Video encoding approaches for multimodal architectures

Andrey Kuznetsov
PhD, Head of FusionBrain Lab, AIRI

Term 5. Modern Methods and Algorithms of Generative Artificial Intelligence

Agenda

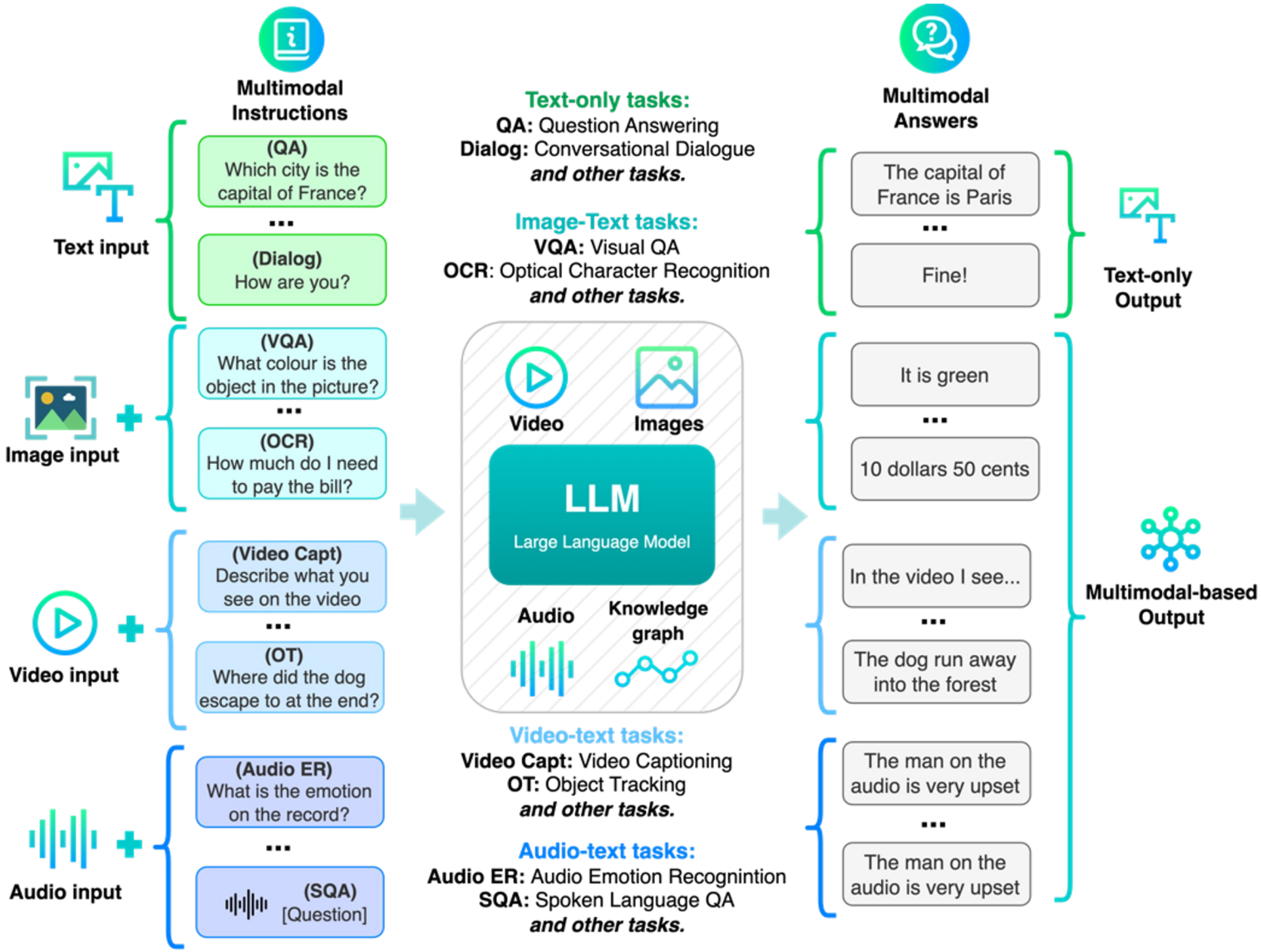
01 LLM and visual information

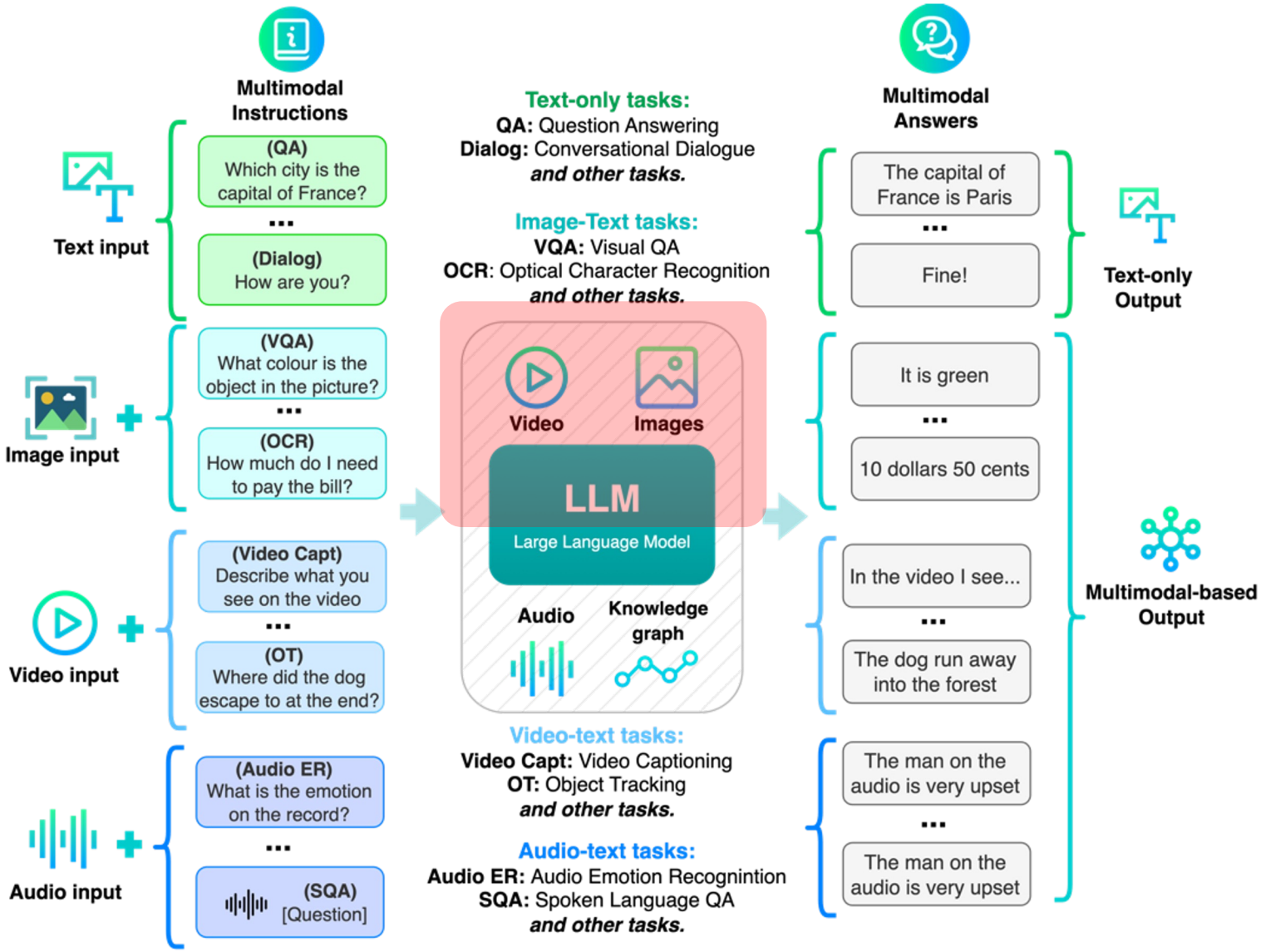
02 Video encoding and fusion with LLM

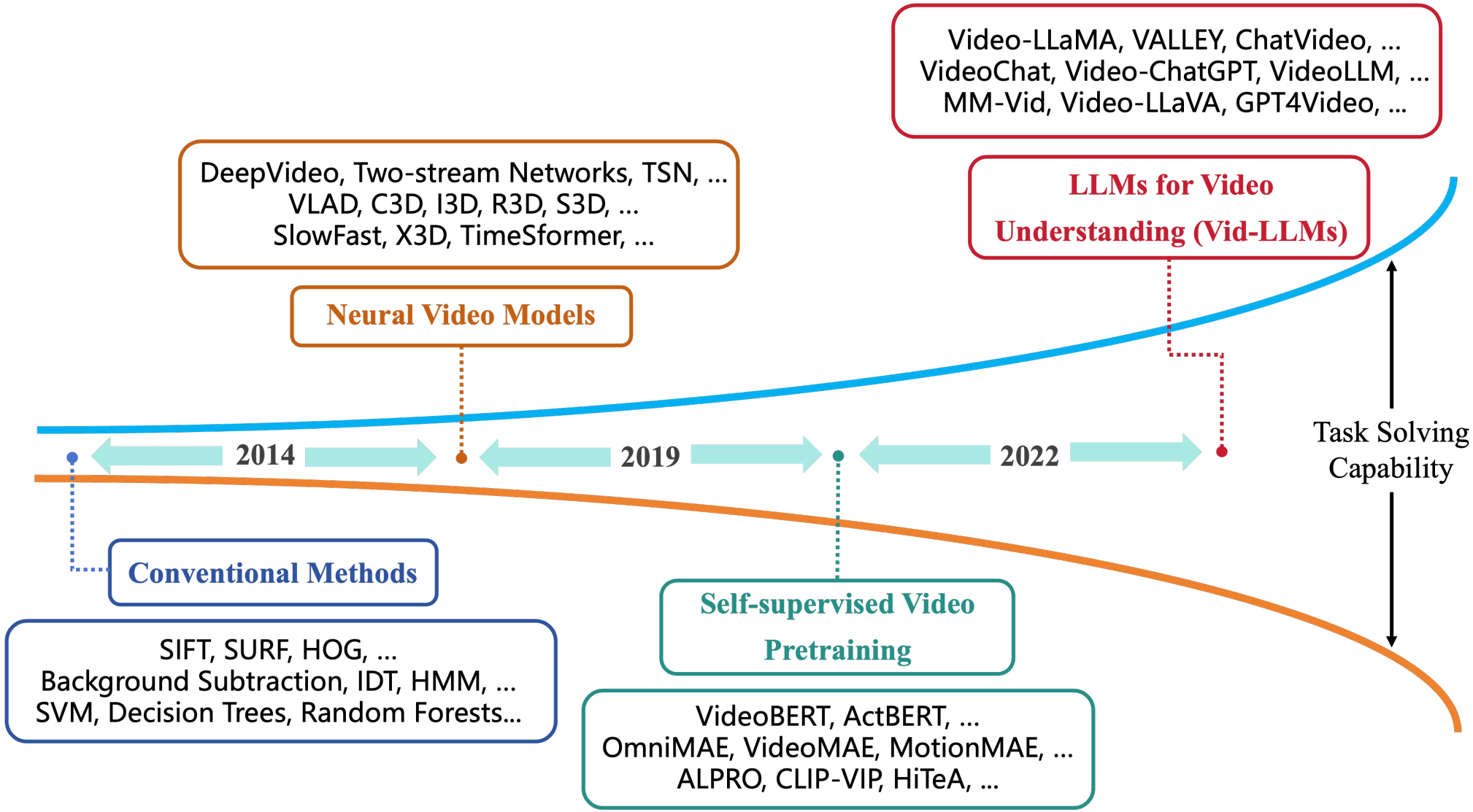
03 Conclusion + 🎁

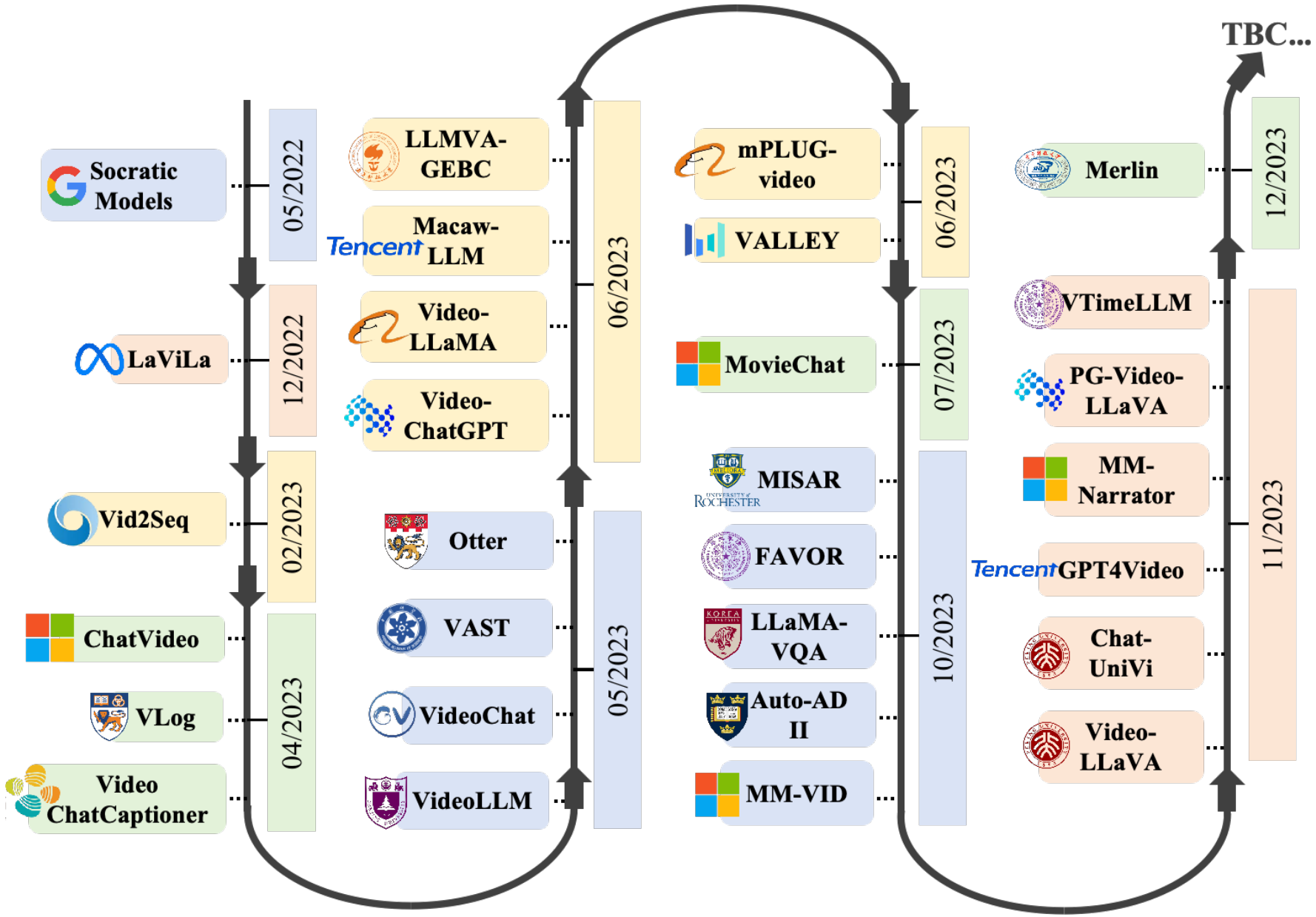
01

LLM and visual information


Multimodal Instructions
Text input
(QA) Which city is the capital of France?
...
(Dialog) How are you?
Image input
(VQA) What colour is the object in the picture?
...
(OCR) How much do I need to pay the bill?
Video input
(Video Capt) Describe what you see on the video
...
(OT) Where did the dog escape to at the end?
Audio input
(Audio ER) What is the emotion on the record?
...
(SQA) [Question]
Text-only tasks:
QA: Question Answering
Dialog: Conversational Dialogue
and other tasks.
Image-Text tasks:
VQA: Visual QA
OCR: Optical Character Recognition
and other tasks.
Video
Images
LLM
Large Language Model
Audio
Knowledge graph
Video-text tasks:
Video Capt: Video Captioning
OT: Object Tracking
and other tasks.
Audio-text tasks:
Audio ER: Audio Emotion Recognintion
SQA: Spoken Language QA
and other tasks.
Multimodal Answers
The capital of France is Paris
...
Fine!
Text-only Output
It is green
...
10 dollars 50 cents
In the video I see...
...
The dog run away into the forest
The man on the audio is very upset
...
The man on the audio is very upset
Multimodal-based Output


Multimodal Instructions
Text input
(QA) Which city is the capital of France?
...
(Dialog) How are you?
Image input
(VQA) What colour is the object in the picture?
...
(OCR) How much do I need to pay the bill?
Video input
(Video Capt) Describe what you see on the video
...
