AIJ Artificial Intelligence Journey

Emotional FusionBrain 4.0

26.09.2024

Irina Abdullaeva

Multimodality + Socialization = AI Assistant

Why we need multimodality?

Why is LLM popular?

→ Text – basis of communication

→ Text – main source of the world knowledge

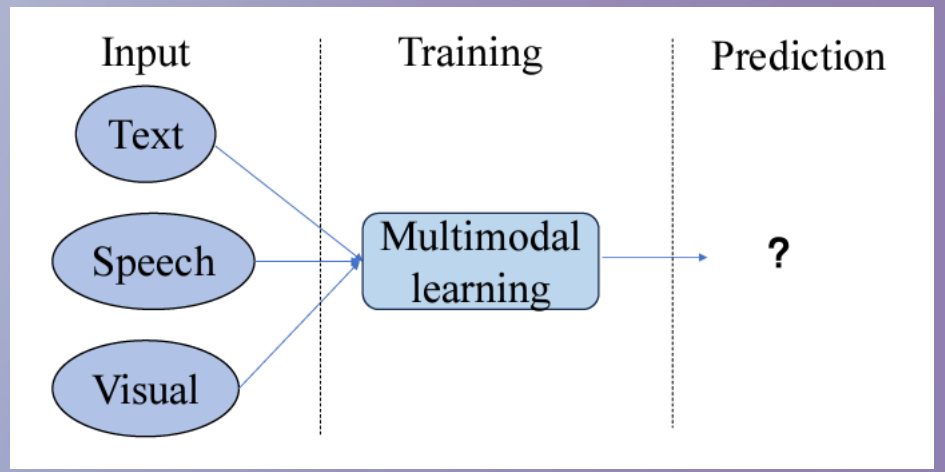
→ Chatbots primarily need text

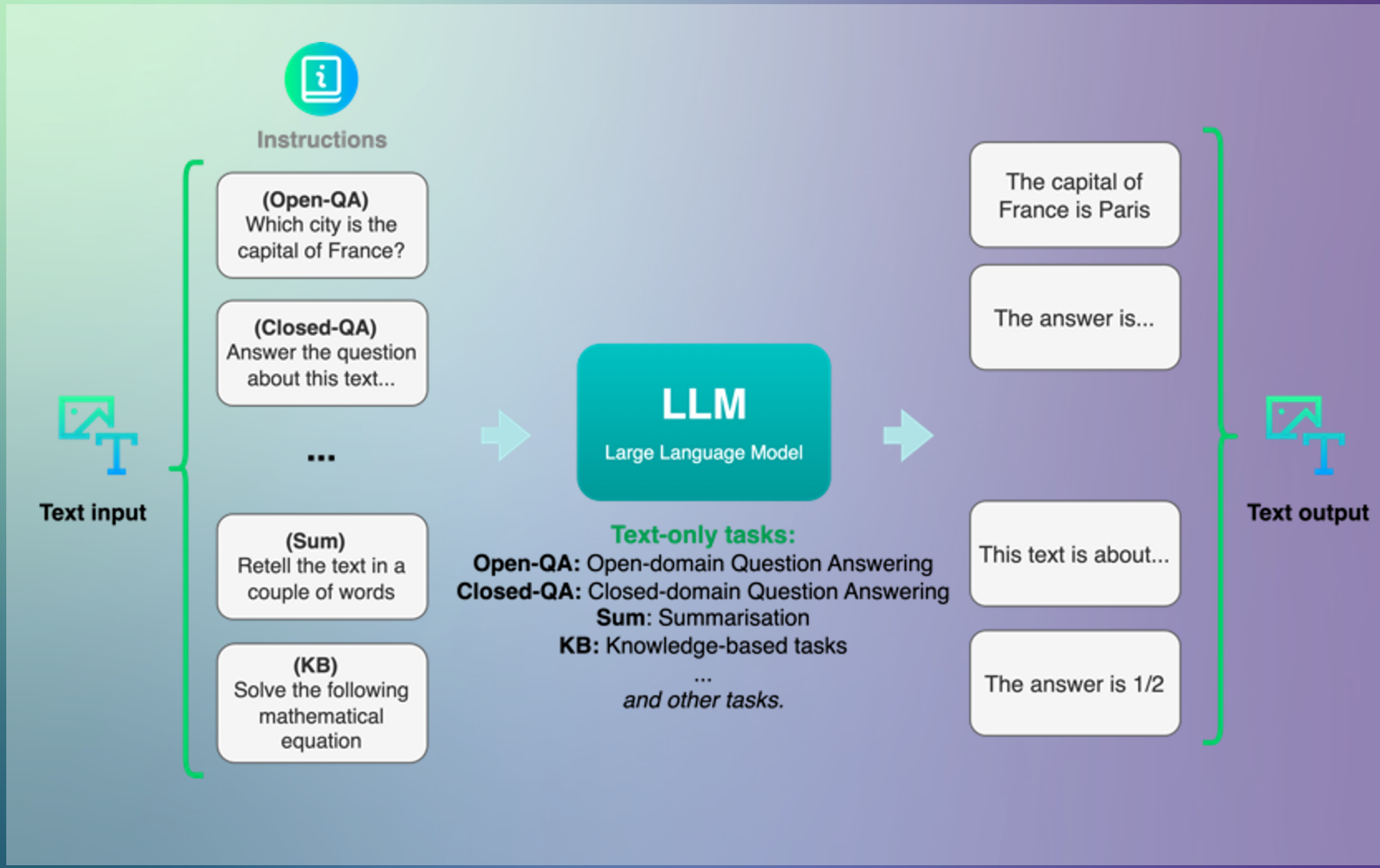
Why do we need something more than text?

