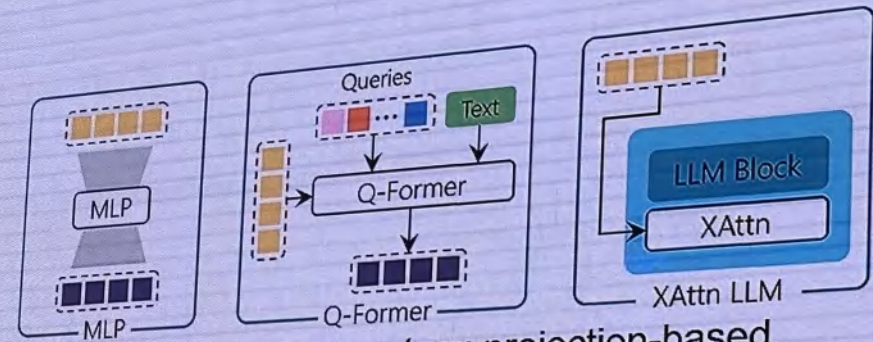
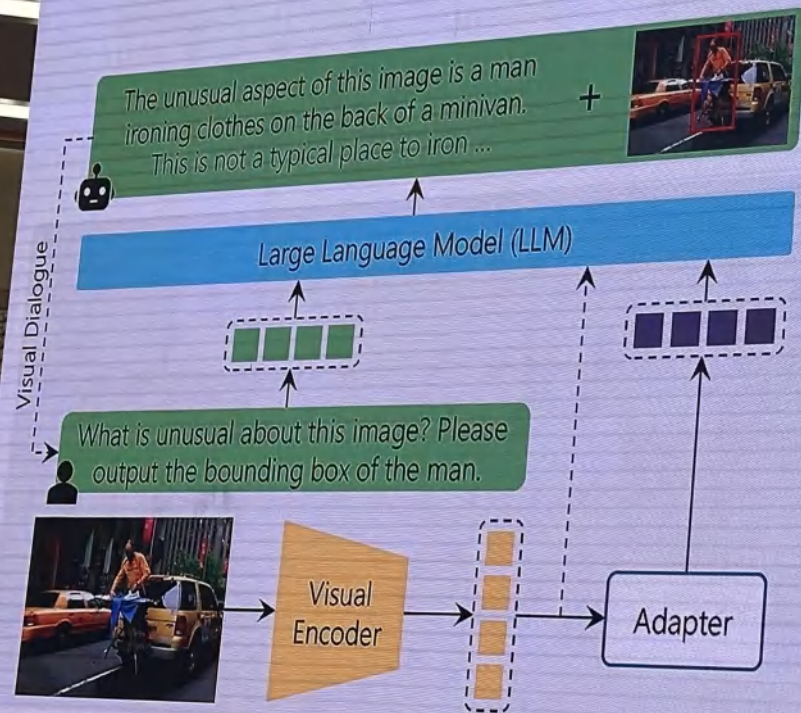


TABLE 1: A summary of commonly used image encoders.

Variants	Pretraining Corpus	Resolution	Samples (B)	Parameter Size (M)
OpenCLIP-ConvNext-L [46]	LAION-2B	320	29	197.4
CLIP-ViT-L/14 [13]	OpenAI's WIT	224/336	13	304.0
EVA-CLIP-ViT-G/14 [47]	LAION-2B, COYO-700M	224	11	1000.0
OpenCLIP-ViT-G/14 [46]	LAION-2B	224	34	1012.7
OpenCLIP-ViT-bigG/14 [46]	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.

