LMM

Understandability of Multi-Modal LLMs

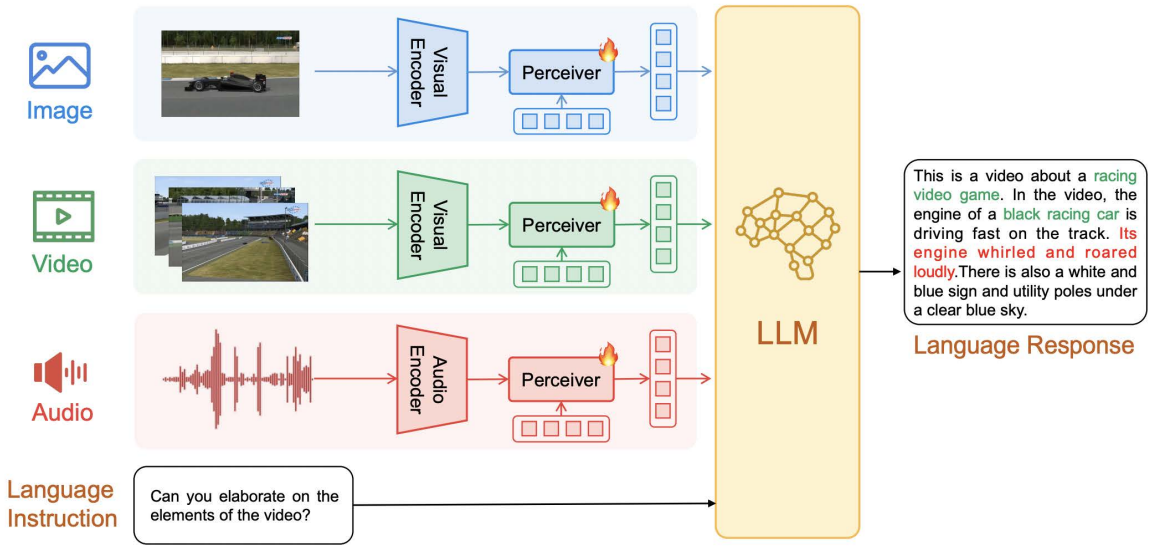
Presented by Hantao Zhou

UHH

December 19, 2023

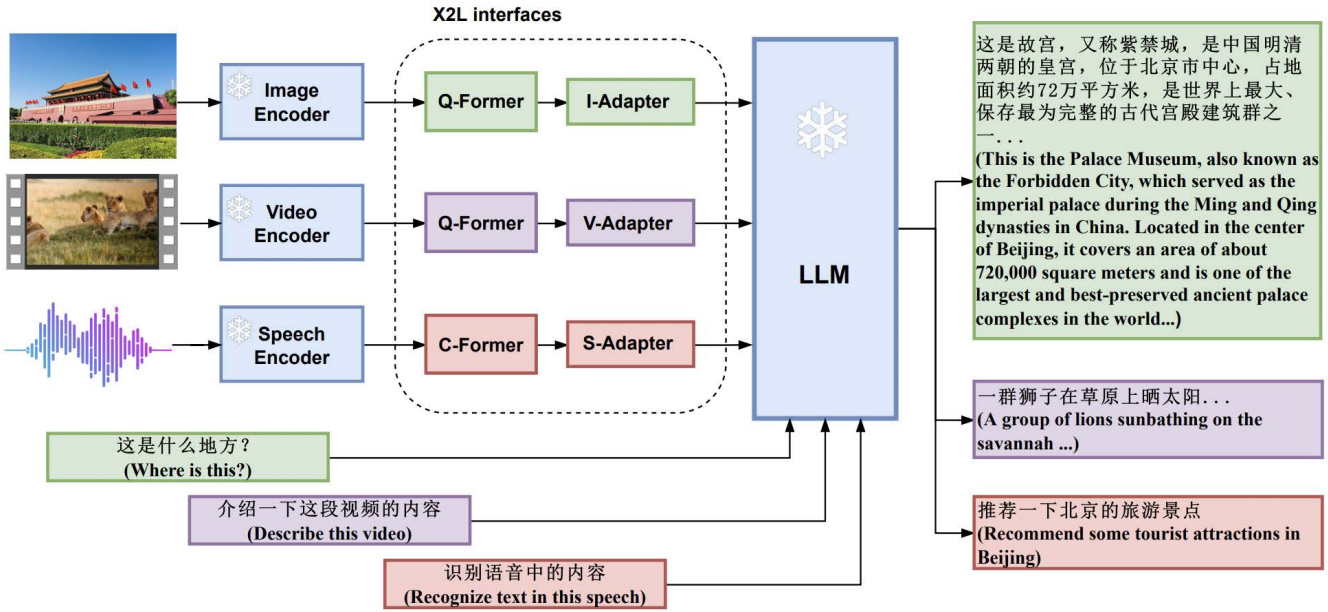
ChatBridge: Zhao et al. (2023)