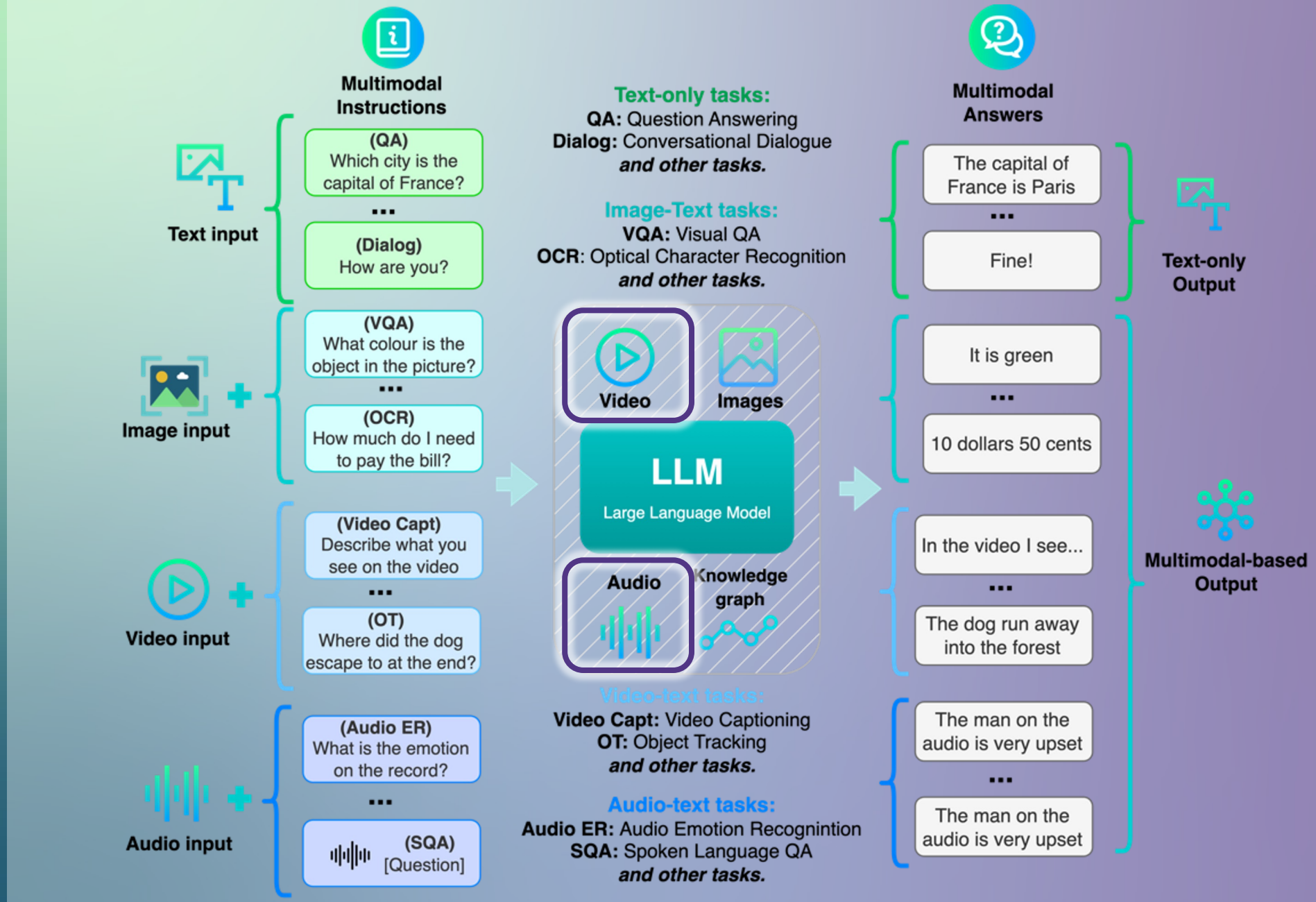
→ People are used to communicate using **more than just text**