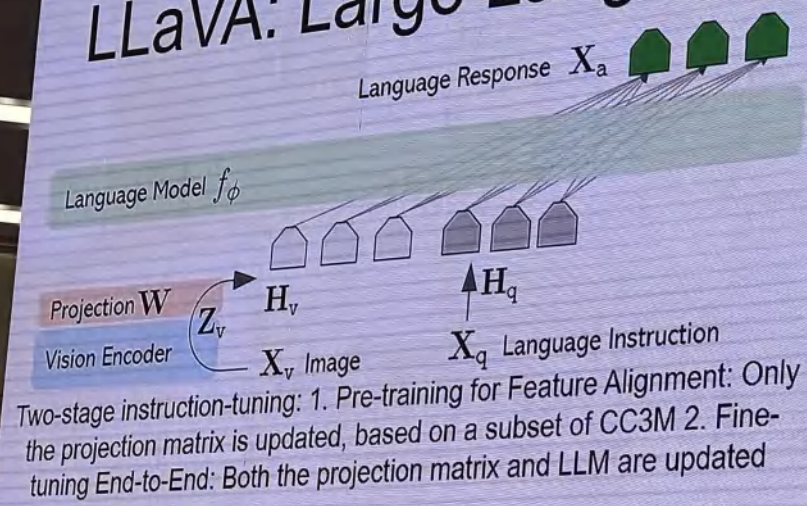
Model	Release Date	Pretrain Data Scale	Parameter Size (B)	Language Support	Architecture
Flan-T5-XL/XXL [56]	Oct-2022	-	3/ 11	en, fr, de	Encoder-Decoder
LLaMA [5]	Feb-2023	1.4T tokens	7/ 13/ 33/ 65	en	Causal Decoder
Vicuna [4]	Mar-2023	1.4T tokens	7/ 13/ 33	en	Causal Decoder
LLaMA-2 [57]	Jul-2023	2T tokens	7/ 13/ 70	en	Causal Decoder
Qwen [58]	Sep-2023	3T tokens	1.8 / 7/ 14/ 72	en, zh	Causal Decoder

Commonly used open-sourced LLMs

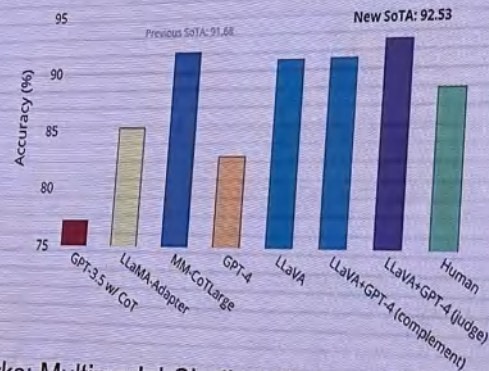
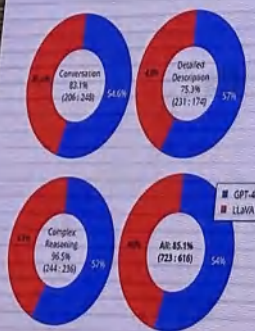
"The (R)Evolution of Multimodal Large Language Models: A Survey", Caffagni et al. arXiv:2312.00785

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

LLaVA: Large Language and Vision Assistant



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 Ironing:



Source: <https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg>

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 [36]

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.

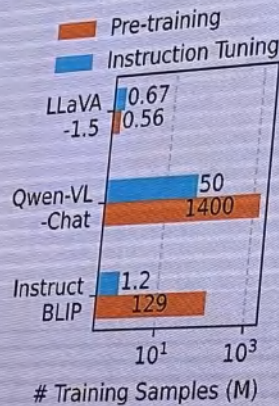
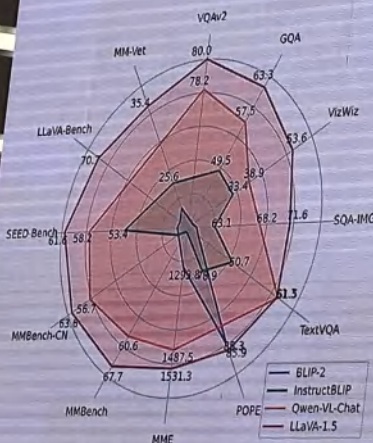
User
BLIP-2

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



language model (Vicuna v1.5 13B)

vision-language connector (MLP)

tokenizer & embedding

vision encoder (CLIP ViT-L/336px)

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 (OCRVQA, 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

