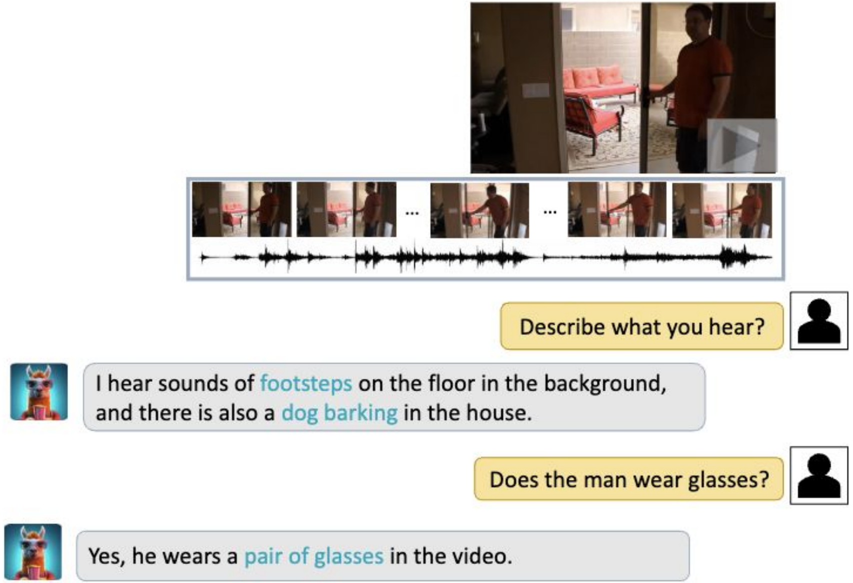
# VideoLlama 2

Seun Eisape
Vision & Language Seminar

# VideoLlama 2



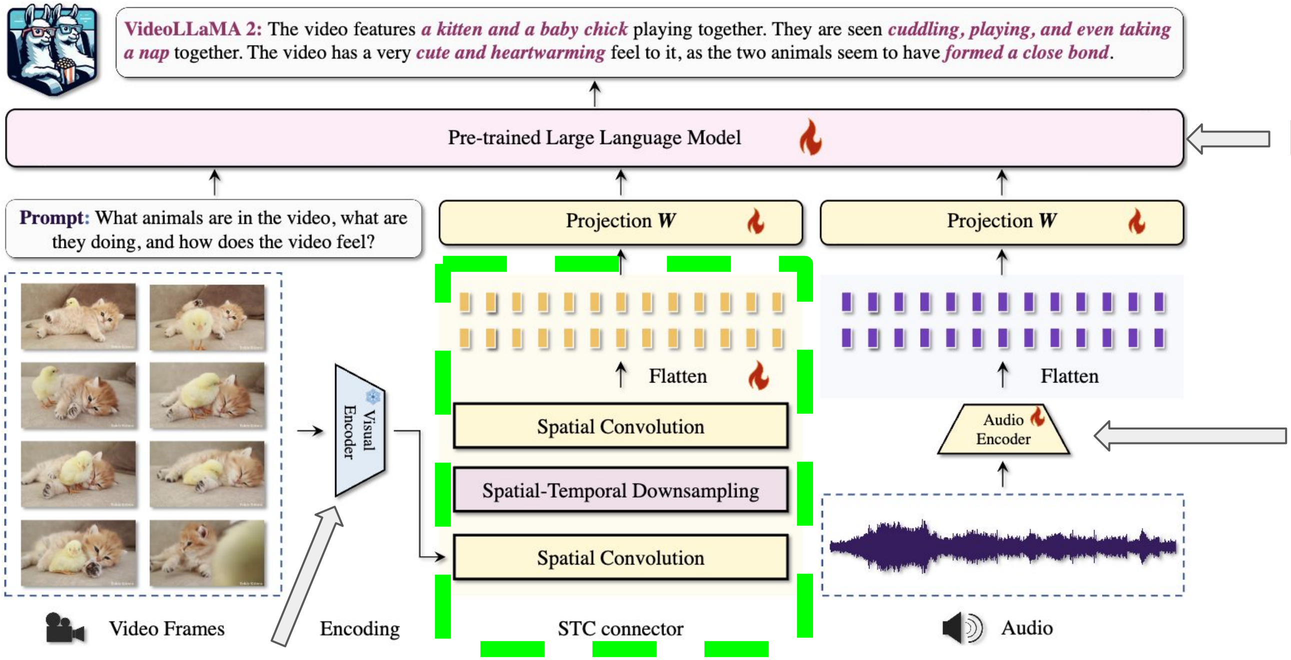
Describe what you hear?

I hear sounds of footsteps on the floor in the background, and there is also a dog barking in the house.

Does the man wear glasses?

Yes, he wears a pair of glasses in the video.

# Architecture



**VideoLLaMA 2:** The video features *a kitten and a baby chick* playing together. They are seen *cuddling, playing, and even taking a nap* together. The video has a very *cute and heartwarming* feel to it, as the two animals seem to have *formed a close bond*.