→ To design a high-quality AI assistant, you need to extend perception abilities far beyond text

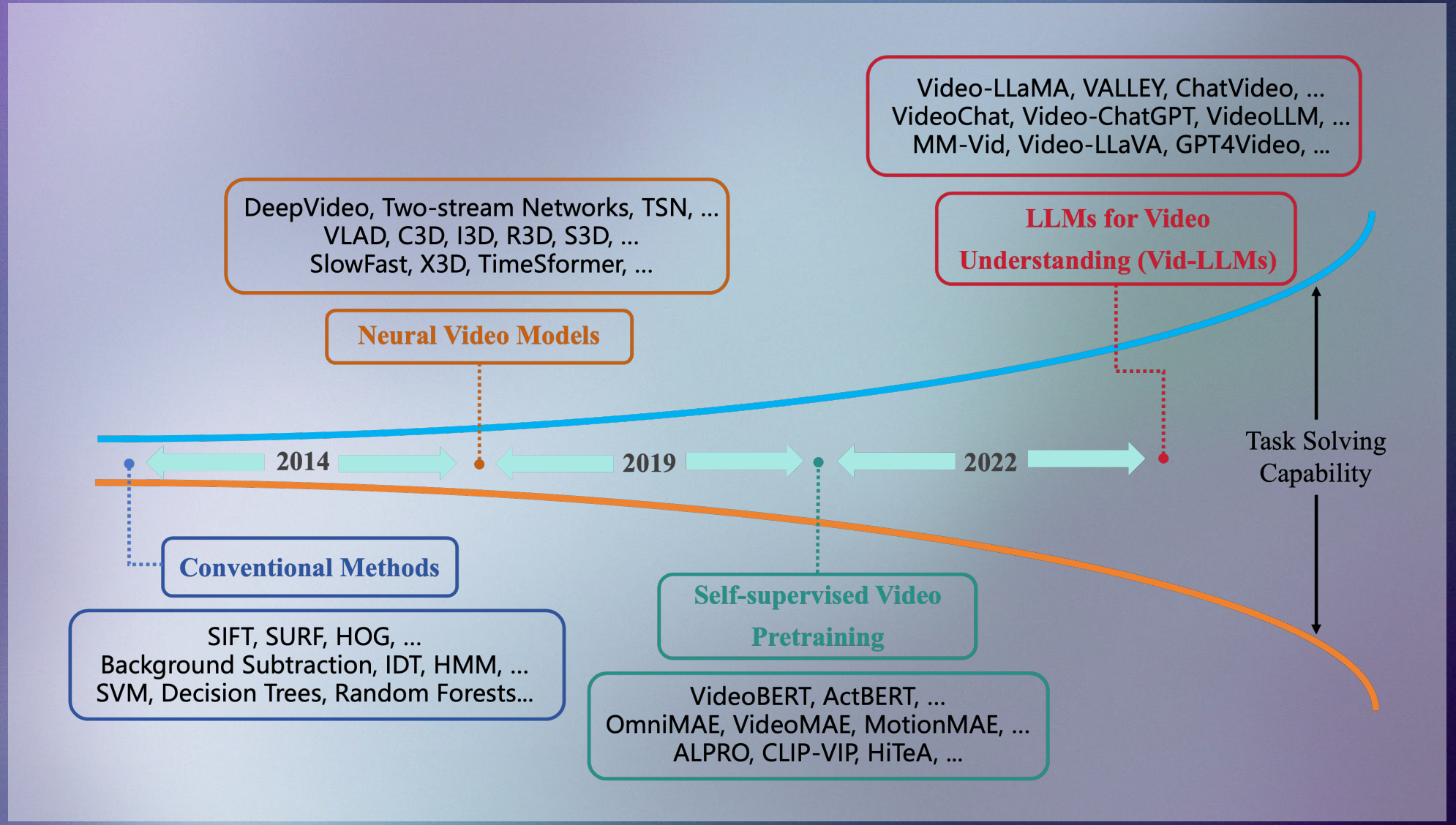
→ Text + images + sound + **video** + ... — this is the perfect NATURAL scenario for interaction




Instructions
(Open-QA)
Which city is the capital of France?
(Closed-QA)
Answer the question about this text...
...
(Sum)
Retell the text in a couple of words
(KB)
Solve the following mathematical equation
Text input
LLM
Large Language Model
Text-only tasks:
Open-QA: Open-domain Question Answering
Closed-QA: Closed-domain Question Answering
Sum: Summarisation
KB: Knowledge-based tasks
...
and other tasks.
The capital of France is Paris
The answer is...
This text is about...
The answer is 1/2
Text output