Model Architecture



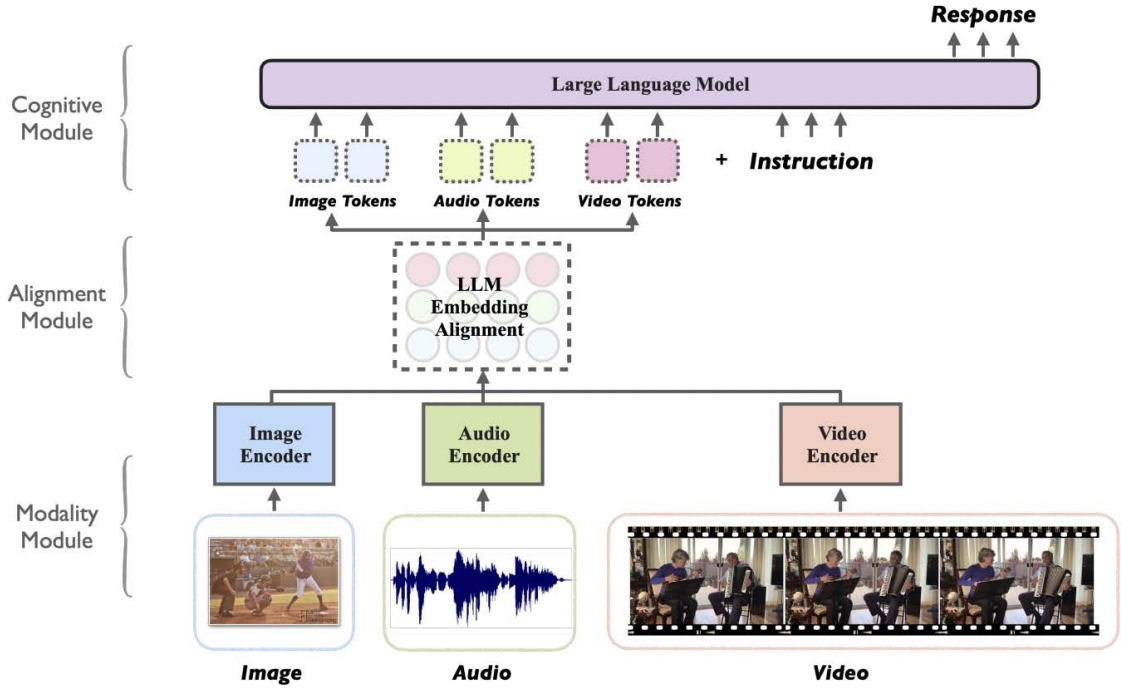
- **ImageEncoder/VideoEncoder:** ViT-G
  - Sample four frames from each video, concatenate frame features as the video features.
- **AudioEncoder:** BEATs - Chen et al. (2022)
  - Divide into 10-second clips, concatenate the clip features as the audio features.
- **Large Language Model:** Vicuna-13B
  - Perceiver: transformer-decoders (only train the perceivers and their learnable query tokens).

X-LLM Chen et al. (2023)



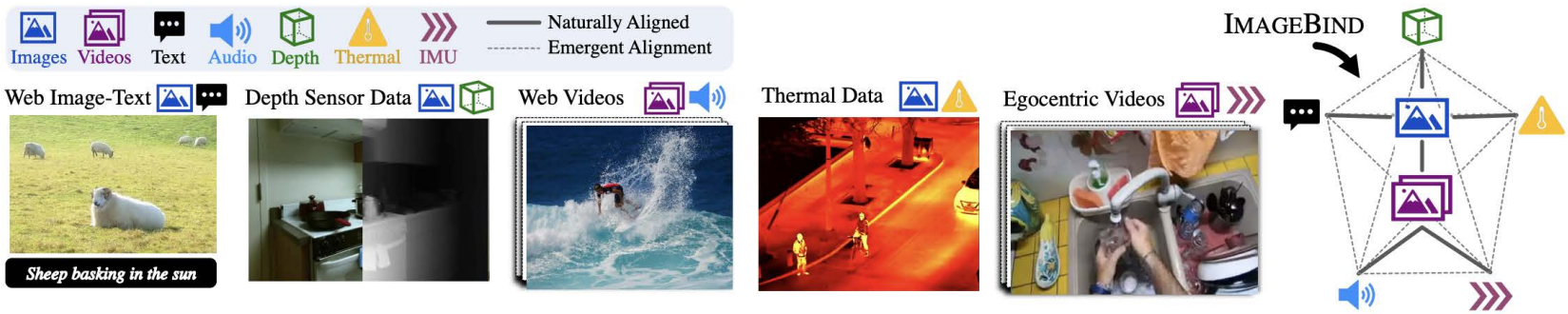
- **Model Architecture**
  - X Encoders
    - The image and video encoders share a pre-trained ViT-G (T frames)
    - The speech encoder consists of convolution layers and a conformer structure Gulati et al. (2020)
  - X2L interface
    - The Image and Video Q-former: initialized from BLIP2's second stage of Q-Former
    - C-former: CIF module Dong and Xu (2020) and a 12-layer transformer structure (frame-level to token-level)
    - I/V/A-Adapter: linear layers
    - LLM: ChatGLM-6B

