OmniLong: A Resource-Effective Context Scaling Framework for Multimodal LLM Customization

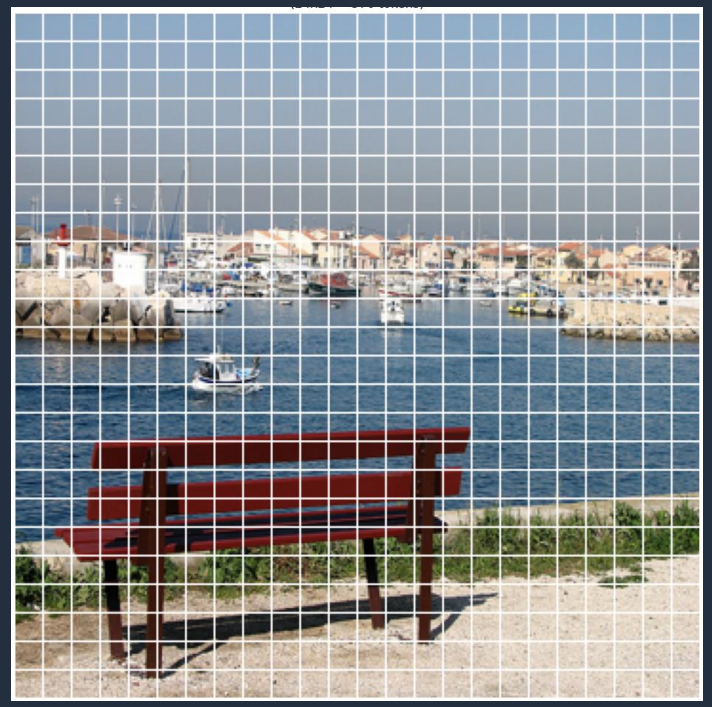
Yin Song
Senior Applied Scientist

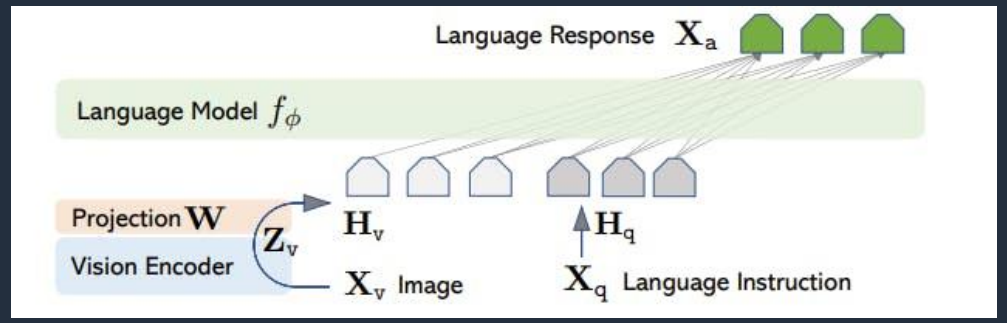
Chen Wu
Principal Applied Scientist

AWS WWSO Prototyping Team

Motivation – Long Contexts for Multi-Modal LLM

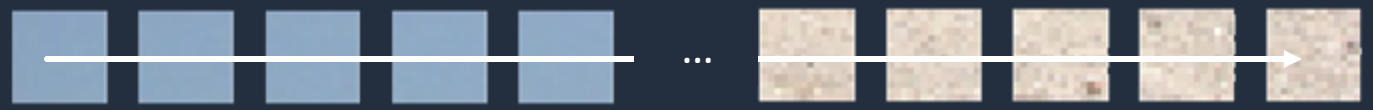
Image of 336 x 336 divided into 14 x 14 patches
=> 24 x 24 = 576 patches





LlaVA Architecture from [Liu et al. 2023]