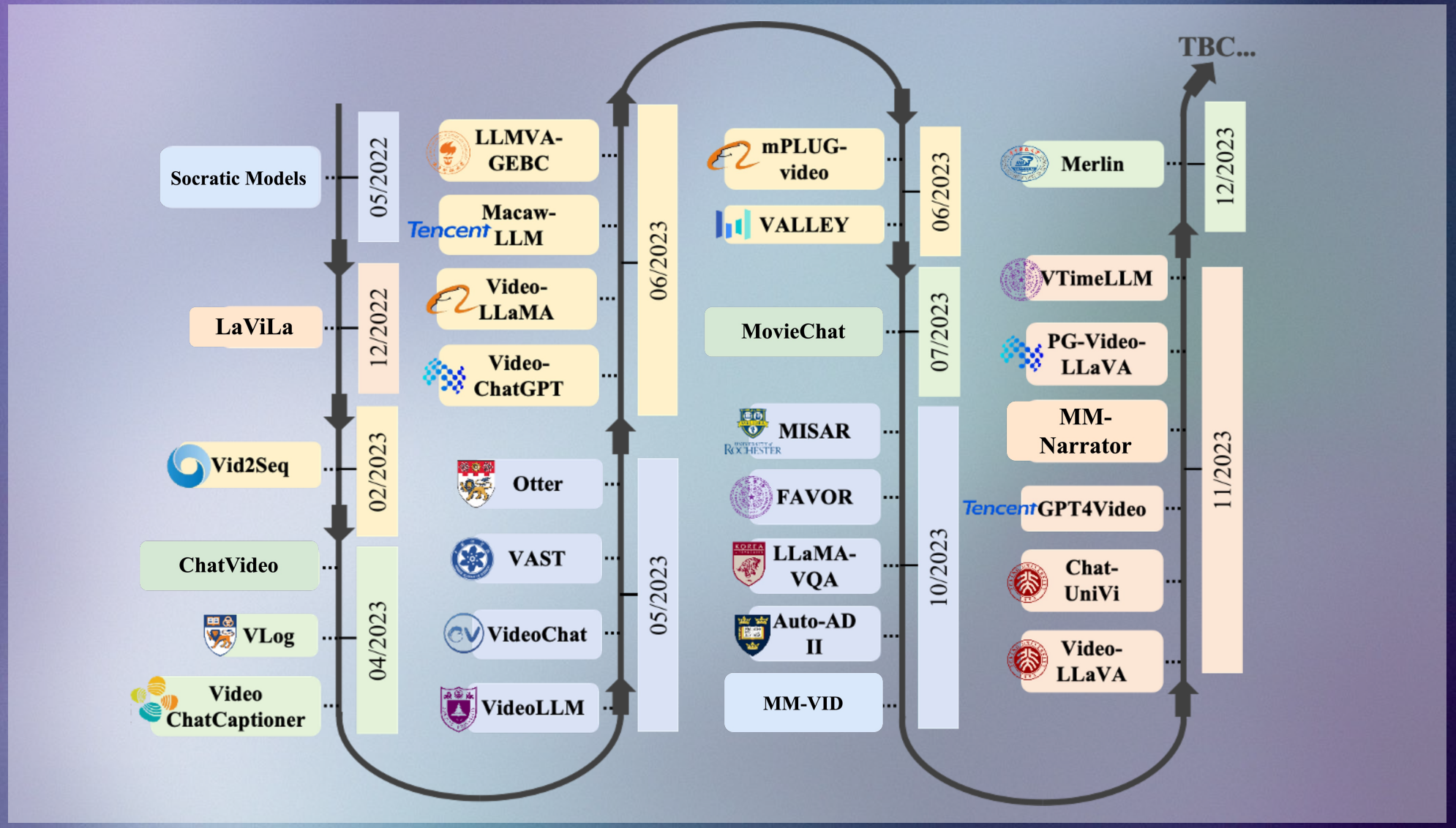

Multimodal Instructions
Text input
(QA) Which city is the capital of France?
...
(Dialog) How are you?
Image input
(VQA) What colour is the object in the picture?
...
(OCR) How much do I need to pay the bill?
Video input
(Video Capt) Describe what you see on the video
...
(OT) Where did the dog escape 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 Models Development