MACAW-LLM Lyu et al. (2023)



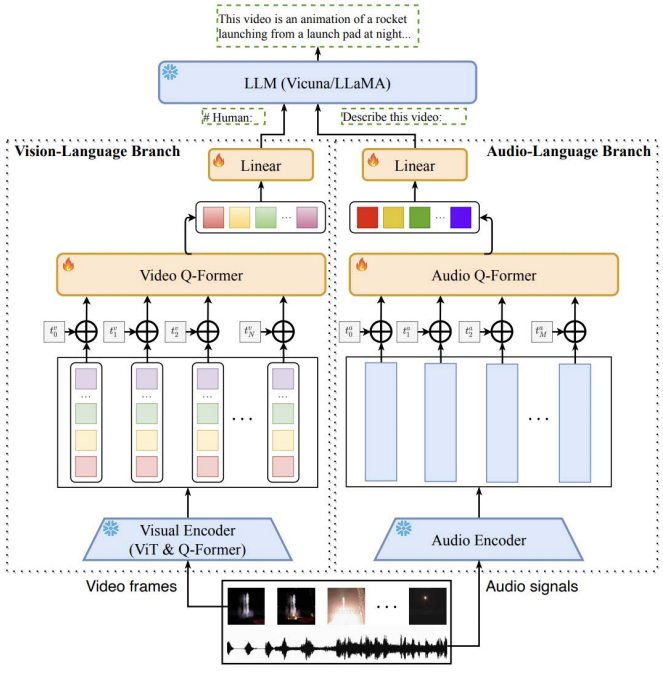
- Model Architecture
  - Modality Encoder
    - Visual Modality Encoder: CLIP (CLIP-VIT-B/16)
    - Audio Modality Encoder: WHISPER (WHISPER-BASE)
  - Large Language Model: LLaMA-7B
  - Alignment Model
    - $h1 = \text{Linear}(\text{Conv1D}(h))$, $ha = \text{Attn}(h1, E, E)$ (1)
  - LLM: LLAMA-7B

ImageBind Girdhar et al. (2023)



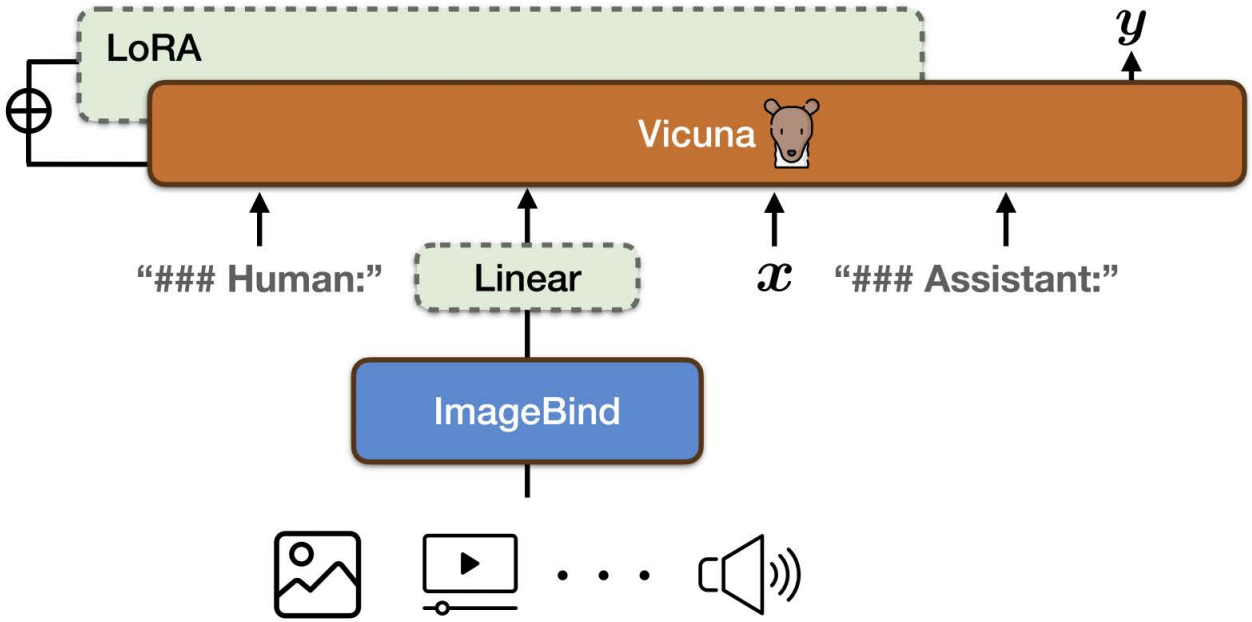
- Motivation
  - The absence of large quantities of multimodal data where all modalities are present together.
  - Utilize binding property of images and show that just aligning each modality's embedding to image embeddings leads to an emergent alignment across all of the modalities.
- Emergent Zero-shot Alignment
  - ImageBind aligns (I,M) by using contrastive learning, aligning every other modality to image I.
  - By training (I,M1) and (I,M2), (M1,M2) will be aligned together.