(OT) Where did the dog escape to at the end?
Audio input
(Audio ER) What is the emotion on the record?
...
(SQA) [Question]
Text-only tasks:
QA: Question Answering
Dialog: Conversational Dialogue
and other tasks.
Image-Text tasks:
VQA: Visual QA
OCR: Optical Character Recognition
and other tasks.
Video
Images
LLM
Large Language Model
Audio
Knowledge graph
Video-text tasks:
Video Capt: Video Captioning
OT: Object Tracking
and other tasks.
Audio-text tasks:
Audio ER: Audio Emotion Recognintion
SQA: Spoken Language QA
and other tasks.
Multimodal Answers
The capital of France is Paris
...
Fine!
Text-only Output
It is green
...
10 dollars 50 cents
In the video I see...
...
The dog run away into the forest
The man on the audio is very upset
...
The man on the audio is very upset
Multimodal-based Output


Video-LLaMA, VALLEY, ChatVideo, ...
VideoChat, Video-ChatGPT, VideoLLM, ...
MM-Vid, Video-LLaVA, GPT4Video, ...
DeepVideo, Two-stream Networks, TSN, ...
VLAD, C3D, I3D, R3D, S3D, ...
SlowFast, X3D, TimeSformer, ...
LLMs for Video Understanding (Vid-LLMs)
Neural Video Models
2014
2019
2022
Task Solving Capability
Conventional Methods
SIFT, SURF, HOG, ...
Background Subtraction, IDT, HMM, ...
SVM, Decision Trees, Random Forests...
Self-supervised Video Pretraining
VideoBERT, ActBERT, ...
OmniMAE, VideoMAE, MotionMAE, ...
ALPRO, CLIP-VIP, HiTeA, ...


TBC...
Socratic Models
05/2022
LaViLa
12/2022
Vid2Seq
02/2023
ChatVideo
VLog
Video ChatCaptioner
04/2023
LLMVA-GEBC
Macaw-LLM
Video-LLaMA
Video-ChatGPT
06/2023
Otter
VAST
VideoChat
VideoLLM
05/2023
mPLUG-video
VALLEY
06/2023
MovieChat
07/2023
MISAR
FAVOR
LLaMA-VQA
Auto-AD II
MM-VID
10/2023
Merlin
12/2023
VTimeLLM
PG-Video-LLaVA
MM-Narrator
GPT4Video
Chat-UniVi
Video-LLaVA
11/2023

OmniFusion

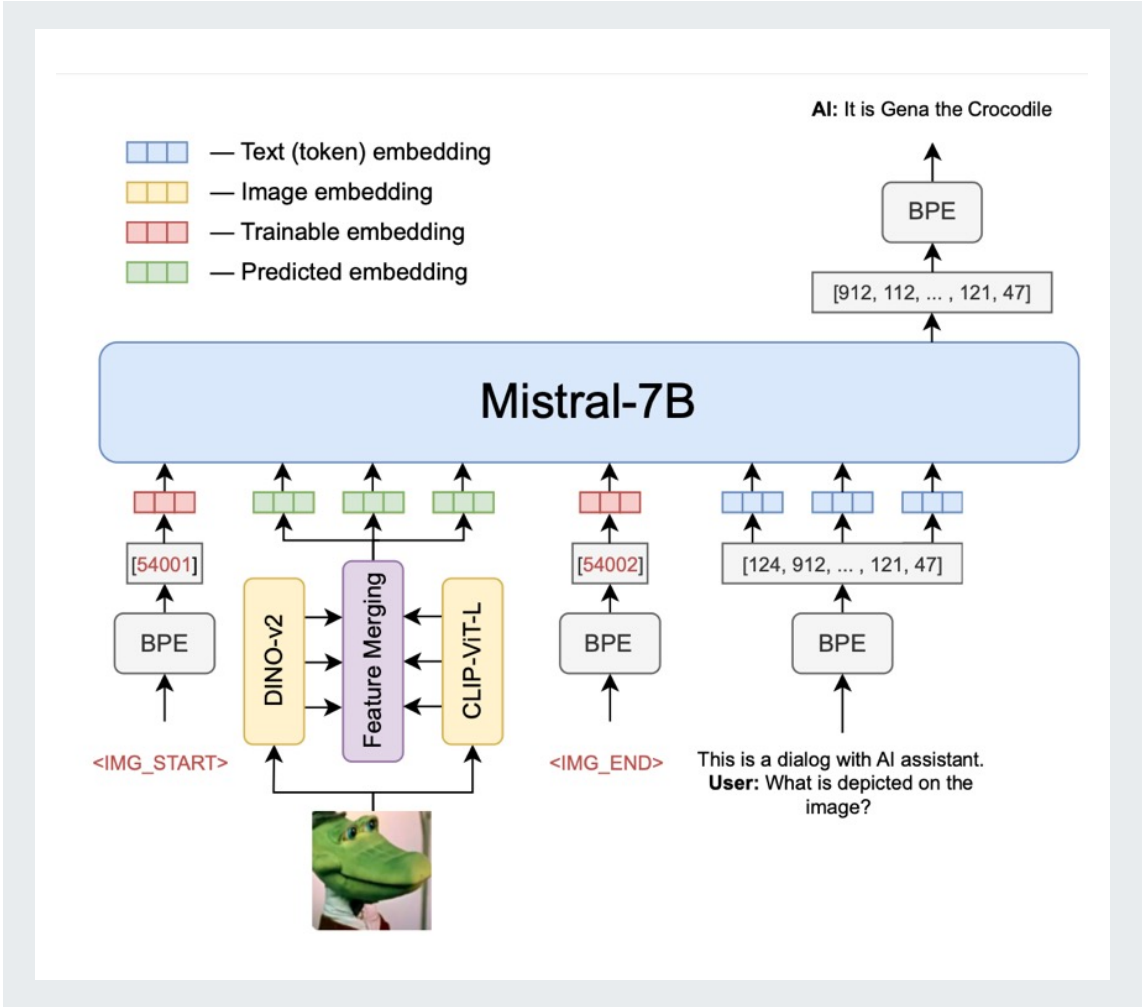
Omnis (lat.) — everything **Fusion** (en.) — merge