1 frame = 576 tokens

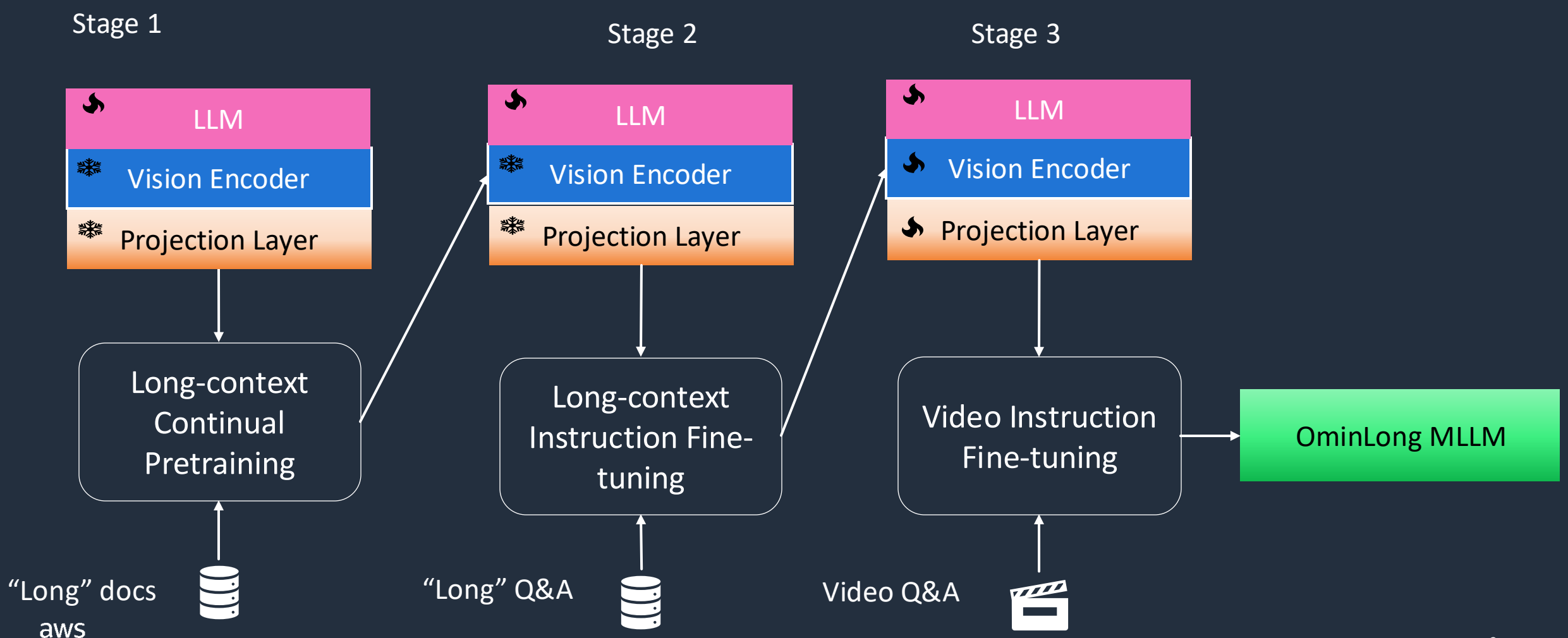


The average length of a YouTube video = 11.7 minutes

1 frame per second sampling rate =>

576 x 11.7 x 60 = 404,352 tokens

OmniLong Training Recipe

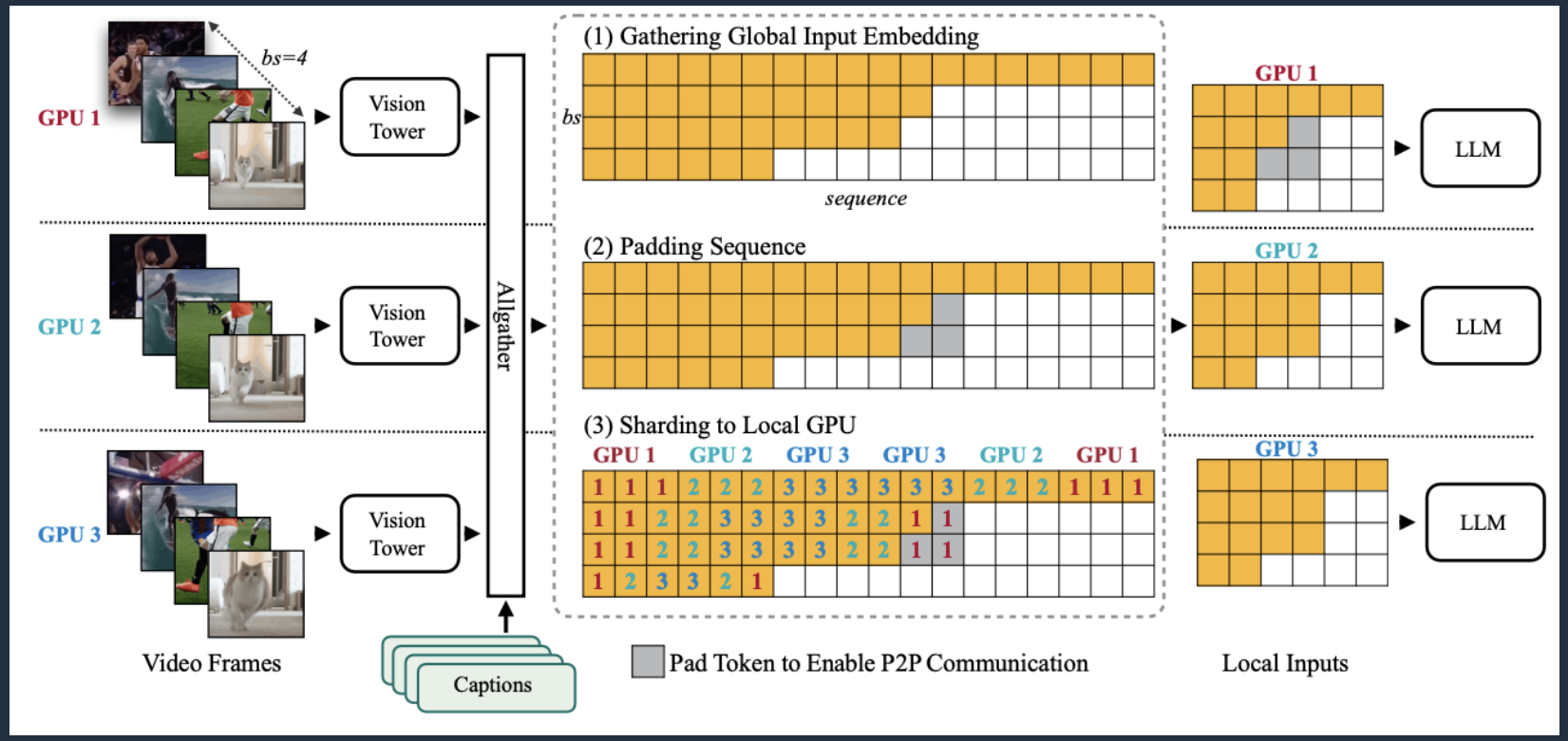


Handling Multimodal Long Context Input

Step 1: Shard Images

Step 2: Shard Tokens