Video-LLaMA Zhang et al. (2023)



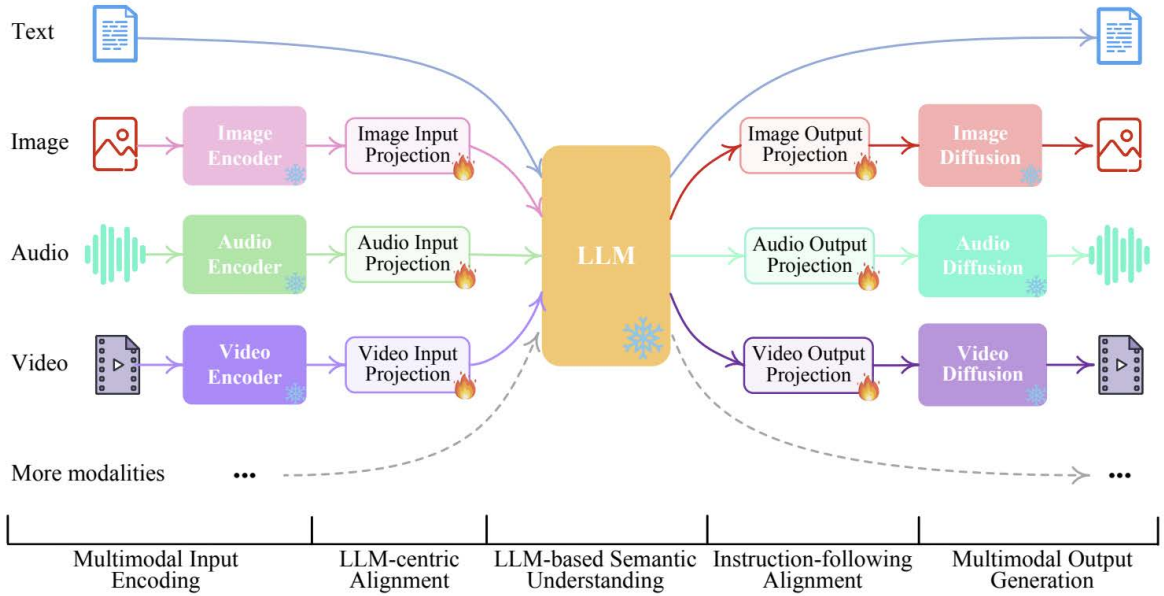
- Model Architecture
  - Vision-Language Branch
    - Frozen image-encoder (ViT-G/14) embeds each video frame (N frame)
  - Audio-Language Branch
    - Scarcity of audio-text data; train the audio-language branch using visual-text data.
    - Pre-trained 'audio' encoder (ImageBind - Image Branch), bridge the ImageBind with the LLM.
  - Q-former (same architecture as Q-Former in BLIP-2)
  - Linear layer to adapt the representations to the input of LLMs.

PandaGPT Su et al. (2023)



- **Motivation**
  - leverages the power of multimodal encoders from ImageBind to bind all the modalities to image
  - then align the ImageBind to the LLM by image-text instruction dataset
- **Model Architecture**
  - Imagebind encoders for all modalities
  - LLM: Vicuna-7b/13b
  - a linear projection layer f to connect the representation produced by ImageBind to Vicuna

NExT-GPT Wu et al. (2023)



- **Motivation**
  - Multimodal Large Language Models (MM-LLMs) have made exciting strides, they mostly fall prey to the limitation of only input-side multimodal understanding, without the ability to produce content in multiple modalities.
- **Model Architecture**
  - Multimodal Encoding Stage: ImageBind + Linear projection layer
  - LLM: Vicuna-7b
  - Multimodal Decoding Stage: Transformer-based projection layer + Diffusion Models (stable diffusion, zeroscope, AudioLDM)
  - To solve the OOD problem, during multimodal alignment training, they use image, video, audio-text pairs.