Key features

- **The first ru+en dialog multimodal model** in Russia
- Operates with **2 modalities**: text, images
- Can describe images and answer questions about them in various domains
- Ranked **#1 in Daily Papers** on Hugging Face on release day April 10, 2024
- Published in **open source**

LLM: Mistral, Phi, Qwen

Visual encoder: CLIP-ViT, InternViT, fusion of encoders



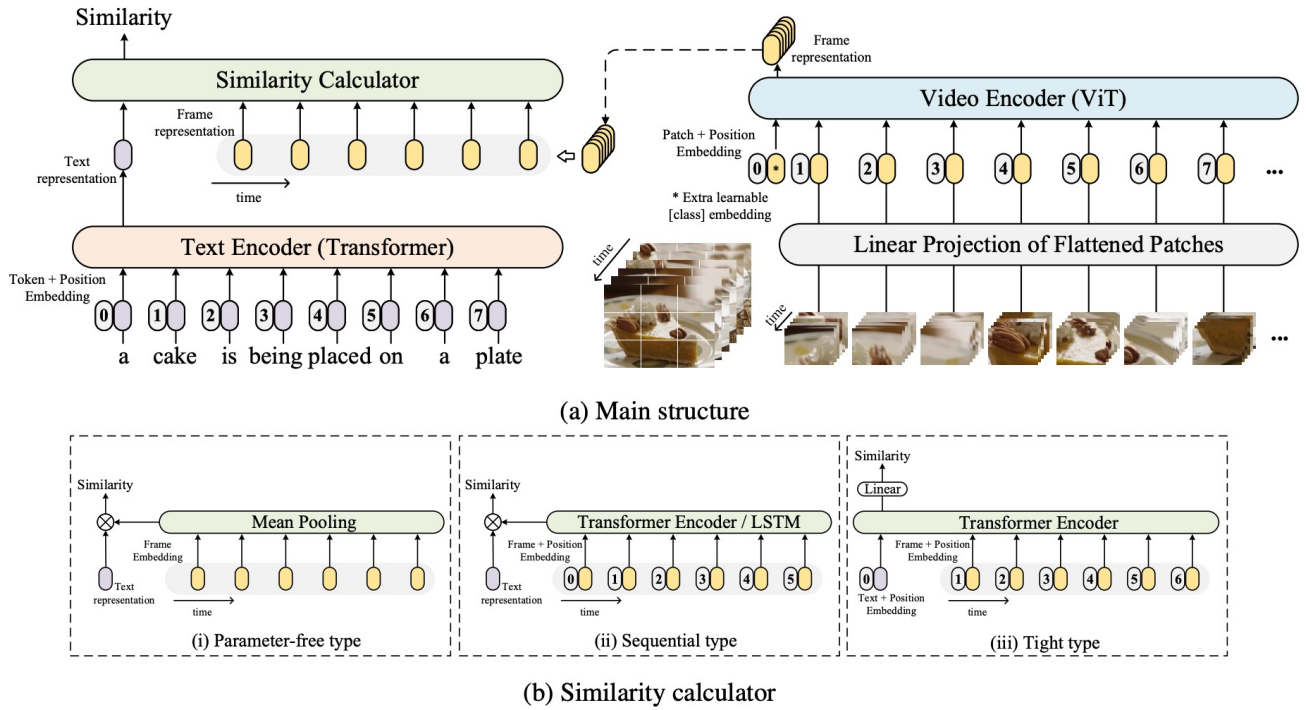
02

Video encoding and fusion with LLM

Spatio-temporal video encoding. CLIP4Clip

CLIP4Clip — CLIP For video Clip retrieval

- Main tasks: video-to-text (rank captions) and text-to-video (rank videos) retrieval
- Input video is sampled into frames
- Frames are reshaped into a sequence of flattened 2D patches
- Patches are mapped to the 1D sequence of embeddings and proceed to the image encoder (ViT-like)
- Similarity calculator predicts the score between the text representation and representation sequence of these frames

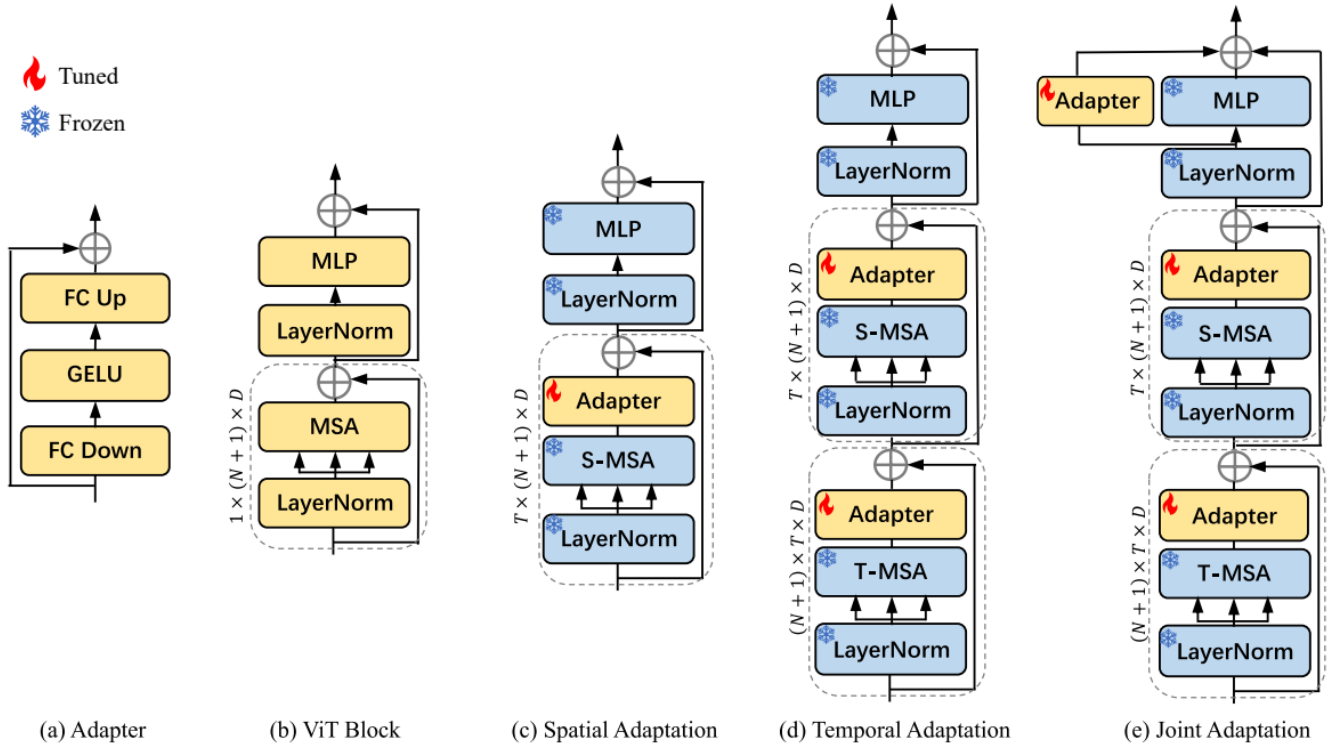


https://arxiv.org/abs/2104.08860

Spatio-temporal video encoding. AIM

AIM — Adapting Image Models

- Reuse the pre-trained self-attention layer in the image model to do temporal modeling
- S-MSA and T-MSA **share** weights, applied to different input dimensions
- Temporal adapter has **no** skip connection



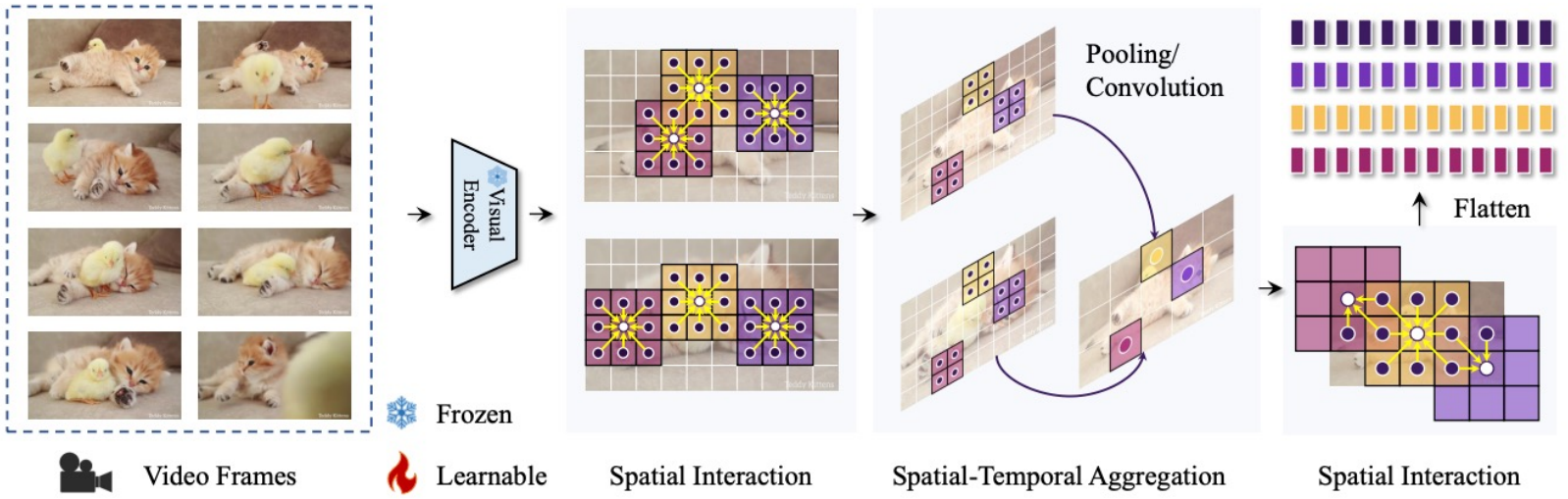
(a) Adapter (b) ViT Block (c) Spatial Adaptation (d) Temporal Adaptation (e) Joint Adaptation