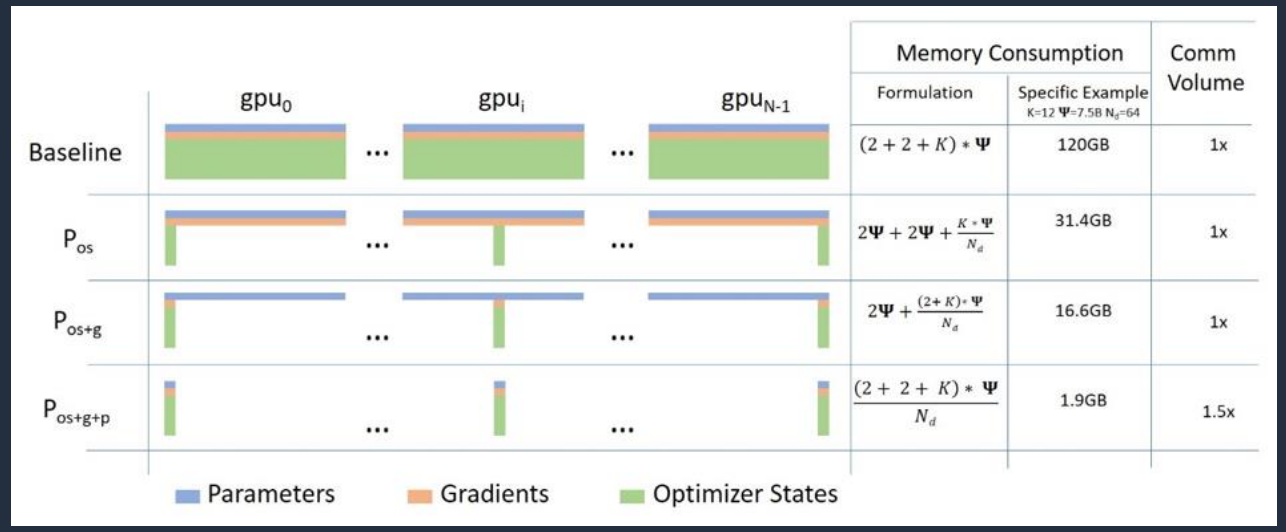
Step 3: Sequence Parallel Training



Ring Attention

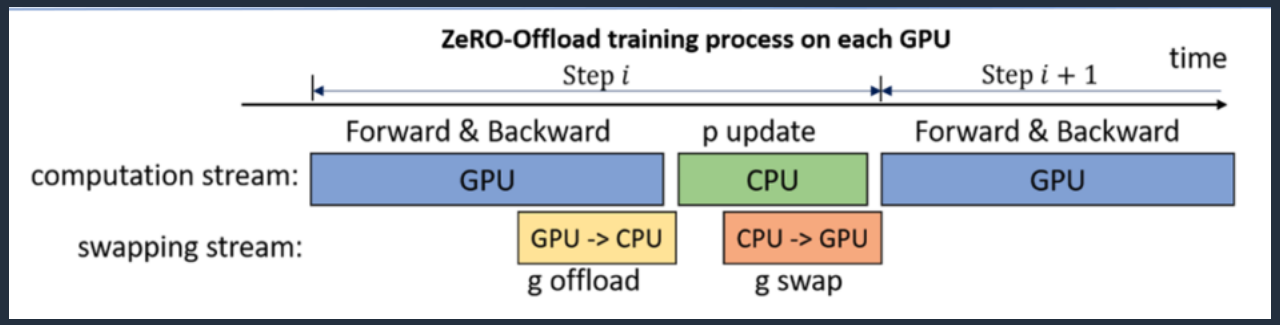
[Figure from chen et al. 2024]

DeepSpeed Zero3 + CPU Offload



	Memory Consumption		Comm Volume
	Formulation	Specific Example K=12 Ψ=7.5B N_d=64	
Baseline	$(2+2+K)*\Psi$	120GB	1x
P_{os}	$2\Psi+2\Psi+\frac{K*\Psi}{N_d}$	31.4GB	1x
P_{os+g}	$2\Psi+\frac{(2+K)*\Psi}{N_d}$	16.6GB	1x
P_{os+g+p}	$\frac{(2+2+K)*\Psi}{N_d}$	1.9GB	1.5x

Stage 1 : Shards optimizer states across data parallel workers/GPUs
Stage 2 : Shards optimizer states + gradients across data parallel workers/GPUs
Stage 3: Shards optimizer states + gradients + model parameters across data parallel workers/GPUs