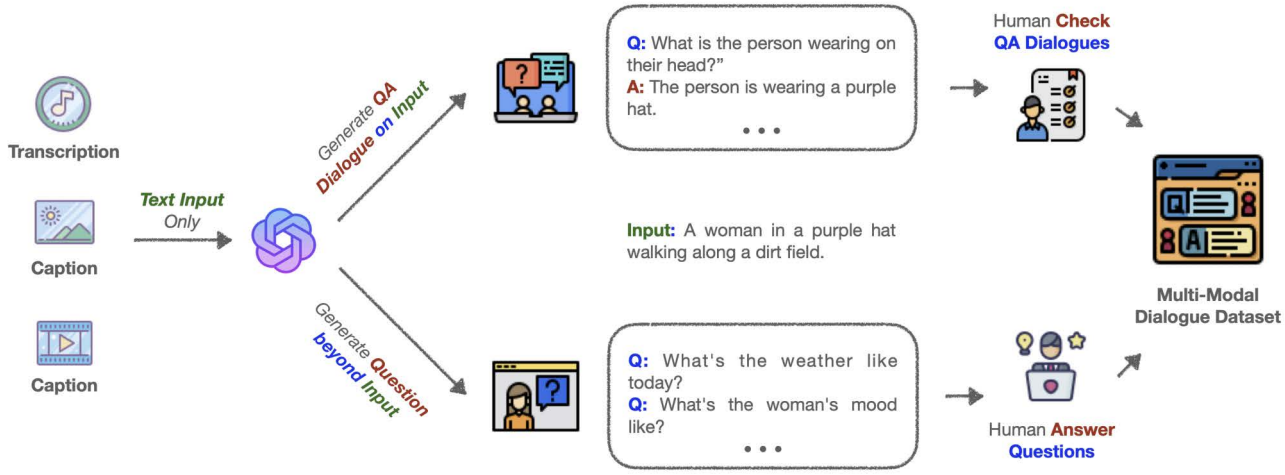
Multi-Modal LLM Training

- **One stage training (Instruction-Tuning)**
  - Macaw-LLM: image, audio, video - language instruction-response data
  - PandaGPT: image-language instruction-response data
- **Two stages training (Multimodal-Alignment + Instruction-Tuning)**
  - **ChatBridge**
    1. image, audio, video - language paired data;
    2. multimodal - language instruction-response data
  - **X-LLM**
    1. image, speech, video - language paired data
    2. multimodal - language instruction-response data;
  - **Video-LLaMA**
    1. image, video - language paired data
    2. image, video - language instruction-response data;
  - **NExT-GPT (Understanding Branch)**
    1. image, audio, video - language paired data
    2. multimodal - language instruction-response data

Categorization

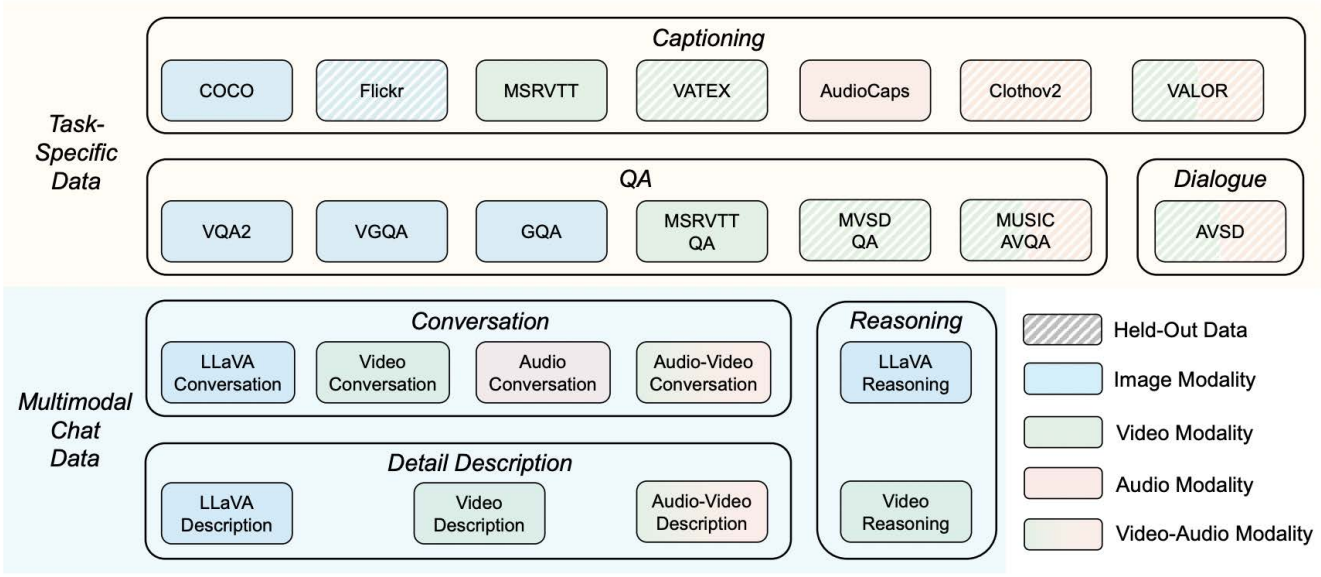
- **Benchmark adaptation**
  - PandaGPT: 160k image-language instruction-following data (MiniGPT-4, LLaVA)
  - Video-LLaMA: 1. Webvid-2M + CC595K; 2. MiniGPT-4 + LLaVa + Video-Chat
- **Expert Tools**
  - ChatGPT-aided
    - X-LLM
    - Macaw-LLM
  - ChatGPT + Other Specific Tools
    - ChatBridge
    - NExT-GPT

Macaw-LLM Data Construction



- Utilize GPT-3.5-TURBO, generate 10 instruction-response pairs within a single query
  - Text Instruction Dataset: Alpaca instruction dataset (52,000 instruction-response examples)
  - Image Instruction Dataset: curate around 69K instruction-response pairs by generating them from COCO image captions
  - Video+Audio Instruction Dataset: generate approximately 50K video instruction-response examples by utilizing the video captions from the Charades and AVSD