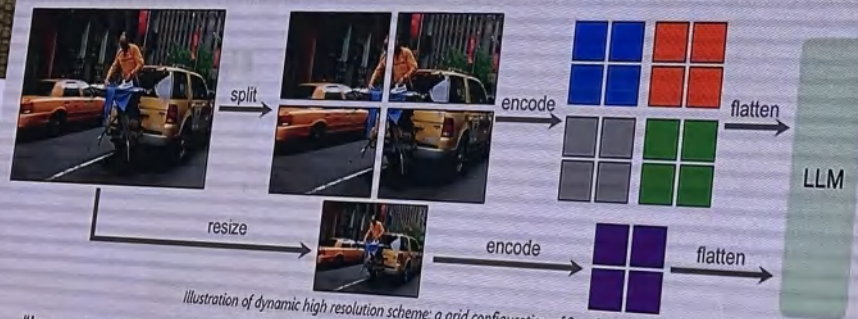


Illustration of dynamic high resolution scheme: a grid configuration of 2×2

"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

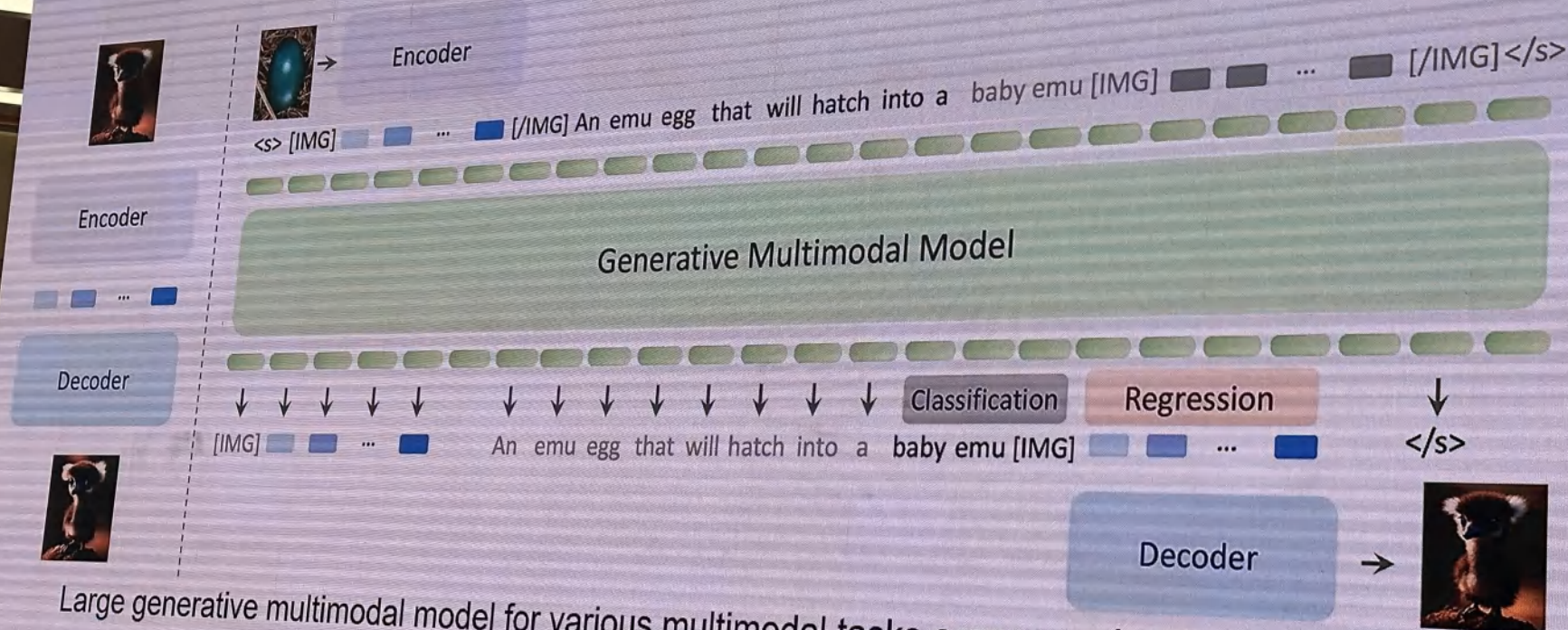
Model	LLM	Visual Encoder	V2L Adapter	VInstr. Tuning	Main Tasks & Capabilities
BLIP-2 (Li et al., 2023f)	FlanT5-XXL-11B*	EVA ViT-g	Q-Former	✓	Visual Dialogue, VQA, Captioning, Retrieval
FROMAGE (Koh et al., 2023b)	OPT-6.7B*	CLIP ViT-L	Linear	✓	Visual Dialogue, Captioning, Retrieval
Kosmos-1 (Huang et al., 2023a)	Magneto-1.3B ^o	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 (Awadalla et al., 2023)	MPT-7B*	CLIP ViT-L	XAttn LLM	✓	VQA, Captioning
Flamingo (Alayrac et al., 2022)	Chinchilla-70B*	NFNet-F6	XAttn LLM	✓	Visual Dialogue, VQA, Captioning
PaLI (Chen et al., 2023i)	mT5-XXL-13B ^o	ViT-e	XAttn LLM	✓	Multilingual, VQA, Captioning, Retrieval
PaLI-X (Chen et al., 2023g)	UL2-32B*	ViT-22B	XAttn LLM	✓	Multilingual, VQA, Captioning
LLaVA (Liu et al., 2023e)	Vicuna-13B ^o	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning
MiniGPT-4 (Zhu et al., 2023a)	Vicuna-13B*	EVA ViT-g	Linear	✓	VQA, Captioning
mPLUG-Owl (Ye et al., 2023c)	LLaMA-7B*	CLIP ViT-L	Q-Former*	✓	Visual Dialogue, VQA
InstructBLIP (Dai et al., 2023)	Vicuna-13B*	EVA ViT-g	Q-Former	✓	Visual Dialogue, VQA, Captioning
MultiModal-GPT (Gong et al., 2023)	LLaMA-7B*	CLIP ViT-L	XAttn LLM	✓	Visual Dialogue, VQA, Captioning
LaVIN (Luo et al., 2023)	LLaMA-13B*	CLIP ViT-L	MLP	✓	Visual Dialogue, VQA, Captioning
Otter (Li et al., 2023a)	LLaMA-7B*	CLIP ViT-L	XAttn LLM	✓	VQA, Captioning
Kosmos-2 (Peng et al., 2023)	Magneto-1.3B ^o	CLIP ViT-L	Q-Former*	✓	Visual Dialogue, VQA, Captioning, Referring, REC
Shikra (Chen et al., 2023f)	Vicuna-13B ^o	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning, Referring, REC, GroundCap
Clever Flamingo (Chen et al., 2023b)	LLaMA-7B*	CLIP ViT-L	XAttn LLM	✓	Visual Dialogue, VQA, Captioning
SVIT (Zhao et al., 2023a)	Vicuna-13B ^o	CLIP ViT-L	MLP	✓	Visual Dialogue, VQA, Captioning