Pre-trained Large Language Model

**Prompt:** What animals are in the video, what are they doing, and how does the video feel?

Projection ***W***

Projection ***W***

Flatten

Flatten

Spatial Convolution

Spatial-Temporal Downsampling

Spatial Convolution

Visual Encoder

Audio Encoder

Video Frames

Encoding

STC connector

Audio

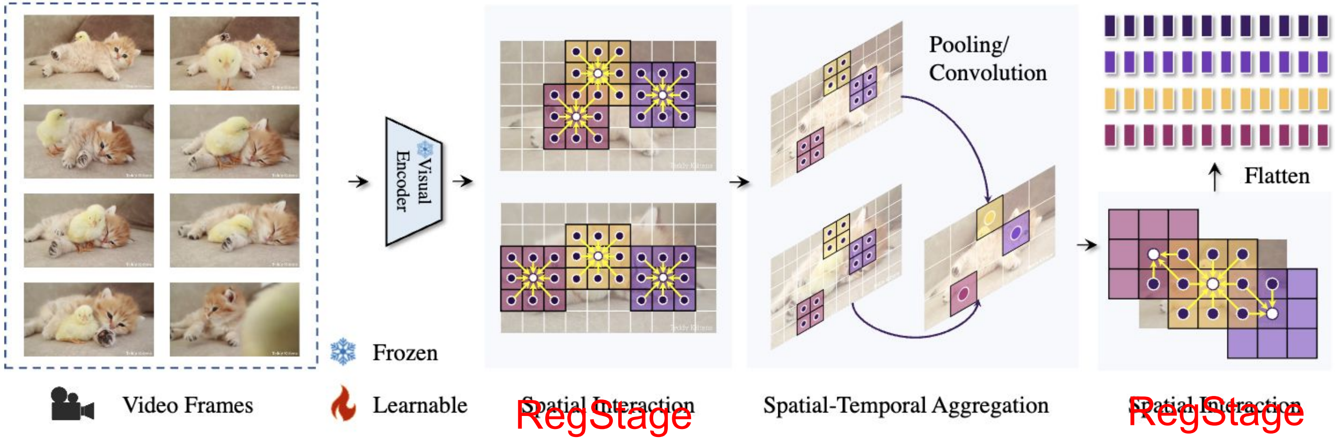
Mistral/Mixtral

BEATS

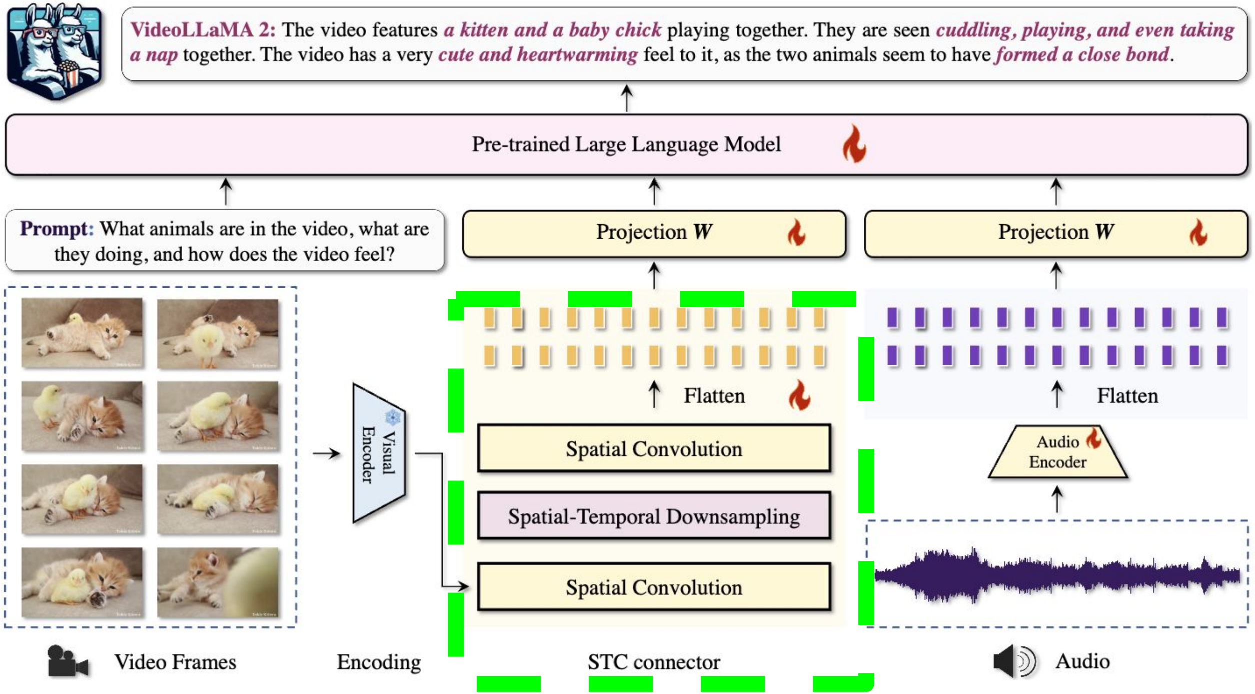
CLIP ViT-L-14

# Spatial-Temporal Convolution Connector

- Convolutions keep some notion of space and time within and across frames
  - They also reduce number of tokens needed across multiple frames
- RegStage “complements the information loss caused by the spatial-temporal downsampling”