ChatBridge Data Construction



- Multi-Modal Alignment Dataset
  - image-text: MS-COCO, SBU Captions, Conceptual Captions, LAION-115M
  - video-text: Webvid10M
  - audio-text: WavCaps
- Uni-Modal Instruction Dataset
- Task-Specific Data: VQA2, VG-QA, COCO Caption; MSRVTTQA, MSRVTT Caption; AudioCaps
- For each task, ChatGPT derive 10-15 unique instruction templates
- Specify desired response style (short, brief, single sentence)

Categorization

- Some of the Multi-Modal LLMs don't have a comprehensive evaluation, only case studies are provided.
  - Case Study
    - Video-LLaMA
    - Macaw-GPT
    - PandaGPT
  - Close-End Evaluation
    - ChatBridge: unimodal + multimodal zero-shot task-specific evaluation
    - X-LLM: ASR ablation study (non zero-shot task-specific evaluation)
    - NExT-GPT: unimodal non zero-shot task-specific evaluation
  - Open-End Evaluation
    - ChatBridge: umimodal + multimodal GPT scoring
    - X-LLM: image-text GPT scoring (same as LLaVA)

Evaluation Results of ChatBridge

Table 1: Zero-shot evaluation of SoTA methods on unimodal input tasks. We report the accuracy for QA tasks and the CIDEr [55] score for captioning tasks.

Methods	Image-Text Tasks				Video-Text Tasks		Audio-Text Tasks
	OKVQA QA	GQA QA	Flickr30k Caption	NoCaps Caption	MSVD QA	VATEX Caption	Clothov2 Caption
Finetuned SoTA	66.1 [18]	65.1 [70]	67.4 [72]	121.6 [31]	60.0 [12]	95.8 [12]	48.8 [38]
Flamingo-9B [3]	44.7	-	61.5	-	30.2	39.5	-
Flamingo-80B [3]	**50.6**	-	67.2	-	35.6	46.7	-
BLIP-2 (FlanT5-XXL) [17]	-	**42.4**	73.7	98.4	34.4	-	-
BLIP-2 (Vicuna-13B) [17]	-	32.3	71.6	103.9	20.3	-	-
ChatBridge w/o MULTIS	41.4	37.4	77.7	107.5	23.5	47.7	22.4
ChatBridge	45.2	41.8	**82.5**	**115.7**	**45.3**	**48.9**	**26.2**

Table 2: Zero-shot evaluation of the effect of multimodal inputs on multimodal input tasks.

Input Modality	AVSD Dialogue		VALOR Captioning		MUSIC-AVQA
	BLEU-4	CIDEr	BLEU-4	CIDEr	Acc.
Finetuned SoTA	40.0 [44]	108.5 [44]	9.6 [12]	61.5 [12]	78.9 [12]
Video	28.3	73.1	2.8	22.3	33.1
Audio	20.2	46.2	0.3	5.2	28.9
Video+Audio	**29.8**	**75.4**	**4.2**	**24.7**	**43.0**

- Uni-Modal Results
  - On image-text, video-text datasets, it achieves comparable performance
- Multi-Modal Results
  - Ablation study for using Multi-Modal dataset, better performance across all three tasks
    - Incorporating both video and audio for solving these tasks.

Evaluation Results of X-LLM and NExT-GPT

Method	B@4	METEOR	CIDEr
Oscar [46]	36.58	30.4	124.12
BLIP-2 [43]	43.7	—	145.8
OFA [86]	44.9	32.5	154.9
CoDi [78]	40.2	31.0	149.9
NExT-GPT	44.3	**32.9**	**156.7**

Table 6: Image-to-text generation (image captioning) results on COCO-caption data [50].

Method	SPIDEr	CIDEr
AudioCaps [38]	0.369	0.593
BART [26]	0.465	0.753
AL-MixGen [39]	0.466	0.755
CoDi [78]	0.480	0.789
NExT-GPT	**0.521**	**0.802**

Table 7: Audio-to-text generation (audio captioning) results on AudioCaps [38].

Method	B@4	METEOR
ORG-TRL [105]	43.6	28.8
GIT [85]	54.8	33.1
mPLUG-2 [91]	57.8	34.9
CoDi [78]	52.1	32.5
NExT-GPT	**58.4**	**38.5**