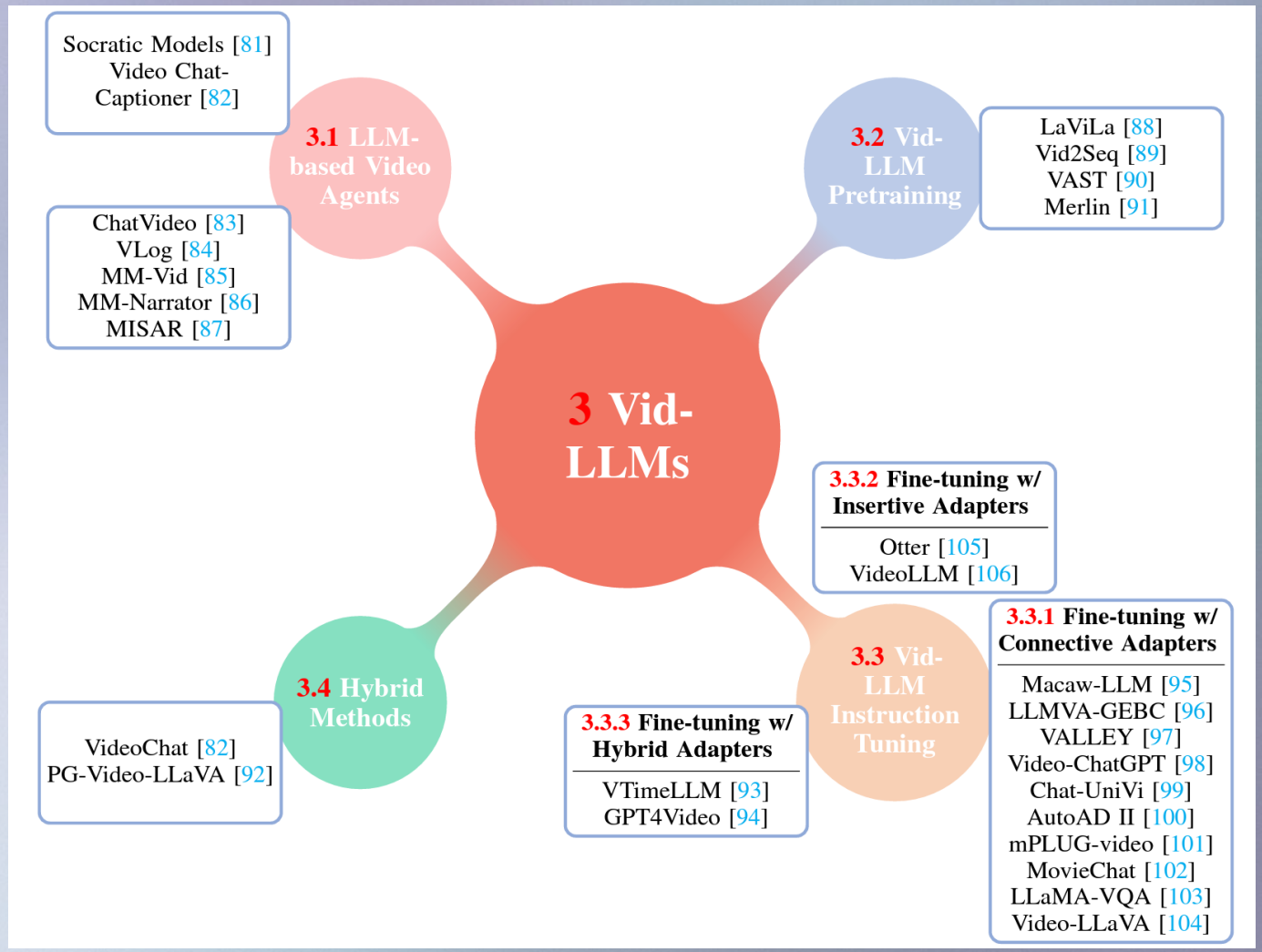


Video-LLMs Evolution

AIJ Artificial Intelligence Journey

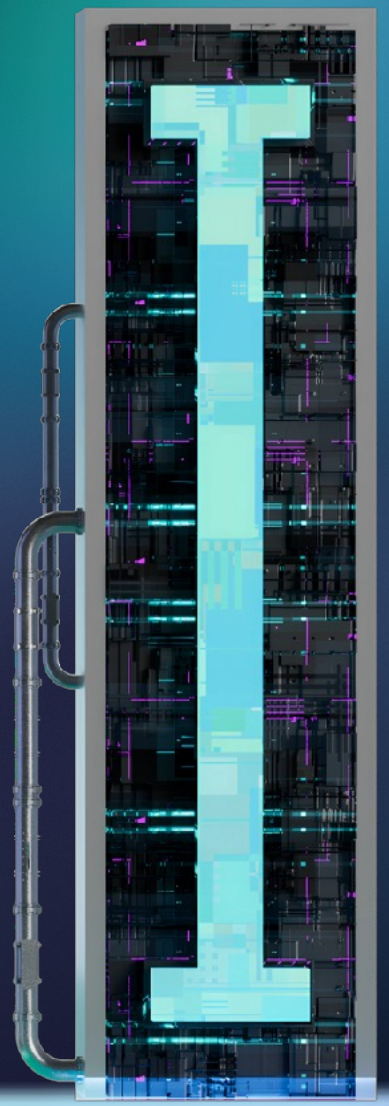


Video-LLMs Taxonomy



https://arxiv.org/abs/2312.17432v2

Emotional Fusion Brain Challenge 4.0



Fusion Brain Challenge Evolution

AIJ Artificial Intelligence Journey

	Fusion Brain Challenge 1.0 AI Journey 2021	Fusion Brain Challenge 2.0 AI Journey 2022	Fusion Brain Challenge 3.0 AI Journey 2023	Emotional Fusion Brain Challenge 4.0 AI Journey 2024
Modalities	2	2 Images Text	3 Images Text & Audio	3+ Video (Images) Text & Audio
Open Tasks	4	6	Multiple	3
Hidden Tasks	-	6		-
Features		All tasks on natural language Has hidden tasks	All tasks on natural language Not limited number of tasks Has hidden metric	All tasks on natural language Social interactions

Task

To develop a universal multimodal model to handle **three input modalities: video, audio and text**.

Contest focuses on videos containing recordings of **social interactions and human emotions**.

Task types

Video Question Answering

This task requires answering a question with a multiple-choice answer based on the video content.

Audio-Video Question Answering

This task requires answering a question based on the video and audio (speech/human sounds/nature) content.

Video Captioning

This task requires a detailed description of the video, including human emotions and interactions.

Metrics

Two metrics: **classification (Accuracy) for QA tasks and generative (METEOR) for Captioning task**.

- **Accuracy metric** (answer accuracy rate) - to evaluate the quality of answers to multiple-choice questions. It is based on an internal assessment of the model's confidence in each of the answers.

- **METEOR metric** - for evaluating the model response generation.
  It is based on the analysis of n-grams and focused on the use of a statistical and accurate evaluation of the source text.

Those metrics will be integrated into final one – **Intergal metric**.

$$I = w_M \times \sum_{j}^{J_0} METEOR_j + w_A \times \sum_{i}^{J_m} Accuracy_i$$

where J_0 and J_m - the number of questions of each type,