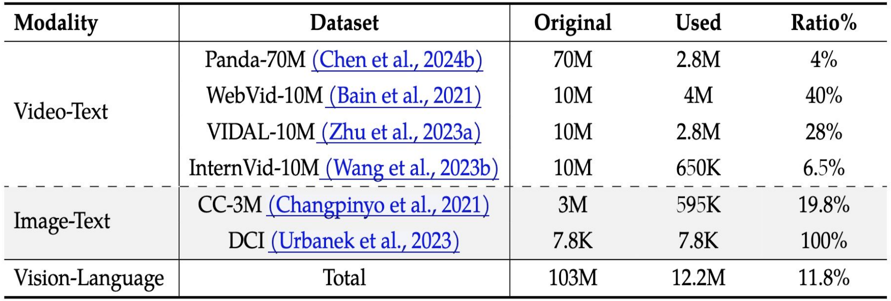
# Vision-Language “Branch”


VideoLLaMA 2: The video features a kitten and a baby chick playing together. They are seen cuddling, playing, and even taking a nap together. The video has a very cute and heartwarming feel to it, as the two animals seem to have formed a close bond.
Pre-trained Large Language Model
Prompt: What animals are in the video, what are they doing, and how does the video feel?
Projection W
Projection W
Flatten
Flatten
Spatial Convolution
Spatial-Temporal Downsampling
Spatial Convolution
Visual Encoder
Audio Encoder
Video Frames
Encoding
STC connector
Audio

# Vision-Language “Branch” Training

## Pretraining

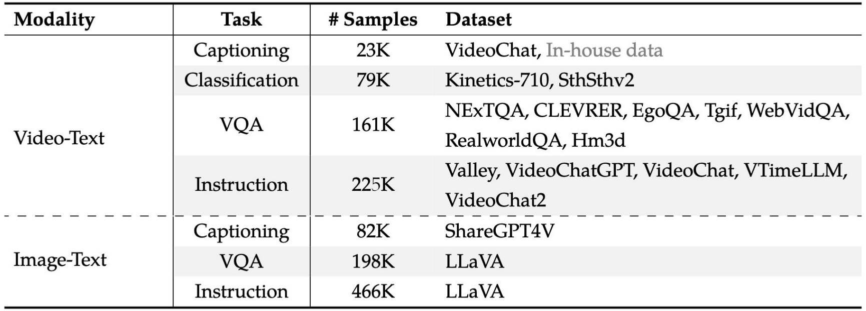
- Minimize Cross-Entropy Loss of Text Tokens



| Modality | Dataset | Original | Used | Ratio% |
|---|---|---|---|---|
| Video-Text | Panda-70M (Chen et al., 2024b) | 70M | 2.8M | 4% |
| | WebVid-10M (Bain et al., 2021) | 10M | 4M | 40% |
| | VIDAL-10M (Zhu et al., 2023a) | 10M | 2.8M | 28% |
| | InternVid-10M (Wang et al., 2023b) | 10M | 650K | 6.5% |
| Image-Text | CC-3M (Changpinyo et al., 2021) | 3M | 595K | 19.8% |
| | DCI (Urbanek et al., 2023) | 7.8K | 7.8K | 100% |
| Vision-Language | Total | 103M | 12.2M | 11.8% |