https://arxiv.org/abs/2302.03024

Spatio-temporal video encoding. STC Connector

STC — Spatio-Temporal Convolution

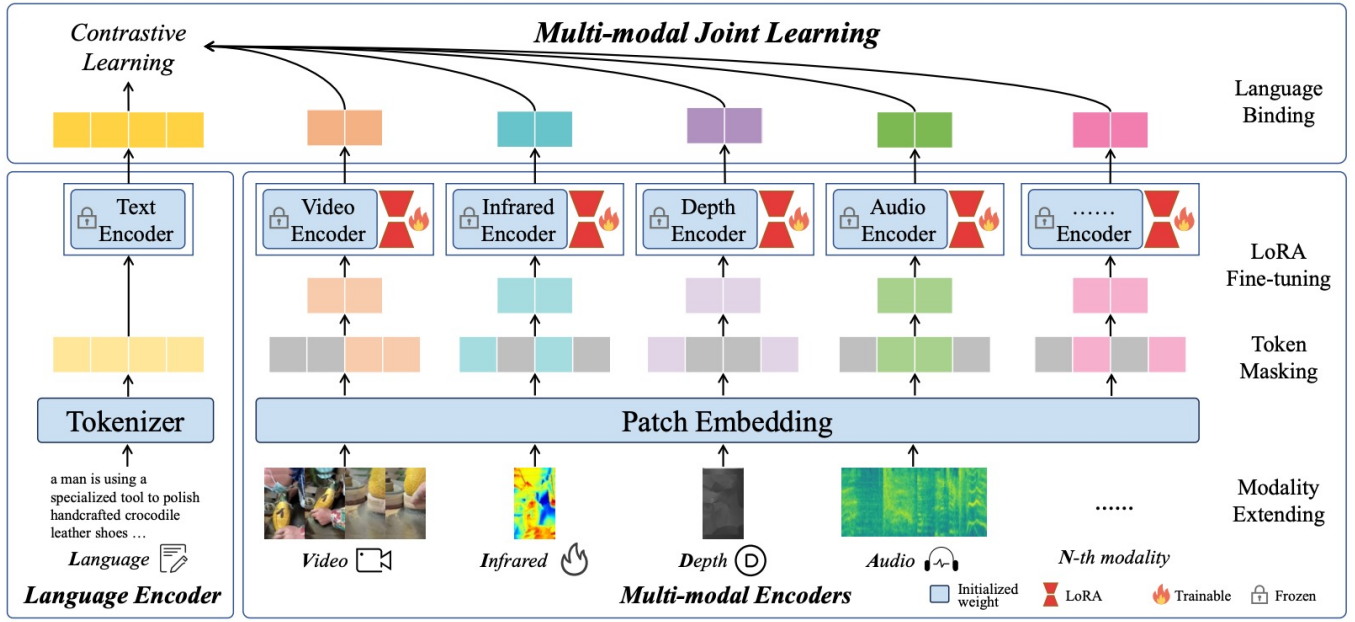
- Implemented in VideoLLaMA2 paper
- STC consists of **two** spatial interaction modules (RegStage adaptation) and **one** spatial-temporal aggregation module (3D convolution adaptation)
- Maintain the spatial-temporal order in the output visual tokens
- Reduce the number of spatial-temporal tokens
- Alleviate information loss during spatial-temporal downsampling



https://arxiv.org/abs/2406.07476

Multimodal encoders. LanguageBind

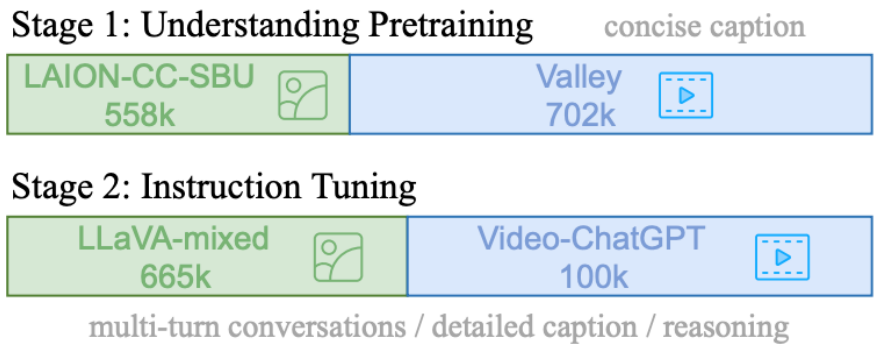
- Language encoder is frozen
- Use contrastive learning between language and other modalities (initialized from OpenCLIP)
- VIDAL-10M dataset, includes 3M *language-video* pairs
- Spatio-temporal video encoder (like in **AIM**)
- Training stages: token masking (Masked AutoEncoder strategy) and LoRA fine-tuning



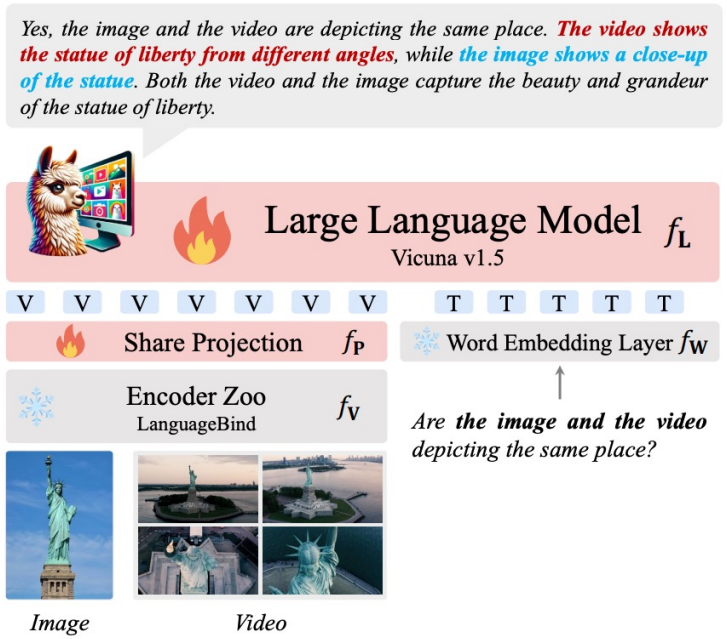
https://arxiv.org/abs/2310.01852

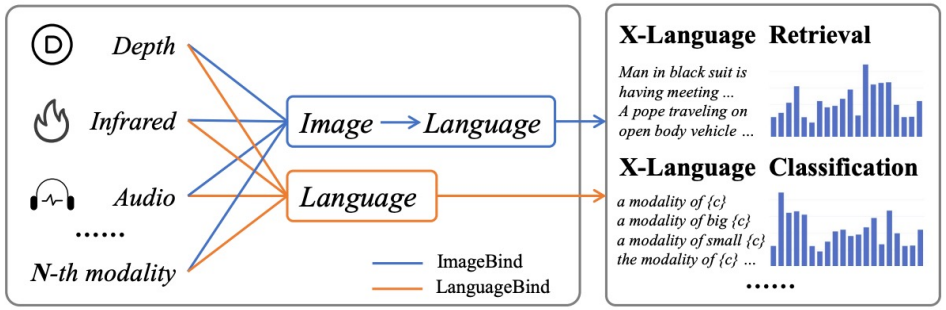
VideoLLaVA = fuse(LLM, video)

- VideoLLaVA — a pioneer work
- Operates both with image and video data
- Use LanguageBind as visual encoders
- Training details:
  - 224 x 224 resolution
  - 8 frames for a video
  - batch combines image and video