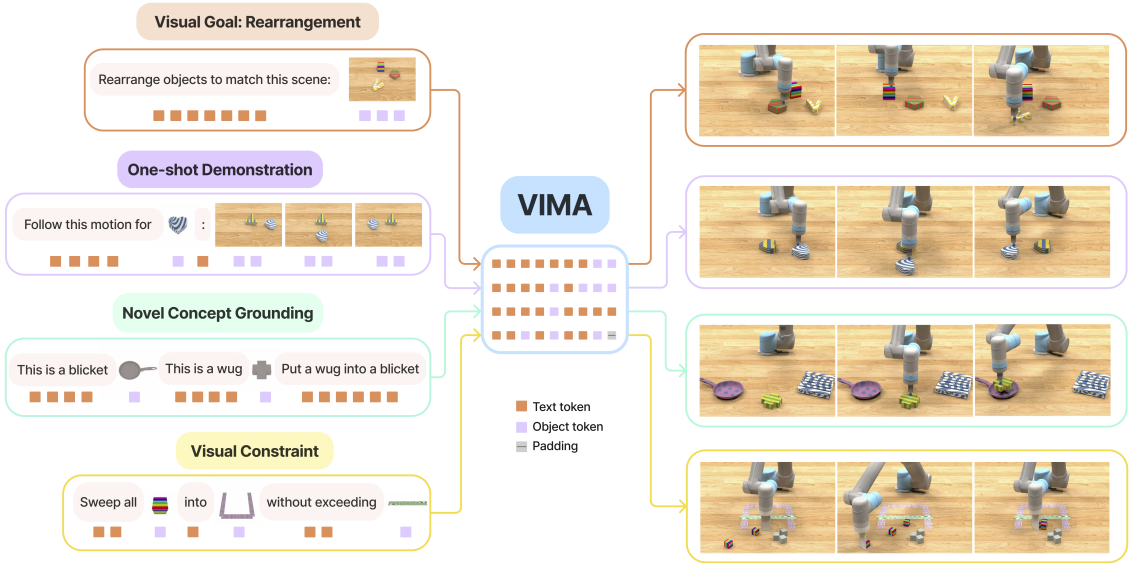
Table 8: Video-to-text generation (video captioning) results on MSR-VTT [92].

- Results of X-LLM
  - Uni-modal evaluation on image-text: comparable results with LLaVA, use of the BLIP2 pretrained Q-Former parameters significantly improves the model's performance.
- Results of NExT-GPT
  - Can mostly achieve much better performance on the X-to-text generation than the CoDi baseline, owing to the direct generation of texts from LLM, which is inherently expertized by the LLM.

Why is it so small

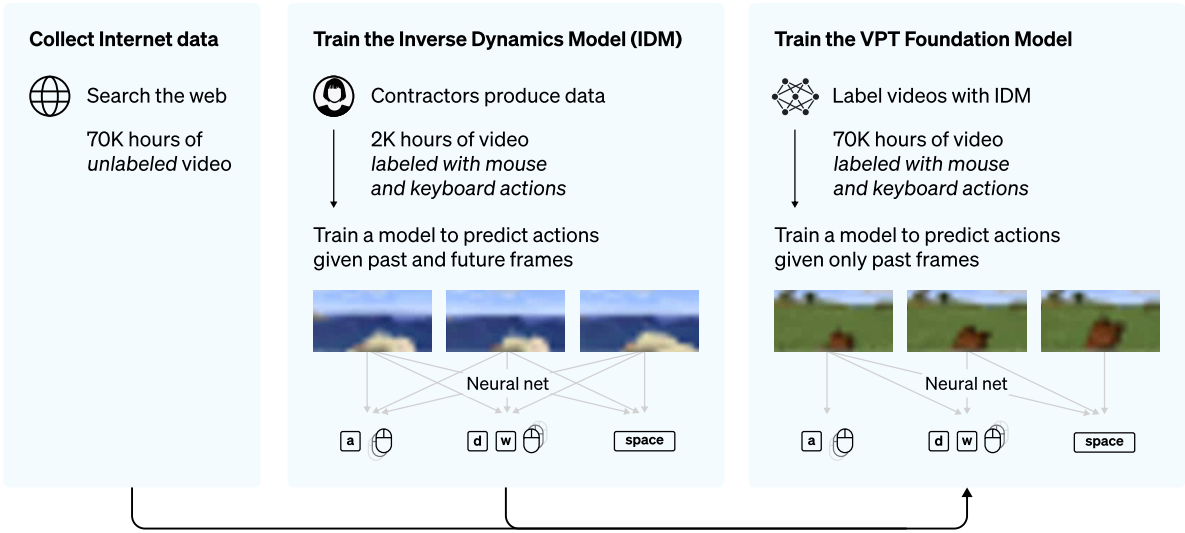
- the main topic today is LMM sorry for the mis-propaganda
- this area is of high potential yet remains lacking of explorations

VIMA



- Multimodal prompting formulation that converts diverse robot manipulation tasks into a uniform sequence modeling problem.
- T-5 for words, Masked RCNN for visual
-

VPT



- Use vast data collected from Internet as the demonstration for the construction of the IDM
- Utilizes the principle of weak supervision to train the ultimate model
-