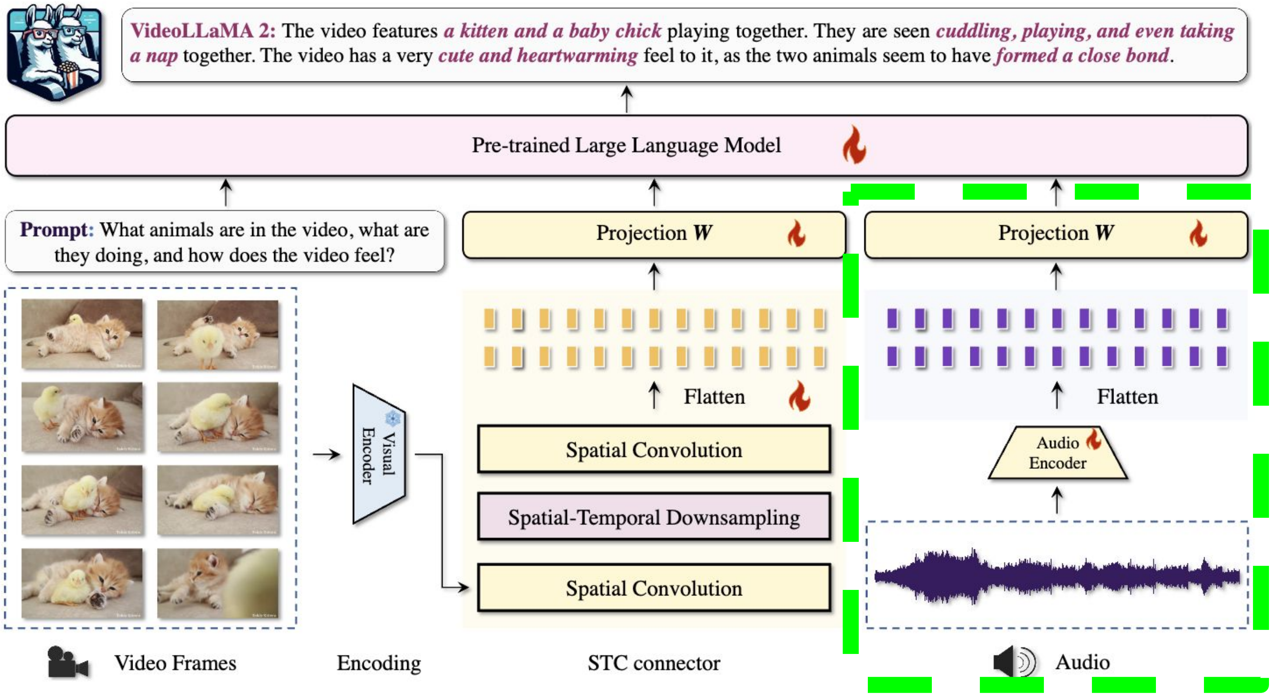
## Multitask finetuning

- Video & Image Captioning
- Video & Image Classification
- Video & Image QA



| Modality | Task | # Samples | Dataset |
|---|---|---|---|
| Video-Text | Captioning | 23K | VideoChat, In-house data |
| | Classification | 79K | Kinetics-710, SthSthv2 |
| | VQA | 161K | NExTQA, CLEVRER, EgoQA, Tgif, WebVidQA, RealworldQA, Hm3d |
| | Instruction | 225K | Valley, VideoChatGPT, VideoChat, VTimeLLM, VideoChat2 |
| Image-Text | Captioning | 82K | ShareGPT4V |
| | VQA | 198K | LLaVA |
| | Instruction | 466K | LLaVA |

# Audio-Language “Branch”


VideoLLaMA 2: The video features a kitten and a baby chick playing together. They are seen cuddling, playing, and even taking a nap together. The video has a very cute and heartwarming feel to it, as the two animals seem to have formed a close bond.
Pre-trained Large Language Model
Prompt: What animals are in the video, what are they doing, and how does the video feel?
Projection W
Projection W
Flatten
Flatten
Spatial Convolution
Spatial-Temporal Downsampling
Spatial Convolution
Visual Encoder
Audio Encoder
Video Frames
Encoding
STC connector
Audio

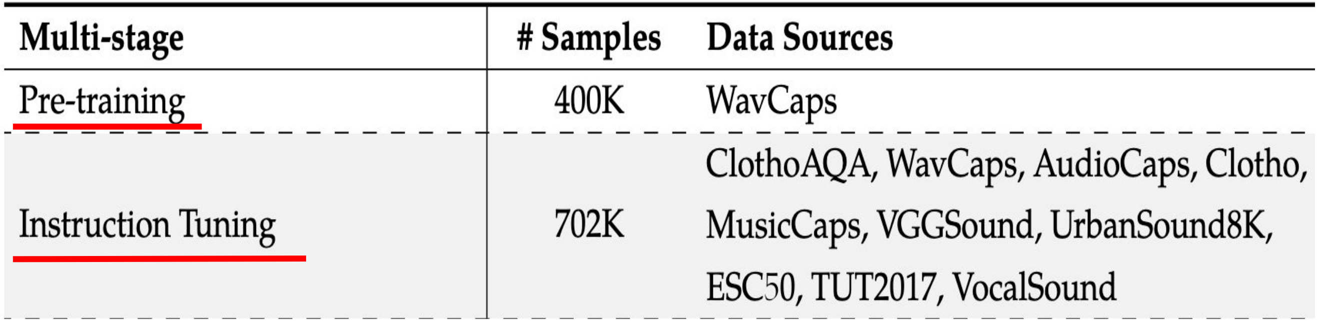
# Audio-Language “Branch”