Training stages

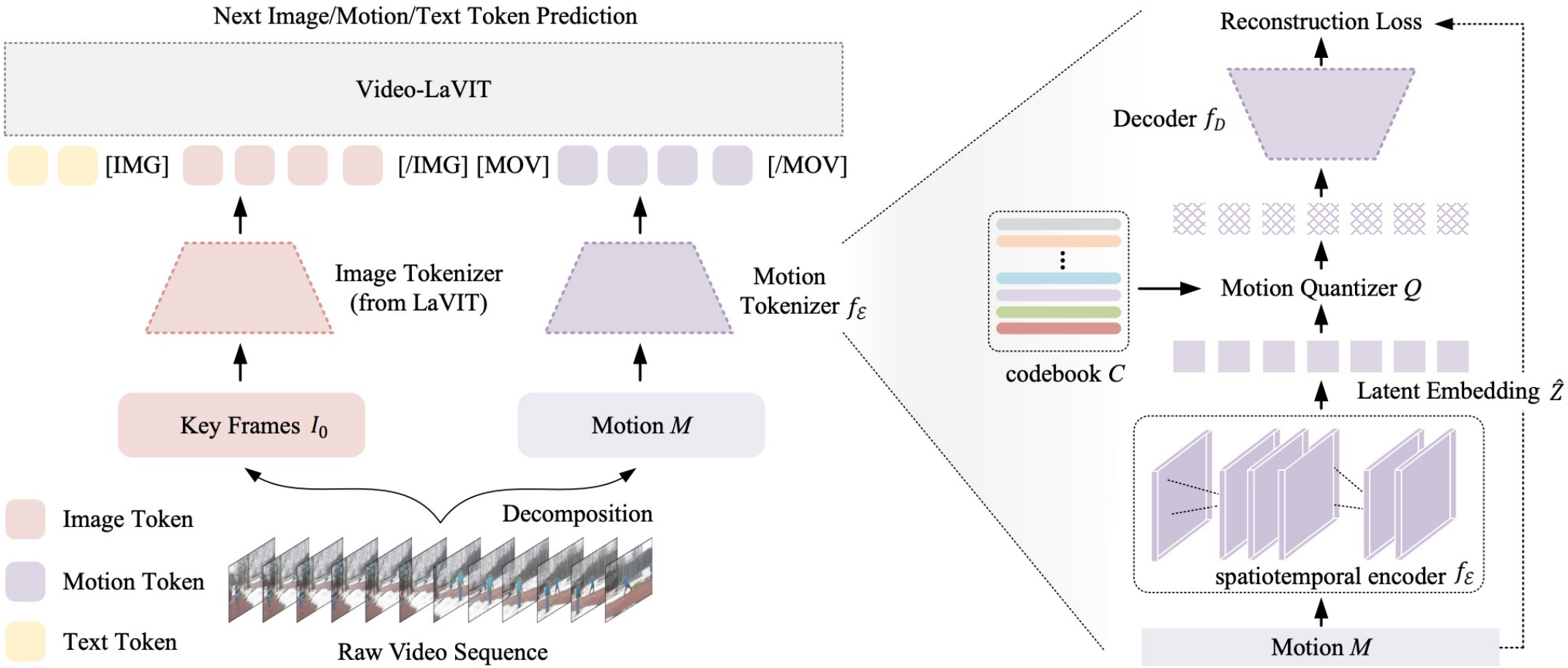




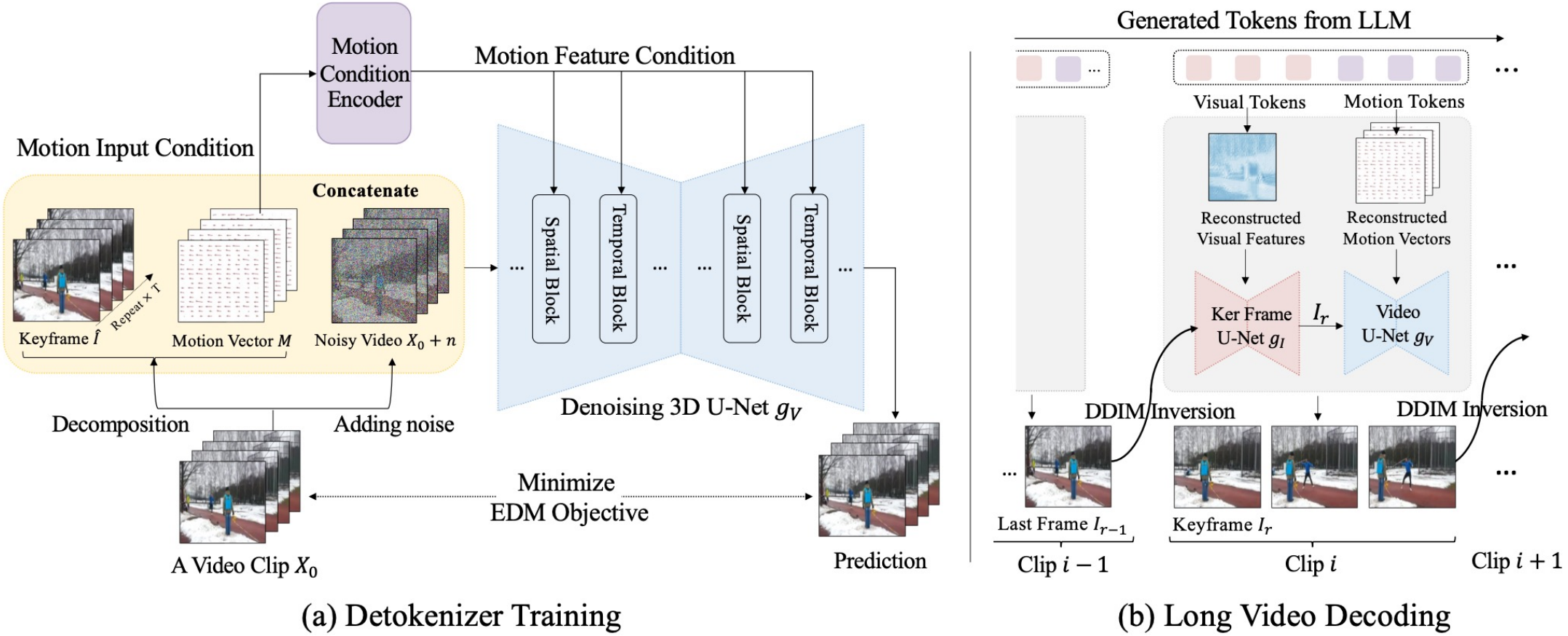
https://arxiv.org/abs/2311.10122

AIRI

Single model for video perception and generation (1)



Single model for video perception and generation (2)



Video modality perception. How about long context?

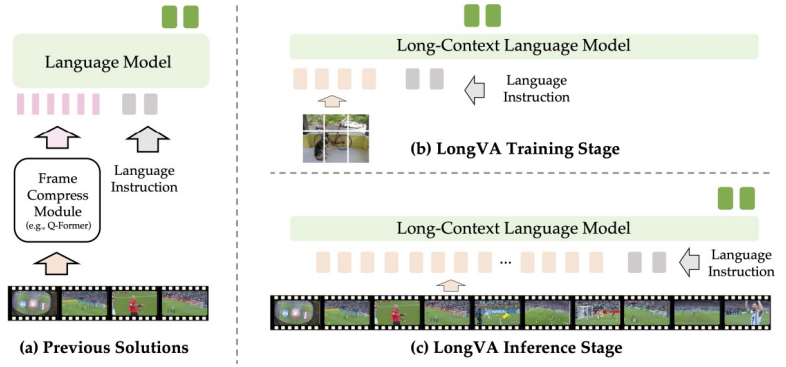
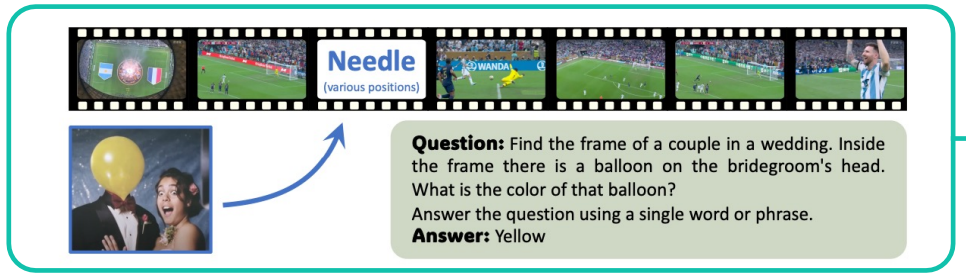
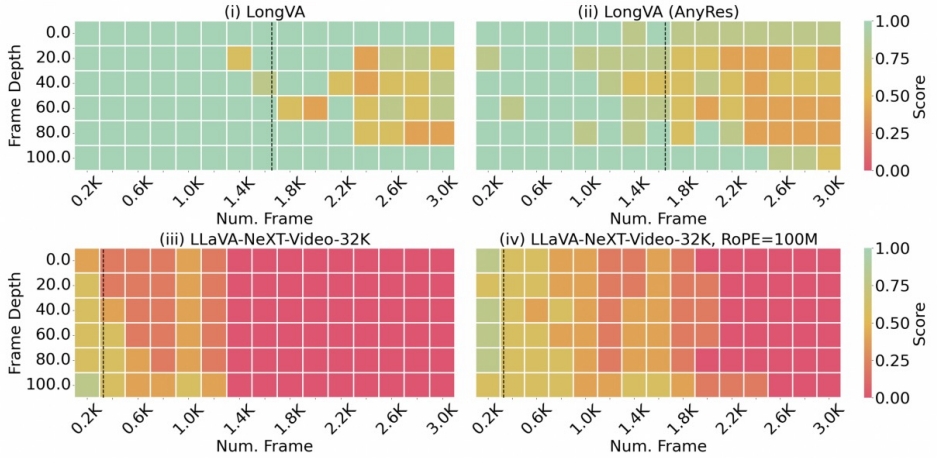


Figure 1: **Left:** to develop long vision models, previous studies proposed better visual resamplers to reduce the number of visual tokens. **Right:** LongVA approaches this problem from the angle of the language model. We leverage image data (short visual input) to align long-context LLM with vision. During the test time, LongVA can zero-shot process extremely long videos, thanks to the property of long context transfer.

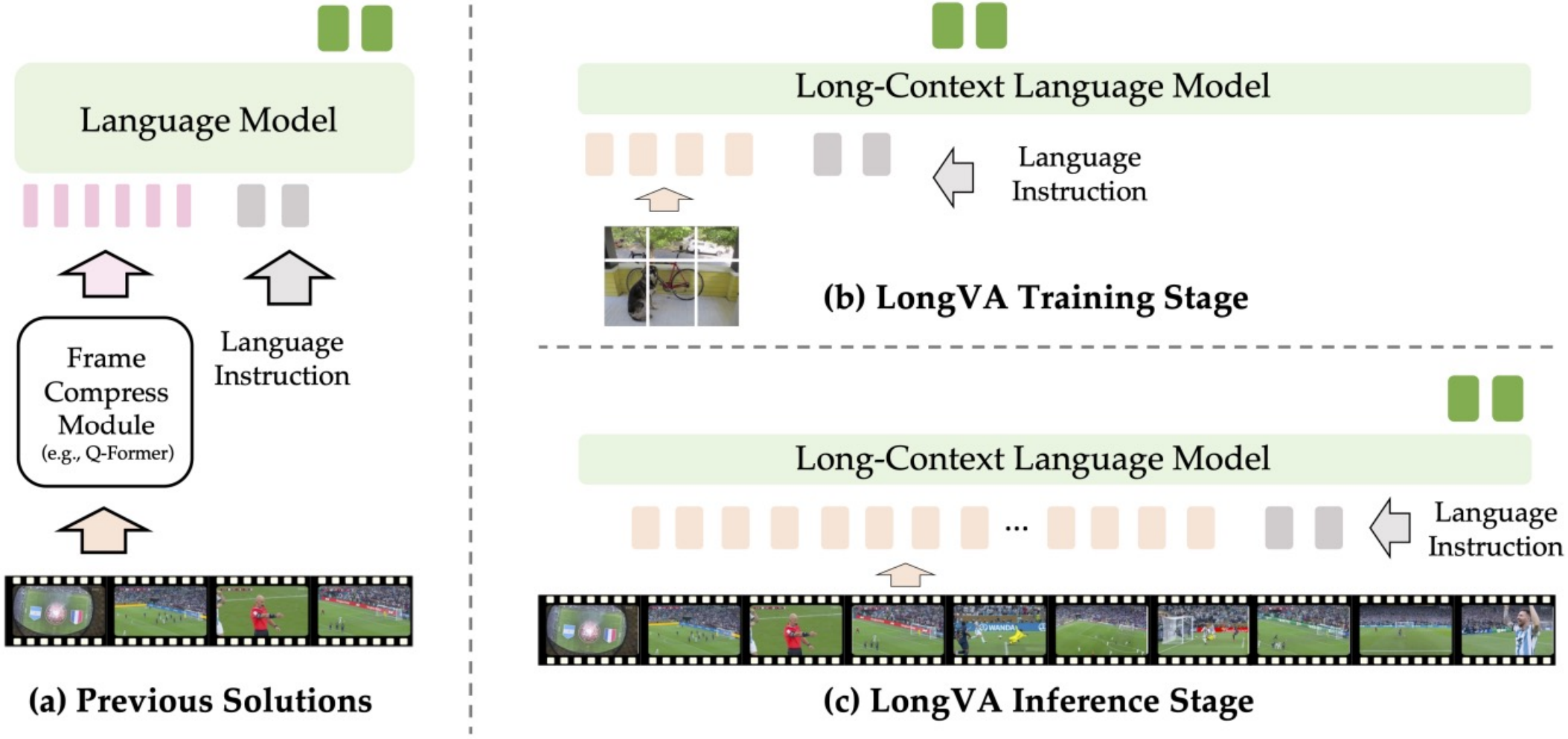
Model	Tokens/Frames*	Training Max Frames*	LM Backbone	LM Context Length
MPLUG-Owl-video [78]	256	4	LLaMA	4K
MovieChat [61]	32	8	Vicuna-v0	2K
Video-LLaVA [83]	49	8	Vicuna-1.5	4K
VideoChat [38]	32/196	8	Vicuna-v0	2K
LLaVA-NeXT-Video [85]	144	16	Vicuna-1.5	4K
ST-LLM [48]	256	16	Vicuna-1.1	2K
Video-LLaMA [15]	32	32	LLaMA-2	4K
Chat-UniVi [29]	112	64	Vicuna-1.5	4K
TimeChat [58]	4	96	LLaMA-2	4K
Video-ChatGPT [49]	256	100	Vicuna-1.1	2K
LLaMA-VID [40]	2	300	Vicuna-1.5	4K
LongVA (Ours)	144	-	Qwen2-Extended	224K+