BLIVA (Hu et al., 2024)	Vicuna-7B*	EVA ViT-g	Q-Former+Linear	✓	Visual Dialogue, VQA, Captioning
IDEFICS (Laurençon et al., 2023)	LLaMA-65B*	OpenCLIP ViT-H	XAttn LLM	✓	Visual Dialogue, VQA, Captioning
Qwen-VL (Bai et al., 2023b)	Qwen-7B ^o	OpenCLIP ViT-bigG	Q-Former*	✓	Visual Dialogue, Multilingual, VQA, Captioning, REC
StableLLaVA (Li et al., 2023h)	Vicuna-13B ^o	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning
Ferret (You et al., 2023)	Vicuna-13B ^o	CLIP ViT-L	Linear	✓	Visual Dialogue, Captioning, Referring, REC, GroundCap
LLaVA-1.5 (Liu et al., 2023d)	Vicuna-13B ^o	CLIP ViT-L	MLP	✓	Visual Dialogue, VQA, Captioning
MiniGPT-v2 (Chen et al., 2023e)	LLaMA-2-7B*	EVA ViT-g	Linear	✓	Visual Dialogue, VQA, Captioning, Referring, REC, GroundCap
Pink (Xuan et al., 2023)	Vicuna-7B*	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning, Referring, REC, GroundCap
CogVLM (Wang et al., 2023c)	Vicuna-7B ^o	EVA ViT-E	MLP	✓	Visual Dialogue, VQA, Captioning, REC
DRESS (Chen et al., 2023j)	Vicuna-13B*	EVA ViT-g	Linear	✓	Visual Dialogue, VQA, Captioning
LION (Chen et al., 2023d)	FlanT5-XXL-11B*	EVA ViT-g	Q-Former+MLP	✓	Visual Dialogue, VQA, Captioning, REC
mPLUG-Owl2 (Ye et al., 2023d)	LLaMA-2-7B ^o	CLIP ViT-L	Q-Former*	✓	Visual Dialogue, VQA, Captioning
SPHINX (Lin et al., 2023b)	LLaMA-2-13B ^o	Mixture	Linear	✓	Visual Dialogue, VQA, Captioning, Referring, REC, GroundCap
Honeybee (Cha et al., 2023)	Vicuna-13B ^o	CLIP ViT-L	ResNet blocks	✓	Visual Dialogue, VQA, Captioning
VILA (Lin et al., 2023a)	LLaMA-2-13B ^o	CLIP ViT-L	Linear	✓	Visual Dialogue, VQA, Captioning
SPHINX-X (Gao et al., 2024)	Mixtral-8x7B ^o	Mixture	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], InfMM-HD [CAS&UCAS&ByteDance], Fuyu [Adept], Otter-HD [NTU], LLaVA-HR [XMU&Pengcheng], LLaVA-UHD [THU&NUS&UCAS], Griffon 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], DetGPT [HKUST&HKU&SJTU], PixelLM [BJTU&USTB&ByteDance], GLaMM [MBZUI etc.], Groma [HKU&ByteDance], Griffon&Griffon 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]
- ...