w_M and w_A are the weights of the METEOR and Accuracy metrics, respectively.

Input Data Format

Each data object is represented as a dictionary with the following fields defined:

- `task_id` - (type: int) represents a unique identifier for the question;
- `task_type` - (type: str) represents the type of task, which can be one of the following options: captioning / qa
- `question` - (type: str) contains the question related to the video;
- `video` - (type: str) describes the path to the video file;
- `audio` - (type: str) describes the path to the audio file, not filled in all tasks;
- `choices` - (type: list) this field contains a list of answer options for the given question.

  Each answer option:
  `{'choice_id': 0, 'choice': 'value'},`
  where `'choice_id'` is the sequential index of the answer option from range [0, 4],
  and `'choice'` is the actual value of the answer option.

Videos for which predictions need to be made are placed in the `` `video` `` folder in **.mp4** format.
Audio recordings are placed in the `` `audio` `` folder in **.mp3** format.

Input Data Format

AIJ Artificial Intelligence Journey

Audio-Video QA

```
{  'task_id': 1,
   'task_type': 'qa',
   'question': 'Who is dancing in the end of the video?',
   'video': 'path_to_video.mp4',
   'audio': 'path_to_audio.mp3',
'choices': [
      {'choice_id': 0, 'choice': 'Woman in red'},
      {'choice_id': 1, 'choice': 'Woman in blue'},
      {'choice_id': 2, 'choice': 'Man in green'},
      {'choice_id': 3, 'choice': 'Man in black'},
      {'choice_id': 4, 'choice': 'Nobody'},
   ]}
```

Video Captioning

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

Submission file structure

Videos for which predictions need to be made are placed in the `video` folder in **.mp4** format.
Audio recordings are placed in the `audio` folder in **.mp3** format.

```
├── run.py
├── video/
├── audio/
├── dataset_file.json
├── README.md
└── your-code-folder
    ├── evaluate.py
    ├── mm_utils.py
    ├── constants.py
    ├── …
```

Output Data Format

The output files should be saved to `output_path_from_job` directory (automatically created in the root of the solution).