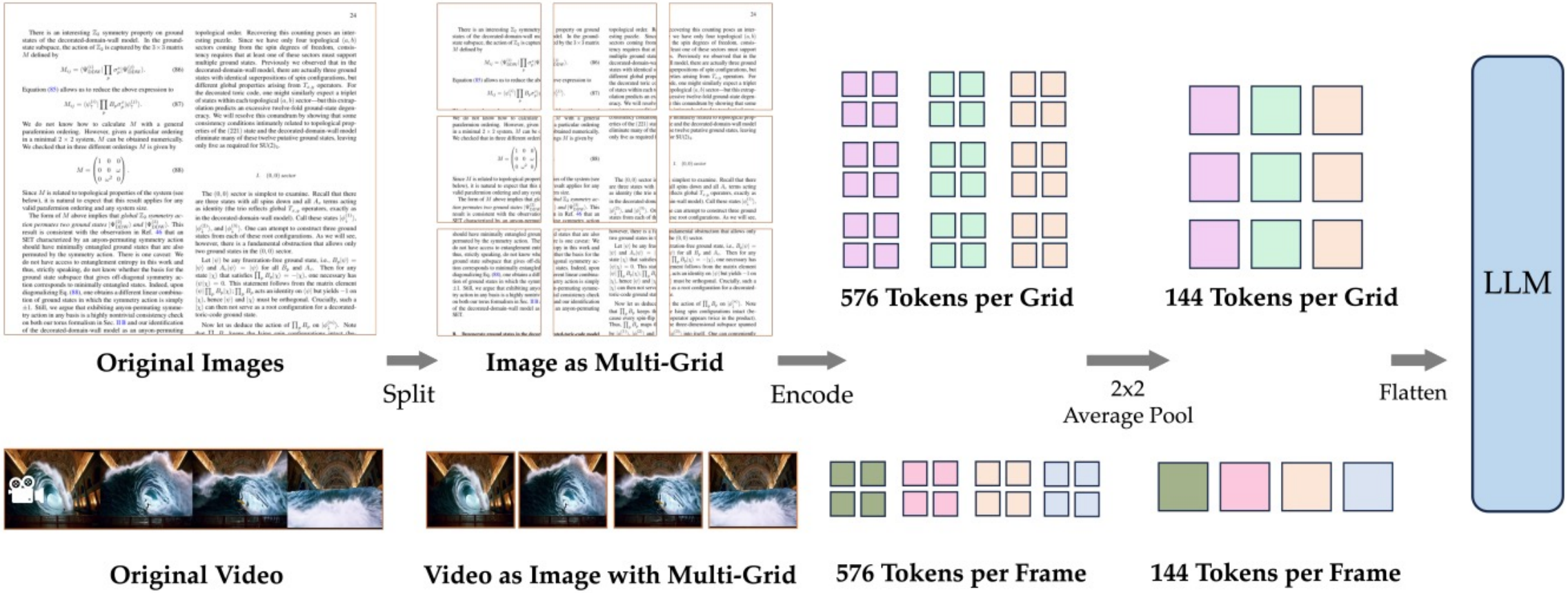


https://github.com/EvolvingLMMs-Lab/LongVA

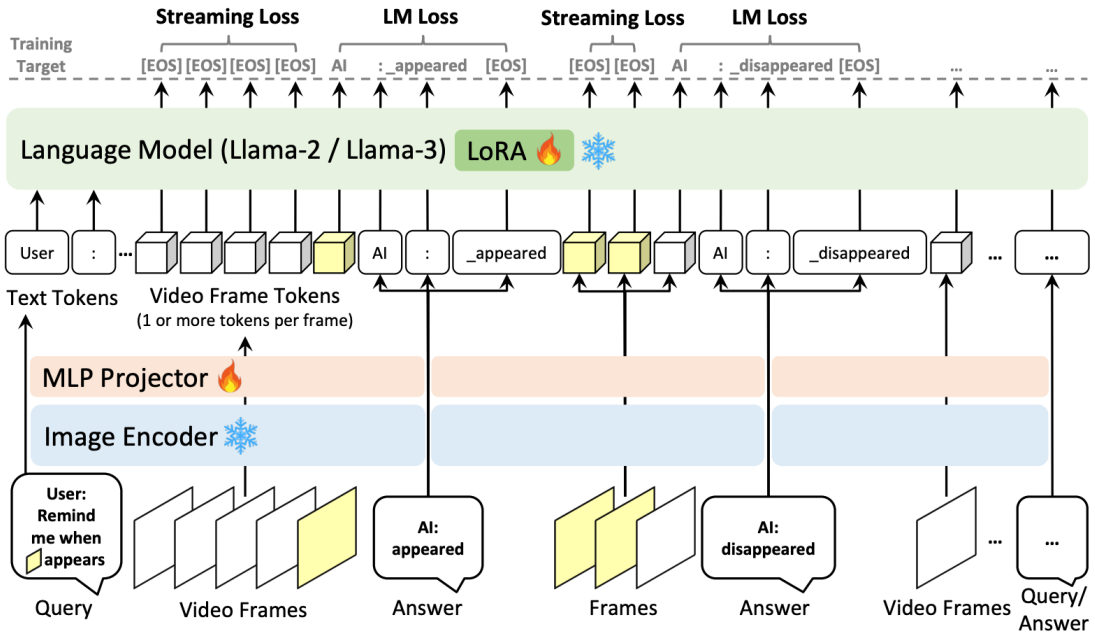
Video modality perception. How about long context?

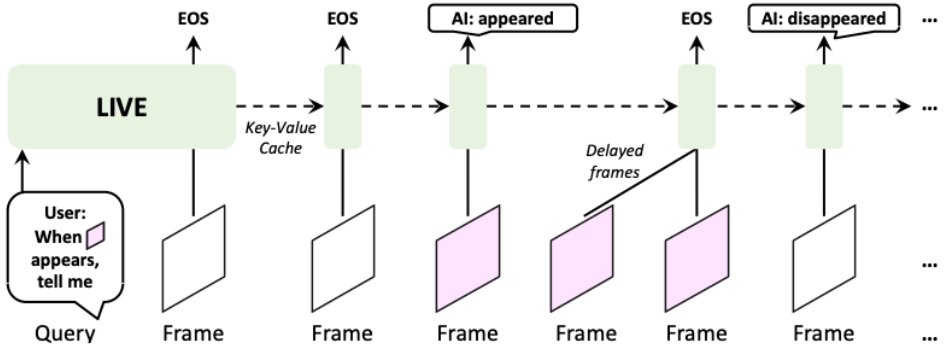


Video modality perception. How about long context?



Video modality perception. Streaming video





Training

- Image encoder — CLIP ViT-L (307M), 2 FPS
- Emb. shape — $(1 + h_p \times w_p) \times c$
  (CLS + avg. pooled spatial tokens)

$$L = \frac{1}{N}\sum_{j=1}^{N}(\underbrace{-\log l_{j+1} P_j^{[\mathrm{Txt}_{j+1}]}}_{LM\,Loss} \underbrace{- w \log f_j P_j^{[\mathrm{EOS}]}}_{Streaming\,Loss})$$

Inference

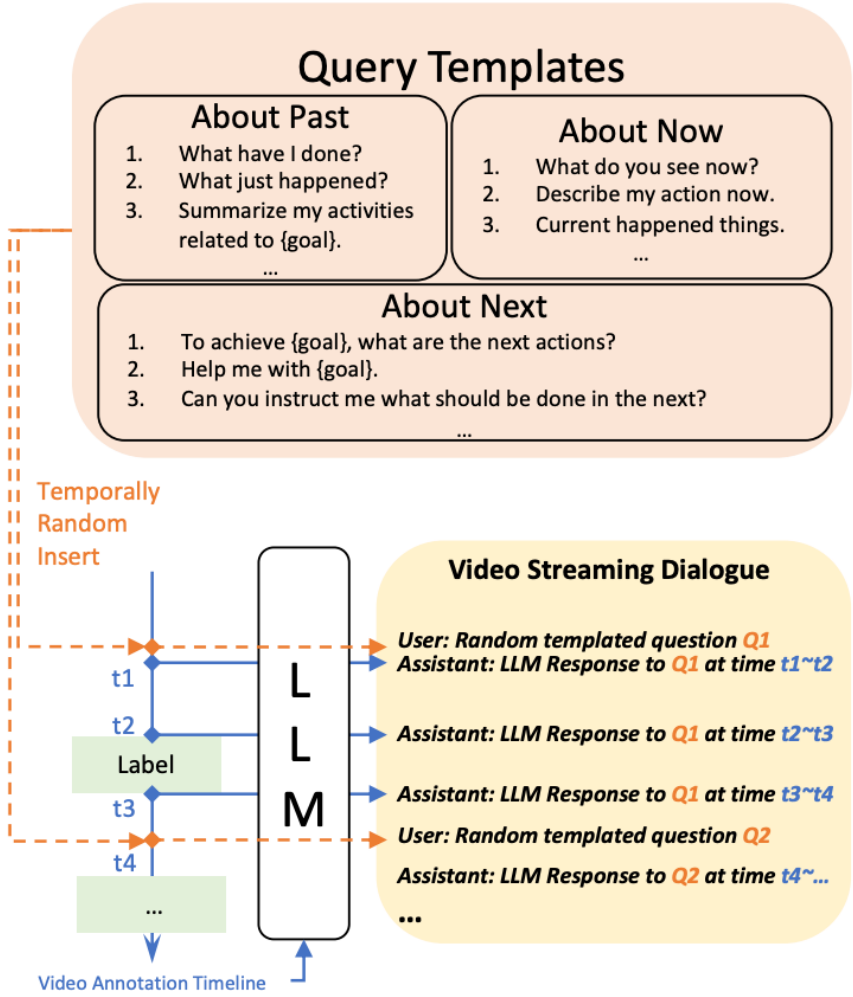
- Continuous KV cache as the input progresses to speed up the inference
- Video frame tokens can be always encoded and buffered — no need to wait the language decoding