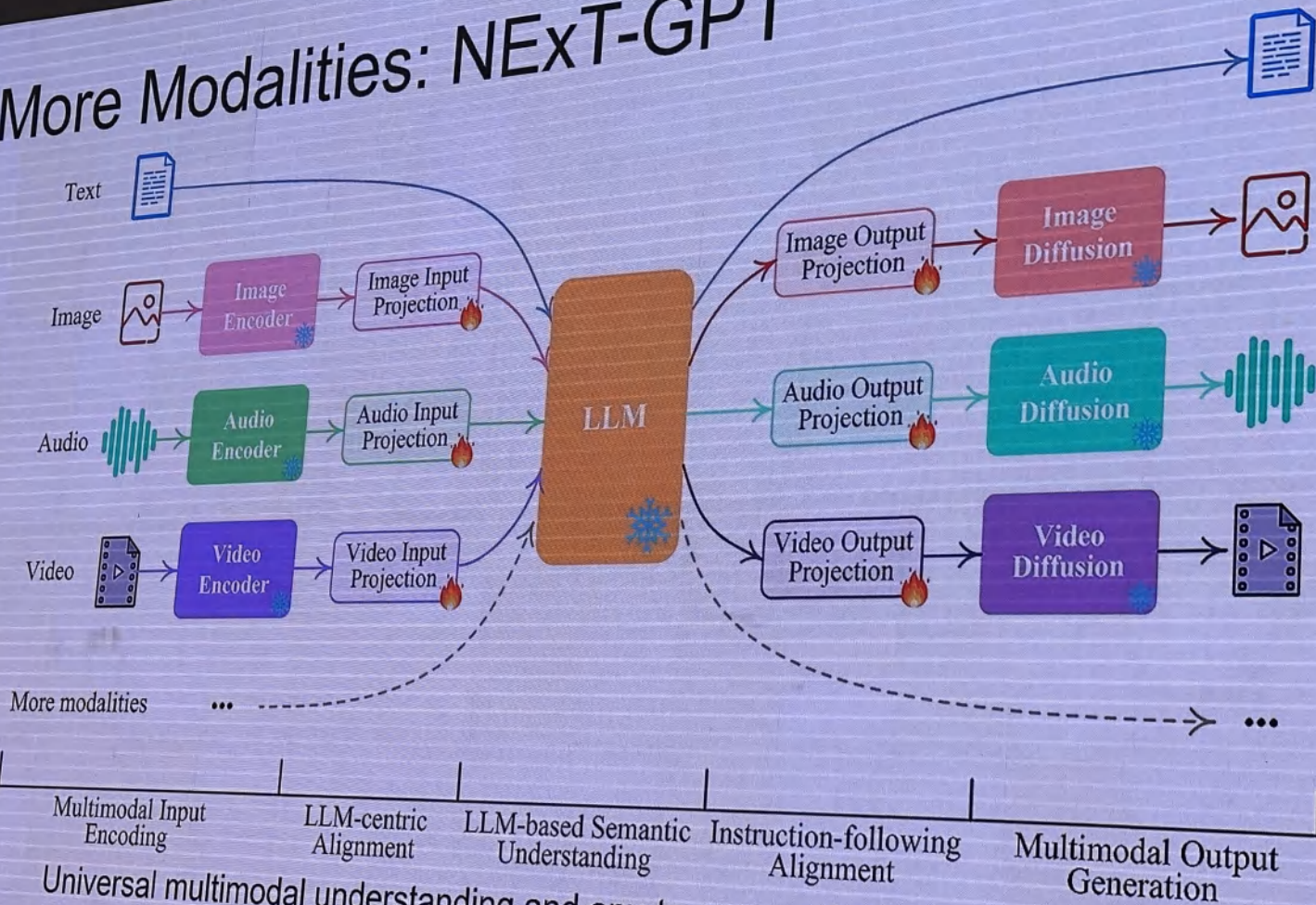
Understanding and Generation: Emu2



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

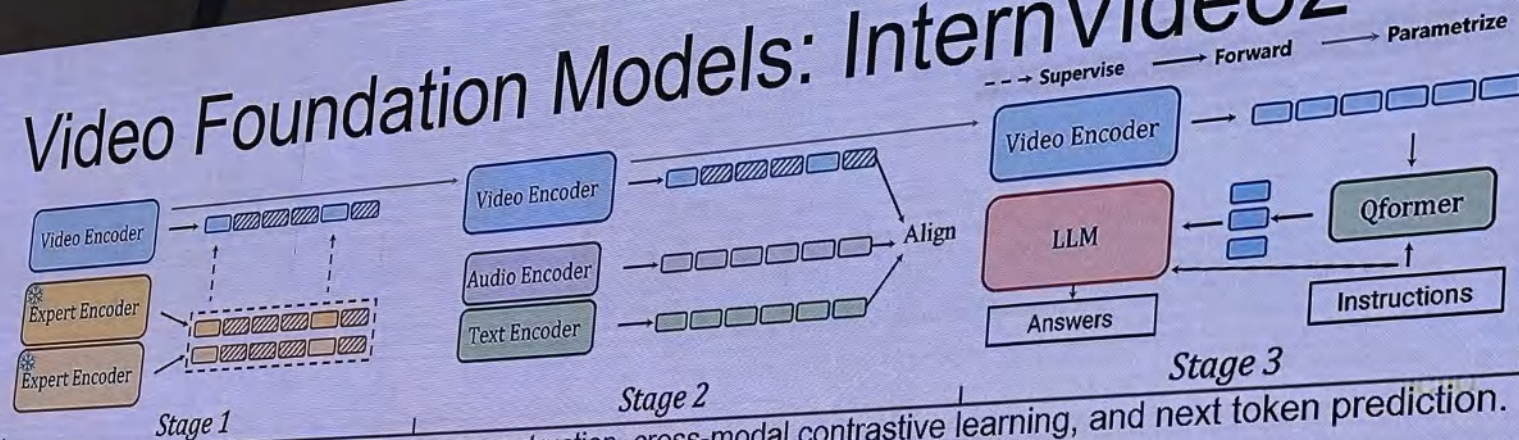
"Generative Multimodal Models are In-Context Learners", Sun et al. CVPR 2024

More Modalities: NExT-GPT



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

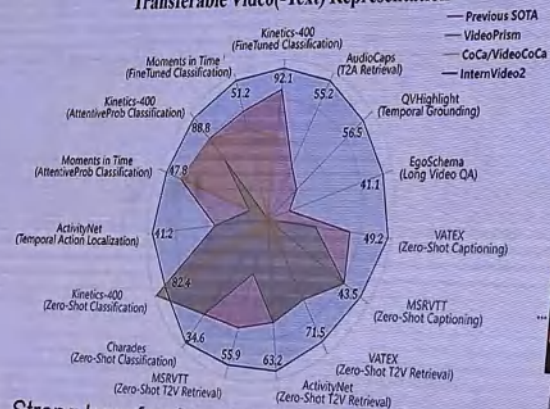
Video Foundation Models: InternVideo2



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)

Transferable Video(-Text) Representation



Support for Long Videos



Temporal Commonsense Reasoning



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