Pretraining

- Minimize next token (text) prediction loss

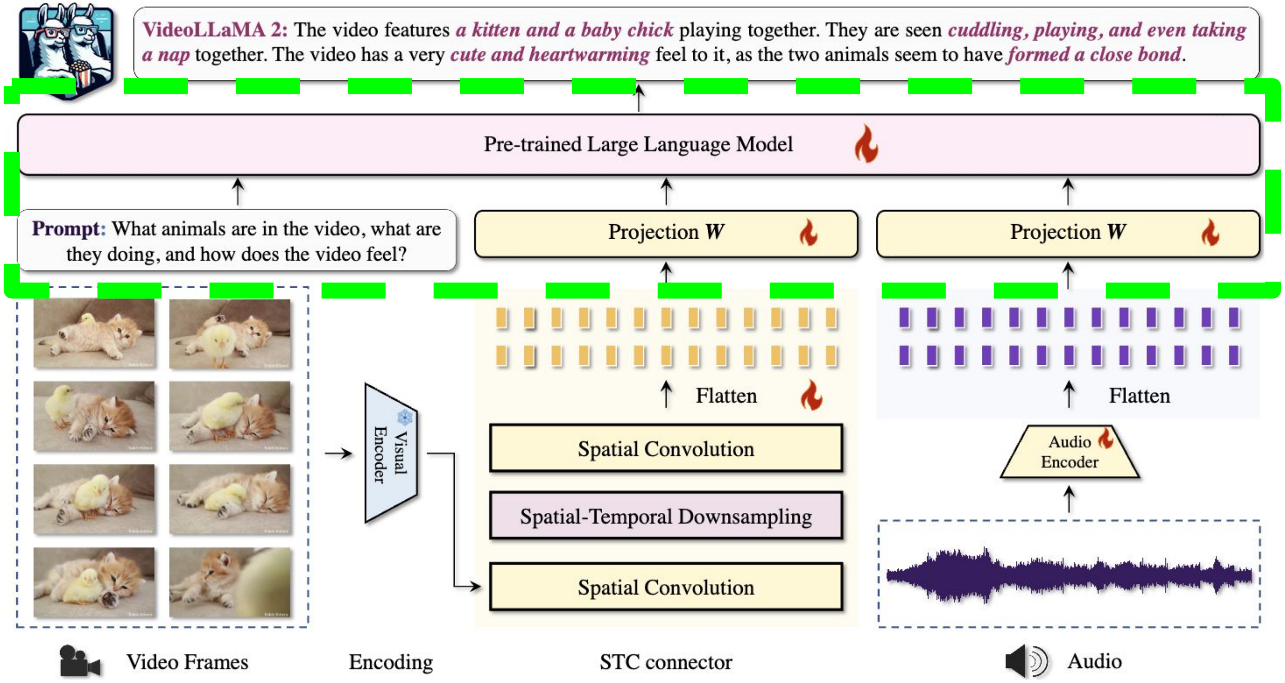
Multitask finetuning

- QA
- Captioning
- Sound Event Classification



| Multi-stage | # Samples | Data Sources |
|---|---|---|
| Pre-training | 400K | WavCaps |
| Instruction Tuning | 702K | ClothoAQA, WavCaps, AudioCaps, Clotho, MusicCaps, VGGSound, UrbanSound8K, ESC50, TUT2017, VocalSound |

# Audio-Video Joint Training

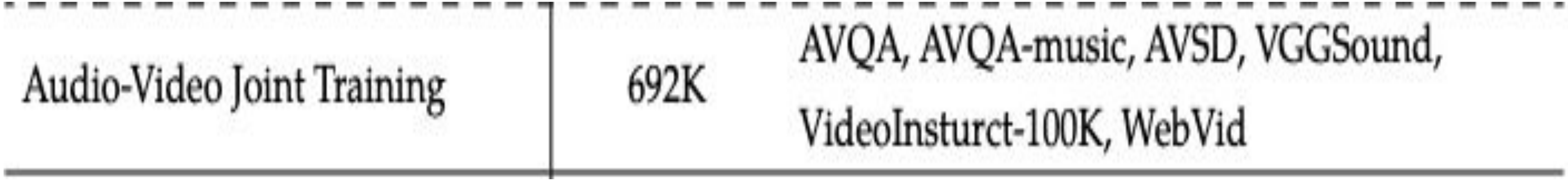

VideoLLaMA 2: The video features a kitten and a baby chick playing together. They are seen cuddling, playing, and even taking a nap together. The video has a very cute and heartwarming feel to it, as the two animals seem to have formed a close bond.
Pre-trained Large Language Model
Prompt: What animals are in the video, what are they doing, and how does the video feel?
Projection W
Projection W
Flatten
Flatten
Visual Encoder
Spatial Convolution
Spatial-Temporal Downsampling
Spatial Convolution
Audio Encoder
Video Frames
Encoding
STC connector
Audio

# Audio-Video Joint Training

Finetune on Aligned Audio & Video

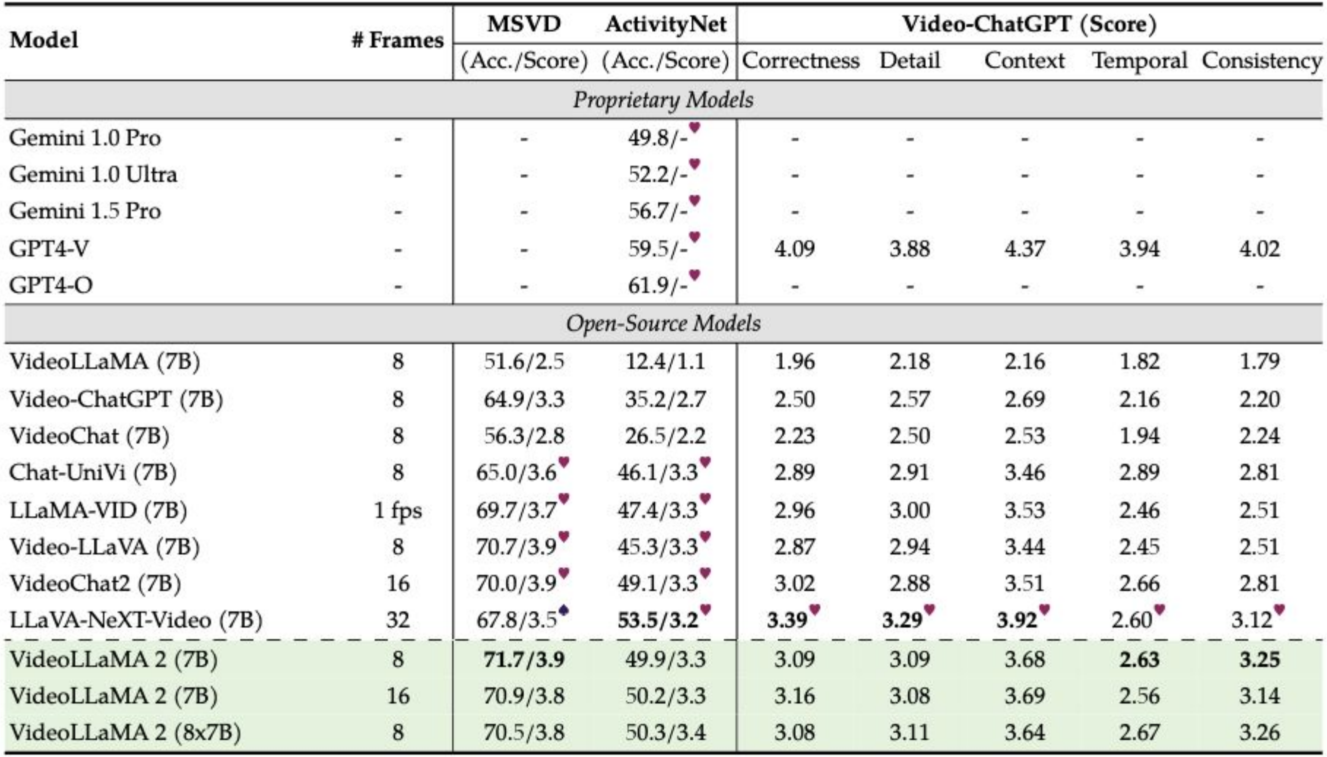
Tasks:

- Audio Visual QA
- Audio Visual Classification



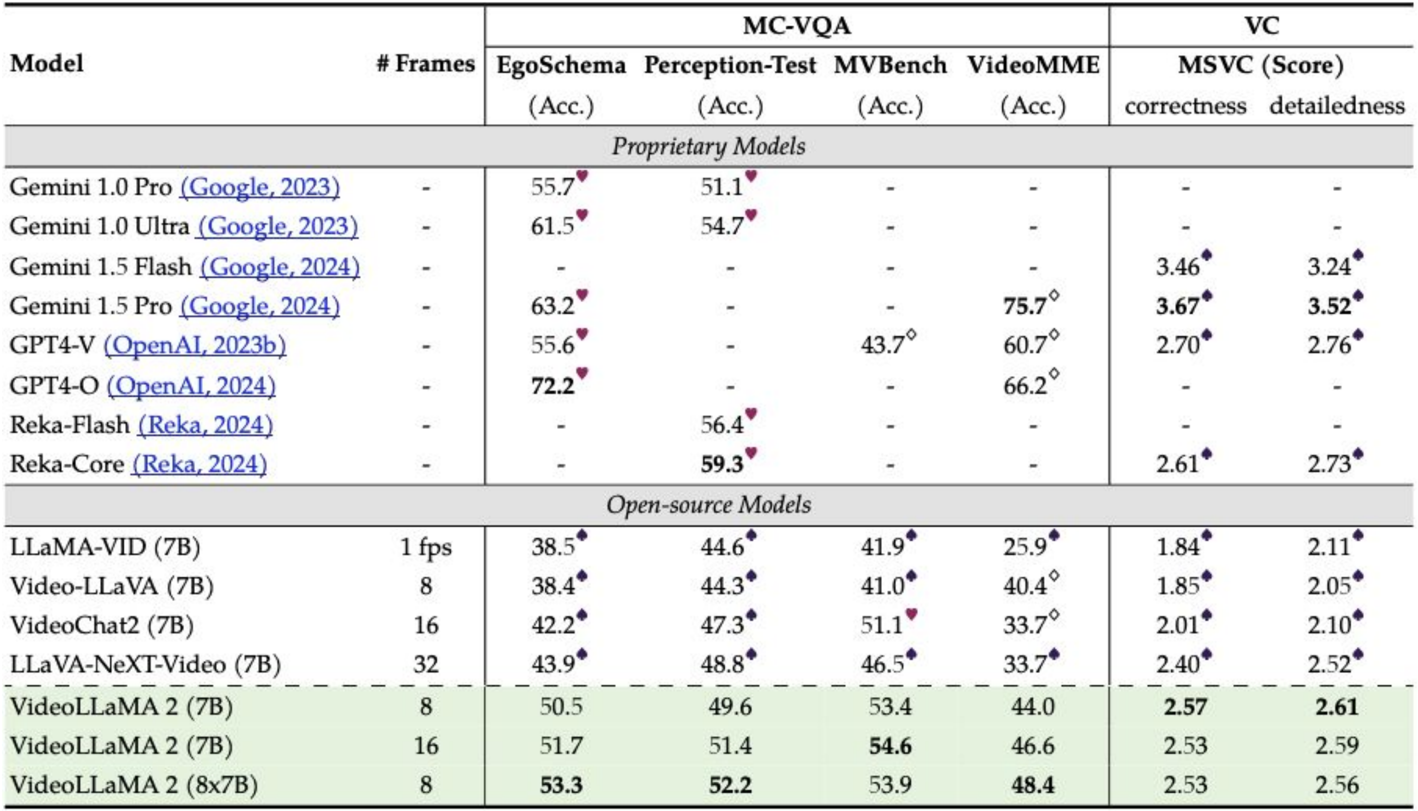
| | | |
|---|---|---|
| Audio-Video Joint Training | 692K | AVQA, AVQA-music, AVSD, VGGSound, VideoInsturct-100K, WebVid |