https://arxiv.org/pdf/2406.11816

Video modality perception. Streaming video

Data annotation pipeline

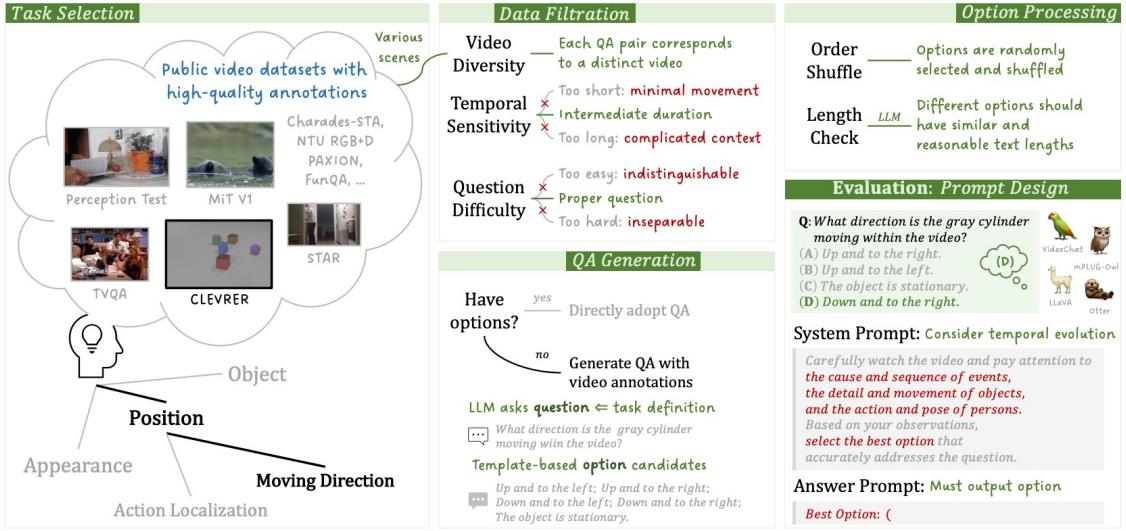
- Prepare a question template library containing various queries about the past, present, and future tenses of the video (Q_i)
- Obtain the video annotation timeline (A) from the offline dataset (e.g., "time t_a~t_b: boiling the water")
- Prompt the LLM to generate responses at every critical timestamp, according to Q_i and A
- During training:
  1. Randomly sample a query and load its responses at critical timestamps
  2. Randomly insert a query into a video timestamp t_r
  3. Discard the responses that occur before t_r, and add a response at t_r



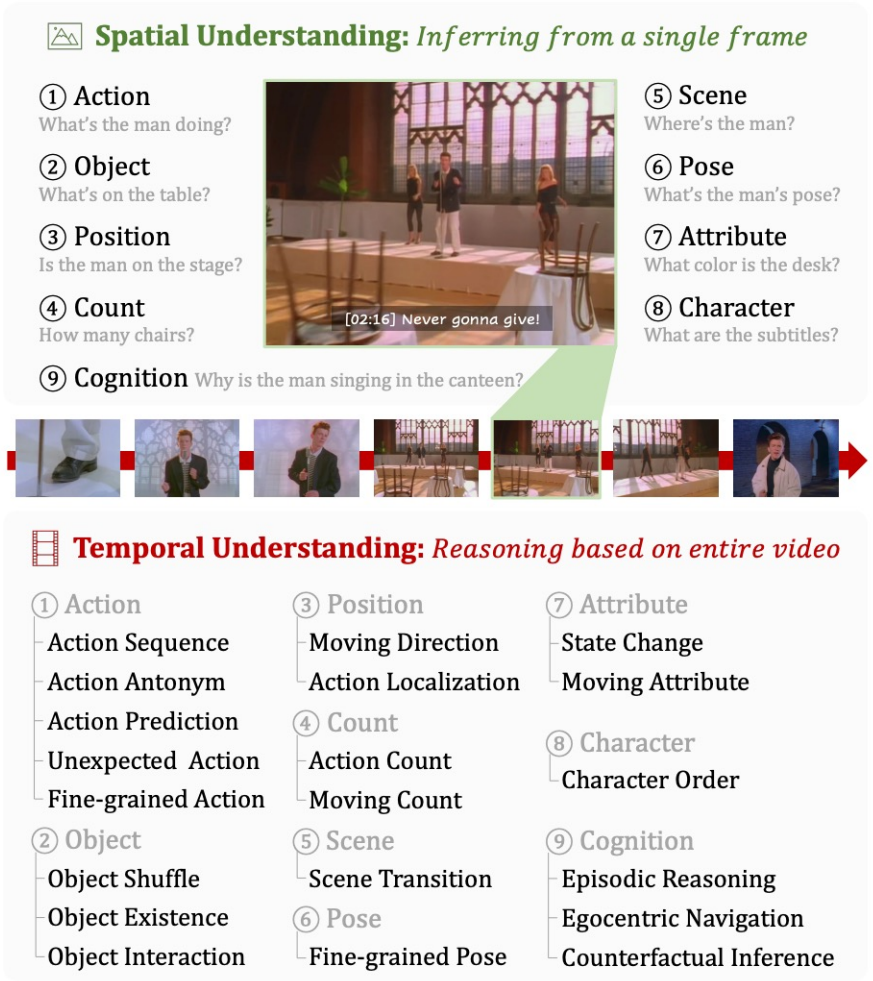
https://arxiv.org/pdf/2406.11816

Benchmarks. MVBench

- 20 challenging tasks
- Cannot be solved by single frame analysis
  ("What direction is the man moving?")



Data generation pipeline



https://arxiv.org/abs/2311.17005

Benchmarks. MV-Bench

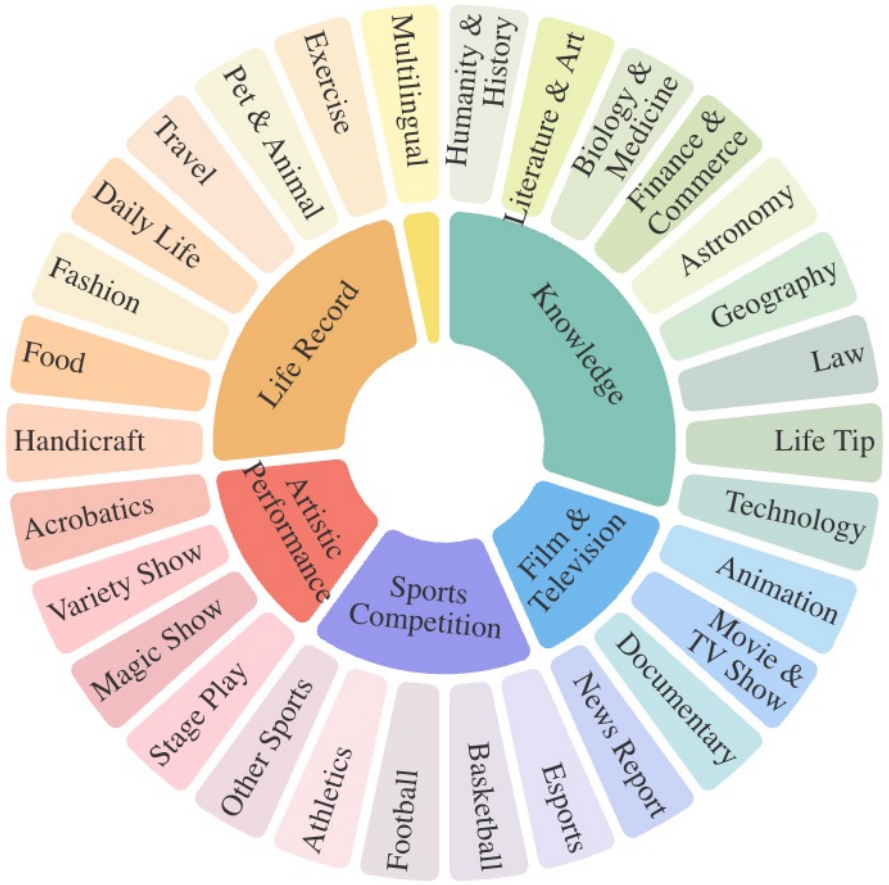
- Tasks are grouped by: **Action**, **Object**, **Position**, **Scene**, **Count**, **Attribute**, **Pose**, **Character**