VILA

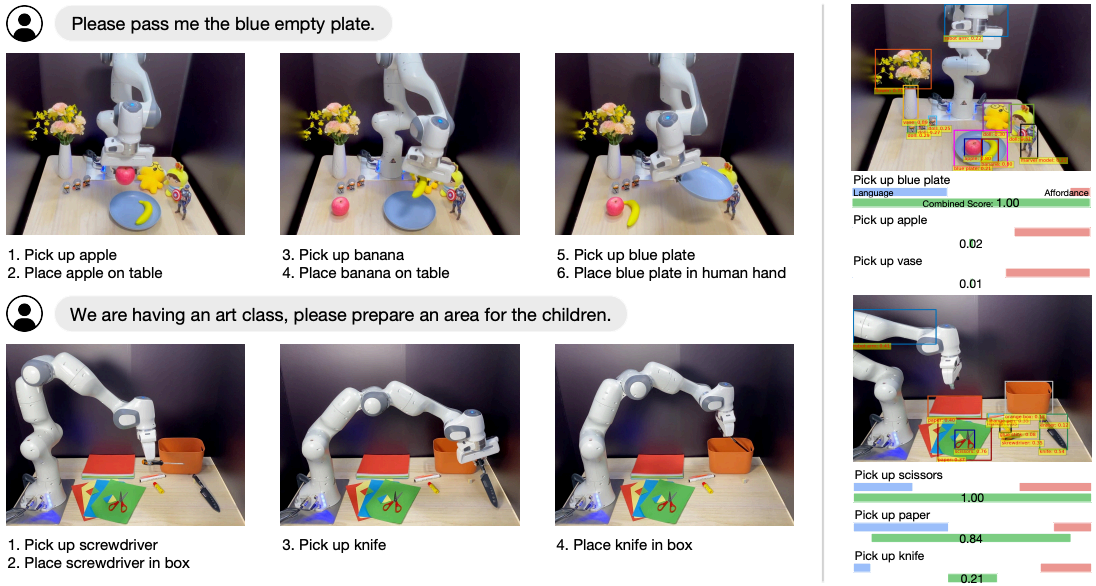
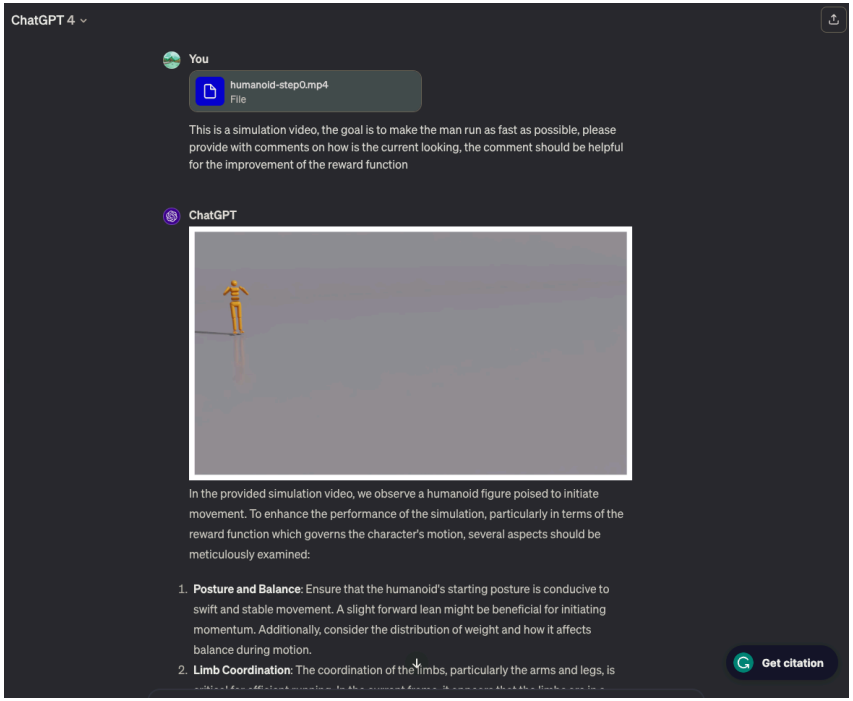


Fig. 3: **Illustration of the execution of VILA (left) and the decision-making process of SayCan (right)**. In the `Bring Empty Plate` task, the robot must first relocate the apple and banana from the blue plate. However, SayCan's initial step is to directly pick up the blue plate. In the `Prepare Art Class` task, while the scissor is supposed to remain on the table, SayCan erroneously picks up the scissor and places it in a box.

- Utilized the COT and GPT4V, add te closed-loop structure
- Mostly focuses on the reasoning performance in complicated scenario and tasks
- However most of the improvement are based on GPT-4V

Pre-experiments based on GPT-4V



- Can separate into several frames and explain accordingly
- Can detect the main object
- Can provide a vague description of the posture
- Have hallucinations, like saying there are two person when the fact being One
- Have off course results, like providing the method rather than the results
- but all of the cons can be solved by training our own model

Main Trend

- The entire Robotic-LLM work tends to combine with LMM
- LMM performs well in the planning task, yet others still remains to explore
- GPT-4V show a potential in recognizing the robot scenarios
- Weak supervision is widely accepted in LMM data constructions

Ongoing Proj. 1.

- Using simulated sceneries to train the models capabilities of providing valuable feedback
- Motivated by the self-supervised type of work in robotics and LMM's capabilities
- Serve as a further exploration based on the idea of EUREKA

Possible Proj. 2.

- Based on real robotics
- Whether the LMM can decide the actual performance in real scenario fits the Human's Desire.