- `./output_path_from_job/for_generative_metric.json` file for **Video Captioning** task ;
- `./output_path_from_job/for_classification_metric.json` for **Question Answering**.
  The answer values are the index of the correct answer from the provided answer choices, **starting with 0 and ending with index 4;**

Video Captioning

```
{
  0: "Video shows ...",
  3: "Here we see ..."
}
```

Video(-Audio) QA

```
{
  1: 1,
  2: 4,
  4: 0
}
```

Submission Technical Details

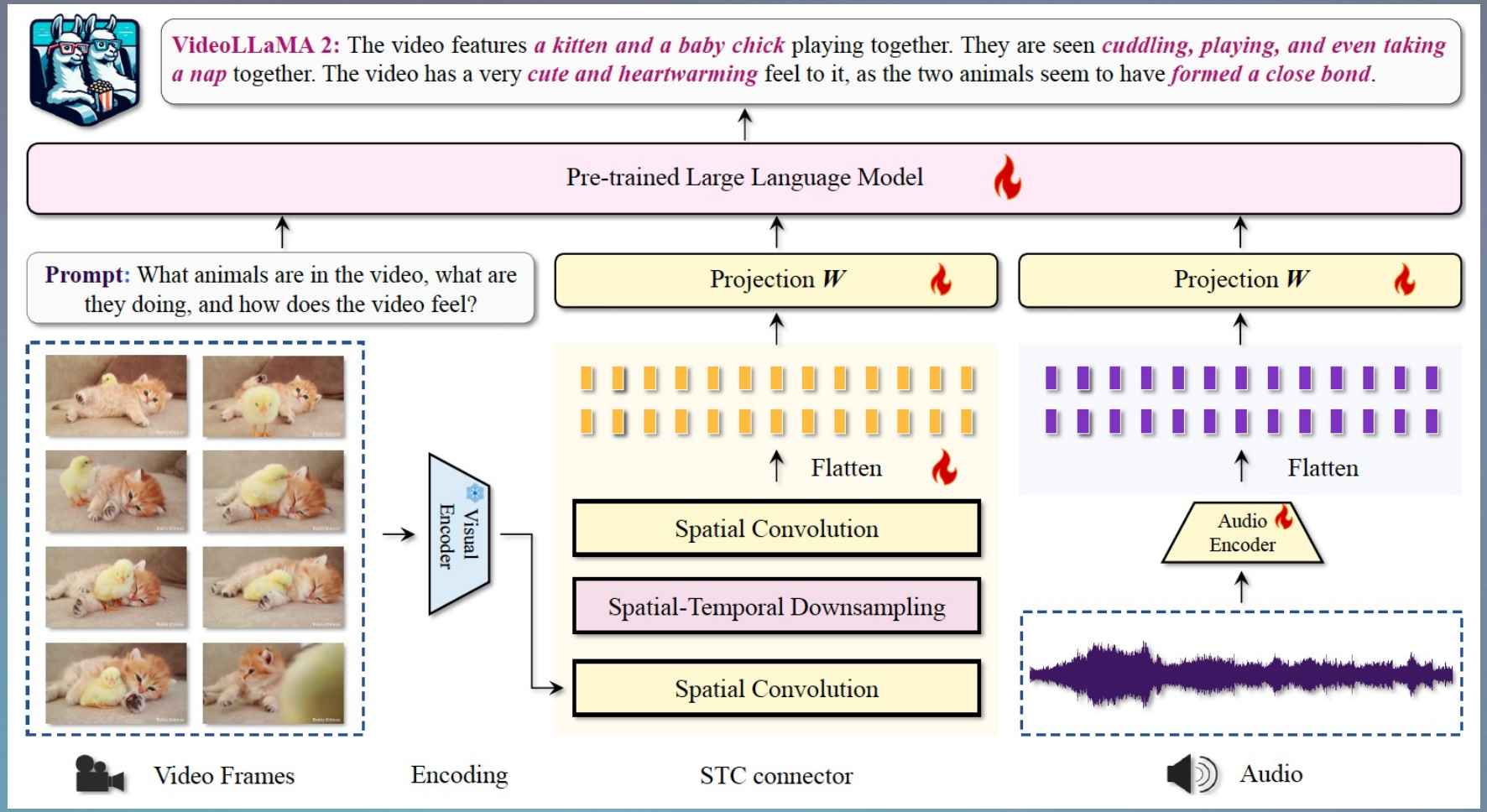
- Default image cannot be changed.
- Dockerfile and requirements for it are on Data tab of Official Contest page.
- Additional requirements can be installed through addition of libraries source code to submission.
- Additional model weights (if needed) should be included in user submission.
- The overall **user submission size should be not more then 5 Gb** (due to technical limits).
- The **disk limit is 10 Gb** (including 2.5 Gb for test dataset).
- In Docker image we stored a list of popular VidLLMs parts. The complete list of them and paths can be also found on Data tab of Official Contest page (PATHS.md).
- Resources: 243 Gb RAM, 16 CPU-cores, 1 GPU Tesla A100 (80 Гб). Max. evaluation time: 3,5 hours