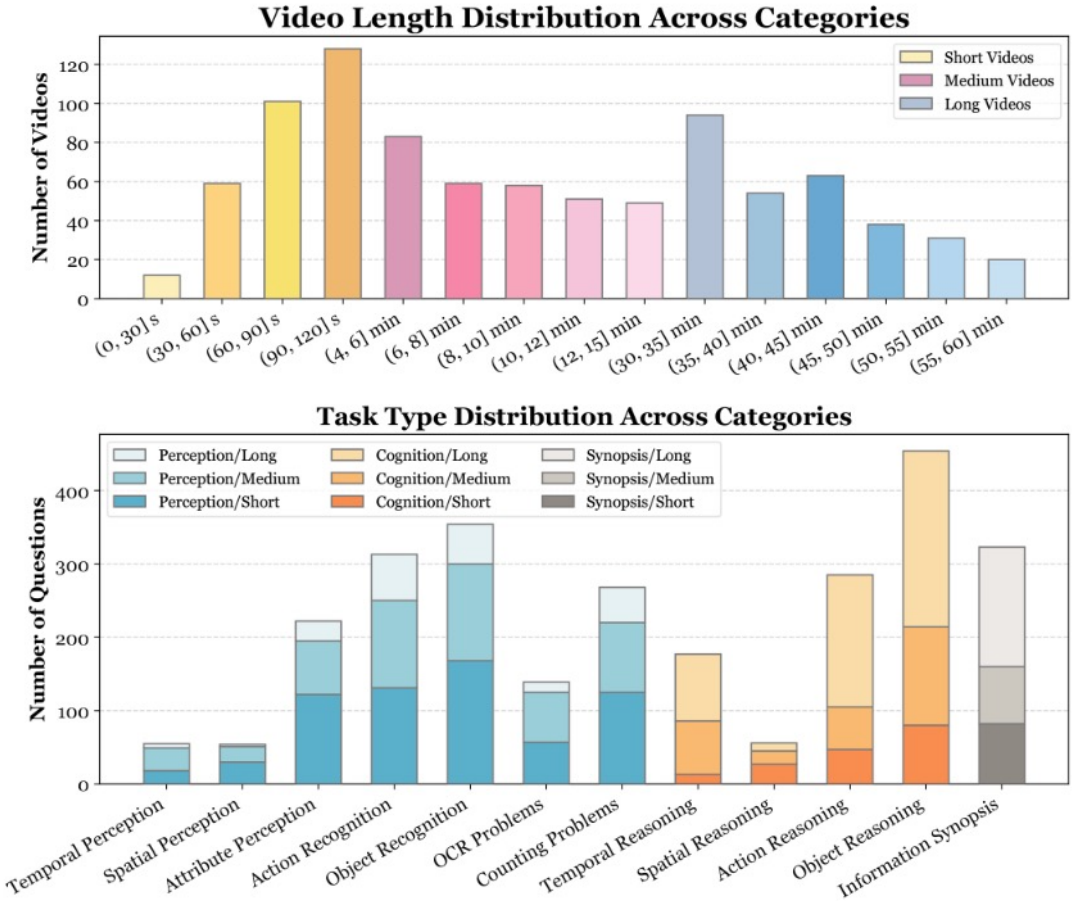
Model	LLM	Avg	AS	AP	AA	FA	UA	OE	OI	OS	MD	AL	ST	AC	MC	MA	SC	FP	CO	EN	ER	CI
Random	-	27.3	25.0	25.0	33.3	25.0	25.0	33.3	25.0	33.3	25.0	25.0	25.0	33.3	25.0	33.3	33.3	25.0	33.3	25.0	20.0	30.9
Image MLLMs: Following [11], all models take 4 frames as input, with the output embeddings concatenated before feeding into the LLM.																						
mPLUG-Owl-I [87]	LLaMA-7B	29.4	25.0	20.0	44.5	27.0	23.5	36.0	24.0	34.0	23.0	24.0	34.5	34.5	22.0	31.5	40.0	24.0	37.0	25.5	21.0	37.0
LLaMA-Adapter [96]	LLaMA-7B	31.7	23.0	28.0	51.0	30.0	33.0	53.5	32.5	33.5	25.5	21.5	30.5	29.0	22.5	41.5	39.5	25.0	31.5	22.5	28.0	32.0
BLIP2 [37]	FlanT5-XL	31.4	24.5	29.0	33.5	17.0	42.0	51.5	26.0	31.0	25.5	26.0	32.5	25.5	30.0	40.0	42.0	27.0	30.0	26.0	37.0	31.0
Otter-I [36]	MPT-7B	33.5	34.5	32.0	39.5	30.5	38.5	48.5	44.0	29.5	19.0	25.5	55.0	20.0	32.5	28.5	39.0	28.0	27.0	32.0	29.0	36.5
MiniGPT-4 [97]	Vicuna-7B	18.8	16.0	18.0	26.0	21.5	16.0	29.5	25.5	13.0	11.5	12.0	9.5	32.5	15.5	8.0	34.0	26.0	29.5	19.0	9.9	3.0
InstructBLIP [11]	Vicuna-7B	32.5	20.0	16.5	46.0	24.5	46.0	51.0	26.0	37.5	22.0	23.0	46.5	**42.5**	26.5	40.5	32.0	25.5	30.0	25.5	30.5	38.0
LLaVA [44]	Vicuna-7B	36.0	28.0	39.5	63.0	30.5	39.0	53.0	41.0	41.5	23.0	20.5	45.0	34.0	20.5	38.5	47.0	25.0	36.0	27.0	26.5	42.0
Video MLLMs: All models take 16 frames as input, with the exception of VideoChatGPT, which uses 100 frames.																						
Otter-V [36]	LLaMA-7B	26.8	23.0	23.0	27.5	27.0	29.5	53.0	28.0	33.0	24.5	23.5	27.5	26.0	28.5	18.0	38.5	22.0	22.0	23.5	19.0	19.5
mPLUG-Owl-V [87]	LLaMA-7B	29.7	22.0	28.0	34.0	29.0	29.0	40.5	27.0	31.5	**27.0**	23.0	29.0	31.5	27.0	40.0	44.0	24.0	31.0	26.0	20.5	29.5
VideoChatGPT [48]	Vicuna-7B	32.7	23.5	26.0	62.0	22.5	26.5	54.0	28.0	40.0	23.0	20.0	31.0	30.5	25.5	39.5	**48.5**	29.0	33.0	29.5	26.0	35.5
VideoLLaMA [94]	Vicuna-7B	34.1	27.5	25.5	51.0	29.0	39.0	48.0	40.5	38.0	22.5	22.5	43.0	34.0	22.5	32.5	45.5	32.5	40.0	30.0	21.0	37.0
VideoChat [39]	Vicuna-7B	35.5	33.5	26.5	56.0	33.5	40.5	53.0	40.5	30.0	25.5	27.0	48.5	35.0	20.5	42.5	46.0	26.5	41.0	23.5	23.5	36.0
VideoChat2$_{\text{text}}$	Vicuna-7B	34.7	24.5	27.0	49.5	27.0	38.0	53.0	28.0	40.0	25.5	27.0	38.5	41.5	27.5	32.5	46.5	26.5	36.0	33.0	32.0	40.0
VideoChat2	Vicuna-7B	**51.1**	**66.0**	47.5	**83.5**	**49.5**	60.0	**58.0**	**71.5**	**42.5**	23.0	23.0	**88.5**	39.0	**42.0**	**58.5**	44.0	**49.0**	36.5	**35.0**	40.5	**65.5**

https://arxiv.org/abs/2311.17005

AIRI

Benchmarks. Video-MME

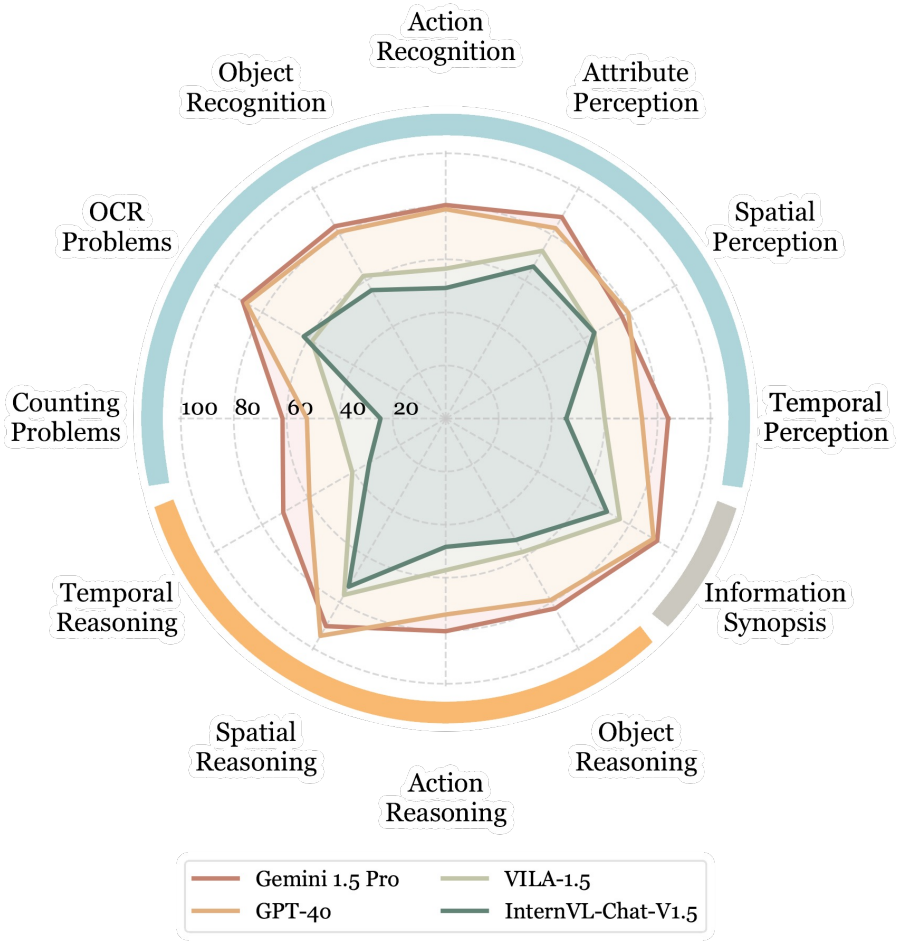




https://arxiv.org/abs/2405.21075

Benchmarks. Video-MME

Models	LLM Params	Short (%)		Medium (%)		Long (%)		Overall (%)	
		w/o subs	w/ subs	w/o subs	w/ subs	w/o subs	w/ subs	w/o subs	w/ subs
Open & Closed-source Image MLLMs									
Qwen-VL-Chat [5]	7B	46.9	47.3	38.7	40.4	37.8	37.9	41.1	41.9
Qwen-VL-Max [5]	-	55.8	57.6	49.2	48.9	**48.9**	**47.0**	**51.3**	51.2
InternVL-Chat-V1.5 [9]	20B	**60.2**	**61.7**	**46.4**	**49.1**	45.6	46.6	50.7	**52.4**
Open-source Video MLLMs									
Video-LLaVA [30]	7B	45.3	46.1	38.0	40.7	36.2	38.1	39.9	41.6
ST-LLM [36]	7B	45.7	48.4	36.8	41.4	31.3	36.9	37.9	42.3
VideoChat2-Mistral [24]	7B	48.3	52.8	37.0	39.4	33.2	39.2	39.5	43.8
Chat-UniVi-V1.5 [19]	7B	45.7	51.2	40.3	44.6	35.8	41.8	40.6	45.9
LLaVA-NeXT-Video [74]	34B	61.7	65.1	50.1	52.2	44.3	47.2	52.0	54.9
VILA-1.5 [31]	34B	**68.1**	**68.9**	**58.1**	**57.4**	**50.8**	**52.0**	**59.0**	**59.4**
Closed-source MLLMs									
GPT-4V [48]	-	70.5	73.2	55.8	59.7	53.5	56.9	59.9	63.3
GPT-4o [49]	-	80.0	82.8	70.3	76.6	65.3	72.1	71.9	77.2
Gemini 1.5 Flash [54]	-	78.8	79.8	68.8	74.7	61.1	68.8	70.3	75.0
Gemini 1.5 Pro [54]	-	**81.7**	**84.5**	**74.3**	**81.0**	**67.4**	**77.4**	**75.0**	**81.3**



https://arxiv.org/abs/2405.21075

03

Conclusion + 

- Video is one of the most complex and informative modalities
- Video + LLM -> New foundation model
- Video comprehension/perception in multimodal paradigm leads to the world model design
- Long video context analysis is a strong challenge for 2025
- Video latent representation encoders can be useful for improving video retrieval tasks

AIJ 2024 Contest. Emotional FusionBrain

Brief info

- Task — develop a **universal multimodal model** for **video question answering** and **video captioning**
- Three input modalities: **video, audio and text**
- Focuses on videos with **social interactions** between people and **human emotions** evaluation

Awards pool

- 1st place – RUB 1,000,000
- 2nd place – RUB 700,000
- 3rd place – RUB 400,000

DS Works

GitHub

Habr

AIJ 2024 Contest. Emotional FusionBrain

Video Question Answering

- Requires answering a question with a multiple-choice answer based on the video content

```
'task_id': 1,
'task_type': 'qa',
'question': 'How did the woman emotionally act in response to the
ringing of her mobile phone?',
'video': 'path_to_video.mp4',
'audio': 'path_to_audio.mp3',
'choices': [
{'choice_id': 0, 'choice': She was happy'},
{'choice_id': 1, 'choice': 'She was upset'},
{'choice_id': 2, 'choice': 'She was angry'},
{'choice_id': 3, 'choice': 'She was irritated'},
{'choice_id': 4, 'choice': 'Not enough information to answer'}]
```

Video Captioning

- Requires a detailed description of the video

```
'task_id': 2,
'task_type': 'captioning',
'question': 'Describe this video in detail.',
'video': 'path_to_video.mp4',
'audio': '',
'choices': []
```

AIRI

Contacts

Contacts

PhD,
Head of FusionBrain Lab, AIRI
Executive director on data science, Sber AI
kuznetsov@airi.net

@KUZNETSOFF87

@COMPLETE_AI