# Opened Ended Video QA



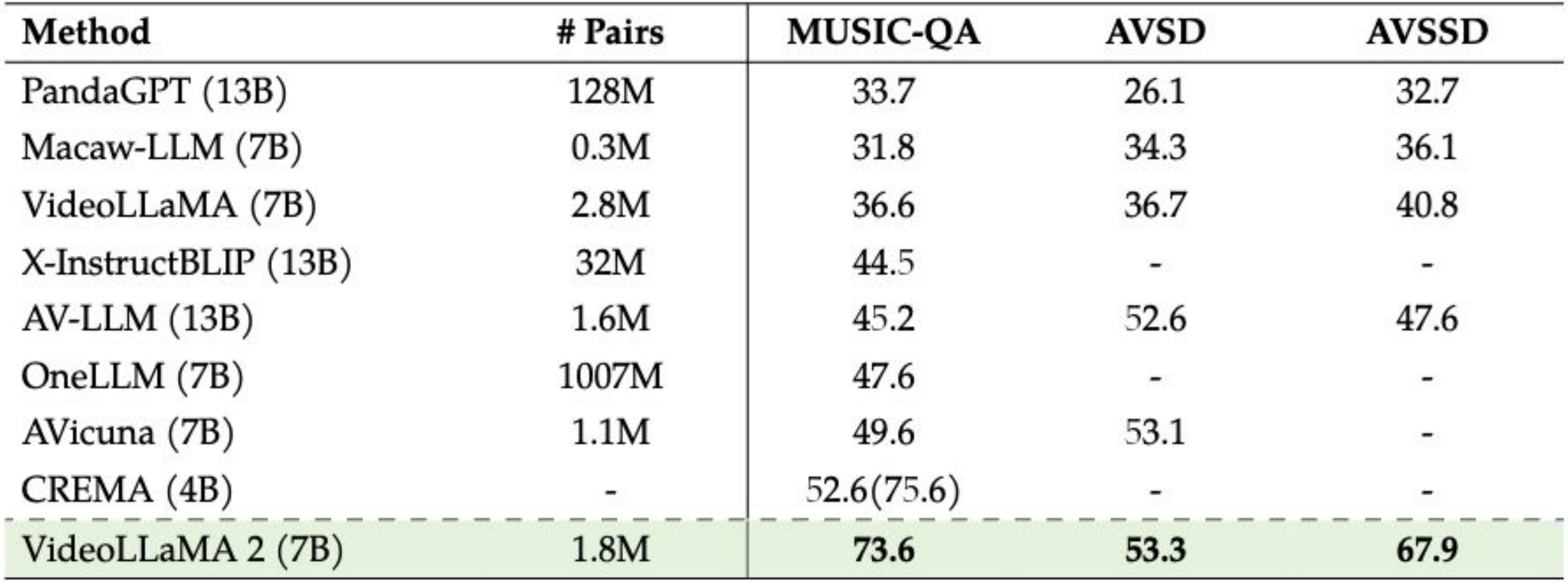
| Model | # Frames | MSVD | ActivityNet | Video-ChatGPT (Score) | | | | |
|---|---|---|---|---|---|---|---|---|
| | | (Acc./Score) | (Acc./Score) | Correctness | Detail | Context | Temporal | Consistency |
| *Proprietary Models* | | | | | | | | |
| Gemini 1.0 Pro | - | - | 49.8/-♥ | - | - | - | - | - |
| Gemini 1.0 Ultra | - | - | 52.2/-♥ | - | - | - | - | - |
| Gemini 1.5 Pro | - | - | 56.7/-♥ | - | - | - | - | - |
| GPT4-V | - | - | 59.5/-♥ | 4.09 | 3.88 | 4.37 | 3.94 | 4.02 |
| GPT4-O | - | - | 61.9/-♥ | - | - | - | - | - |
| *Open-Source Models* | | | | | | | | |
| VideoLLaMA (7B) | 8 | 51.6/2.5 | 12.4/1.1 | 1.96 | 2.18 | 2.16 | 1.82 | 1.79 |
| Video-ChatGPT (7B) | 8 | 64.9/3.3 | 35.2/2.7 | 2.50 | 2.57 | 2.69 | 2.16 | 2.20 |
| VideoChat (7B) | 8 | 56.3/2.8 | 26.5/2.2 | 2.23 | 2.50 | 2.53 | 1.94 | 2.24 |
| Chat-UniVi (7B) | 8 | 65.0/3.6♥ | 46.1/3.3♥ | 2.89 | 2.91 | 3.46 | 2.89 | 2.81 |
| LLaMA-VID (7B) | 1 fps | 69.7/3.7♥ | 47.4/3.3♥ | 2.96 | 3.00 | 3.53 | 2.46 | 2.51 |
| Video-LLaVA (7B) | 8 | 70.7/3.9♥ | 45.3/3.3♥ | 2.87 | 2.94 | 3.44 | 2.45 | 2.51 |
| VideoChat2 (7B) | 16 | 70.0/3.9♥ | 49.1/3.3♥ | 3.02 | 2.88 | 3.51 | 2.66 | 2.81 |
| LLaVA-NeXT-Video (7B) | 32 | 67.8/3.5◆ | **53.5/3.2**♥ | **3.39**♥ | **3.29**♥ | **3.92**♥ | 2.60♥ | 3.12♥ |
| VideoLLaMA 2 (7B) | 8 | **71.7/3.9** | 49.9/3.3 | 3.09 | 3.09 | 3.68 | **2.63** | **3.25** |
| VideoLLaMA 2 (7B) | 16 | 70.9/3.8 | 50.2/3.3 | 3.16 | 3.08 | 3.69 | 2.56 | 3.14 |
| VideoLLaMA 2 (8x7B) | 8 | 70.5/3.8 | 50.3/3.4 | 3.08 | 3.11 | 3.64 | 2.67 | 3.26 |