Model Options

AIJ Artificial Intelligence Journey

Model name	Path in image	Link to HF
DAMO-NLP-SG/VideoLLaMA2-7B	/app/DAMO-NLP-SG/VideoLLaMA2-7B	link
imagebind_huge.pth	/app/imagebind_huge.pth	link
InternViT-300M-448px	/app/InternViT-300M-448px	link
internlm2-chat-1_8b	/app/internlm2-chat-1_8b	link
LanguageBind_Audio_FT	/app/LanguageBind_Audio_FT	link
llava-onevision-qwen2-0_5b-sis	/app/llava-onevision-qwen2-0_5b-sis	link
Qwen2-0.5B-Instruct	/app/Qwen2-0.5B-Instruct	link
llava-onevision-qwen2-7b-sis	/app/llava-onevision-qwen2-7b-sis	link
clip-vit-large-patch14-336	/app/clip-vit-large-patch14-336	link

Baseline Solution

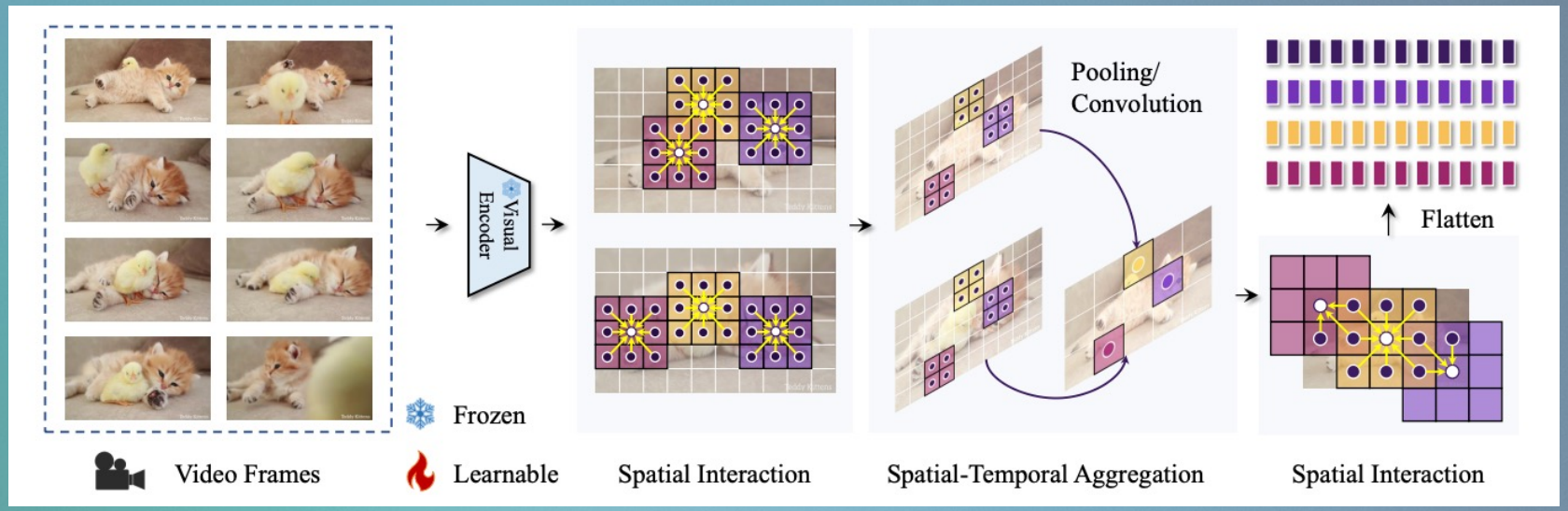


VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs

Baseline Solution

Spatio-temporal video encoding. STC Connector

- 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



Additional nominations

AIJ Artificial Intelligence Journey

The Role Game*

- to apply the developed multimodal model to **the video recordings of the role game**
- the model will have to answer questions about:
  - the course of the game,
  - determine the participants' roles
  - the likelihood of plausibility of the points voiced and demonstrated by the players.
- the Accuracy classification metric

The Fastest Solution*

- the models will be evaluated based on the lowest-time inference metric T

$$T = \frac{1}{N}\sum_{i=1}^{N}\tau_i$$

where τ_i is the inference time for task I, sec.
N is the number of tasks.

*for the Top-10 Participants or Teams of Participants indicated on the private leaderboard