# Multiple Choice Video QA



| Model | # Frames | MC-VQA |  |  |  | VC |  |
|---|---|---|---|---|---|---|---|
|  |  | EgoSchema | Perception-Test | MVBench | VideoMME | MSVC (Score) |  |
|  |  | (Acc.) | (Acc.) | (Acc.) | (Acc.) | correctness | detailedness |
| *Proprietary Models* |  |  |  |  |  |  |  |
| Gemini 1.0 Pro (Google, 2023) | - | 55.7♥ | 51.1♥ | - | - | - | - |
| Gemini 1.0 Ultra (Google, 2023) | - | 61.5♥ | 54.7♥ | - | - | - | - |
| Gemini 1.5 Flash (Google, 2024) | - | - | - | - | - | 3.46♠ | 3.24♠ |
| Gemini 1.5 Pro (Google, 2024) | - | 63.2♥ | - | - | **75.7**◇ | **3.67**♠ | **3.52**♠ |
| GPT4-V (OpenAI, 2023b) | - | 55.6♥ | - | 43.7◇ | 60.7◇ | 2.70♠ | 2.76♠ |
| GPT4-O (OpenAI, 2024) | - | **72.2**♥ | - | - | 66.2◇ | - | - |
| Reka-Flash (Reka, 2024) | - | - | 56.4♥ | - | - | - | - |
| Reka-Core (Reka, 2024) | - | - | **59.3**♥ | - | - | 2.61♠ | 2.73♠ |
| *Open-source Models* |  |  |  |  |  |  |  |
| LLaMA-VID (7B) | 1 fps | 38.5♠ | 44.6♠ | 41.9♠ | 25.9♠ | 1.84♠ | 2.11♠ |
| Video-LLaVA (7B) | 8 | 38.4♠ | 44.3♠ | 41.0♠ | 40.4◇ | 1.85♠ | 2.05♠ |
| VideoChat2 (7B) | 16 | 42.2♠ | 47.3♠ | 51.1♥ | 33.7◇ | 2.01♠ | 2.10♠ |
| LLaVA-NeXT-Video (7B) | 32 | 43.9♠ | 48.8♠ | 46.5♠ | 33.7♠ | 2.40♠ | 2.52♠ |
| VideoLLaMA 2 (7B) | 8 | 50.5 | 49.6 | 53.4 | 44.0 | **2.57** | **2.61** |
| VideoLLaMA 2 (7B) | 16 | 51.7 | 51.4 | **54.6** | 46.6 | 2.53 | 2.59 |
| VideoLLaMA 2 (8x7B) | 8 | **53.3** | **52.2** | 53.9 | **48.4** | 2.53 | 2.56 |

# Open Ended Audio Video QA



| **Method** | **# Pairs** | **MUSIC-QA** | **AVSD** | **AVSSD** |
|---|---|---|---|---|
| PandaGPT (13B) | 128M | 33.7 | 26.1 | 32.7 |
| Macaw-LLM (7B) | 0.3M | 31.8 | 34.3 | 36.1 |
| VideoLLaMA (7B) | 2.8M | 36.6 | 36.7 | 40.8 |
| X-InstructBLIP (13B) | 32M | 44.5 | - | - |
| AV-LLM (13B) | 1.6M | 45.2 | 52.6 | 47.6 |
| OneLLM (7B) | 1007M | 47.6 | - | - |
| AVicuna (7B) | 1.1M | 49.6 | 53.1 | - |
| CREMA (4B) | - | 52.6(75.6) | - | - |
| VideoLLaMA 2 (7B) | 1.8M | **73.6** | **53.3** | **67.9** |

# Discussion

“If you were to design a suite of analyses similar to the Idefics2 or Prismatic VLM ablations but for Video-LLMs, what design decisions would you ablate and why?”

- *Stephanie Fu*

“In the real-world, video tasks may involve much longer time scales than typically used in the benchmarks. How might the STC connector need to be adapted or extended to handle very long videos?”

- Anish Kachinthaya

# Discussion continued

“The paper appears to imply that the architectural design of the STC component played an important role in yielding the benchmark improvements presented. Is this a fair comparison? What about the role of data – if training data choices are no longer standardized, how can we reliably differentiate the impact of different architectural decisions across models?”

- *Rudy Corona*

“Maybe normalize performance on a task by number of samples seen during training”

- *Seun (comment)*

“What is RegStage?” *happens before and after convolution to ‘complement information loss’

- *Seun Eisape*

# Discussion Continued

“Given that the STC module improves temporal modelling, would it make sense to have a model that does something similar to fuse the audio and video modalities? They don’t explore speech tasks or ASR tasks which are highly dependent on the ability to model low level video and audio features in a temporal manner.”

- *Giscard Biamby*

Any questions that were not answered?