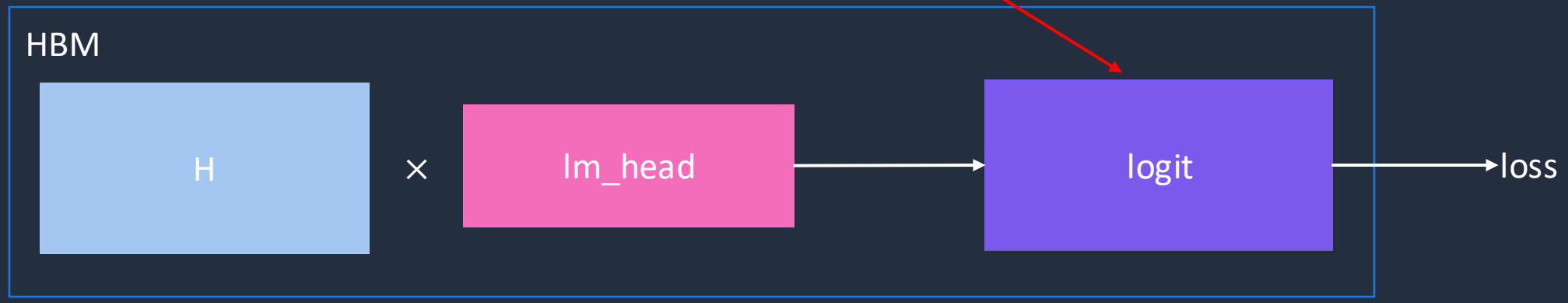
Optimizer Offload: Offloads the gradients + optimizer states to CPU/Disk building on top of ZERO Stage 2
Param Offload: Offloads the model parameters to CPU/Disk building on top of ZERO Stage 3

Figures from [DeepSpeed Team et al 2020]

Out of Memory (OOM) issue – Root Cause

torch.OutOfMemoryError: CUDA out of memory.

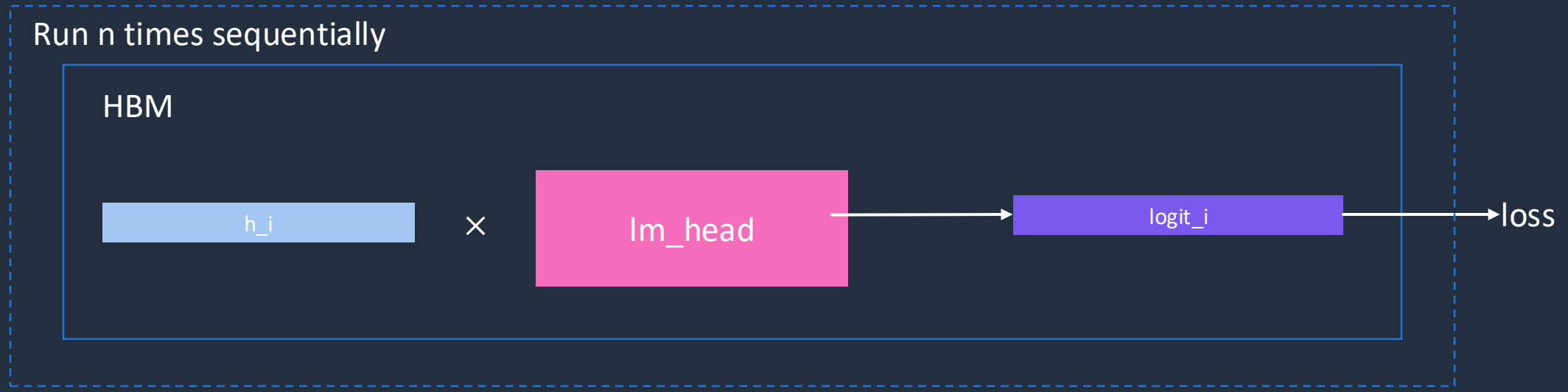
Caused by fully materializing the logits matrix and its gradients!!!



Out of Memory (OOM) issue - Solution

torch.OutOfMemoryError: CUDA out of memory.

Optimised Loss Calculation



Datasets and Base Models

Base Models
- llava-onevision-qwen2-7b-ov-hf
- Qwen2.5-VL-7B-Instruct

Fine-tuned Models
- OmniLong-LLaVA-OneVision-7B
- OmniLong-Qwen2.5-VL-7B

Stage 1

source code
research papers
open web content
public domain books

1.2B tokens

Long-context Continual Pretraining

Stage 2

64 K tokens
128 K tokens
256 K tokens
512 K tokens

22M tokens

Long-context Instruction Fine-tuning

Stage 3

256 Frames
512 Frames
1024 Frames
2048 Frames

1200 videos

Video Instruction Fine-tuning

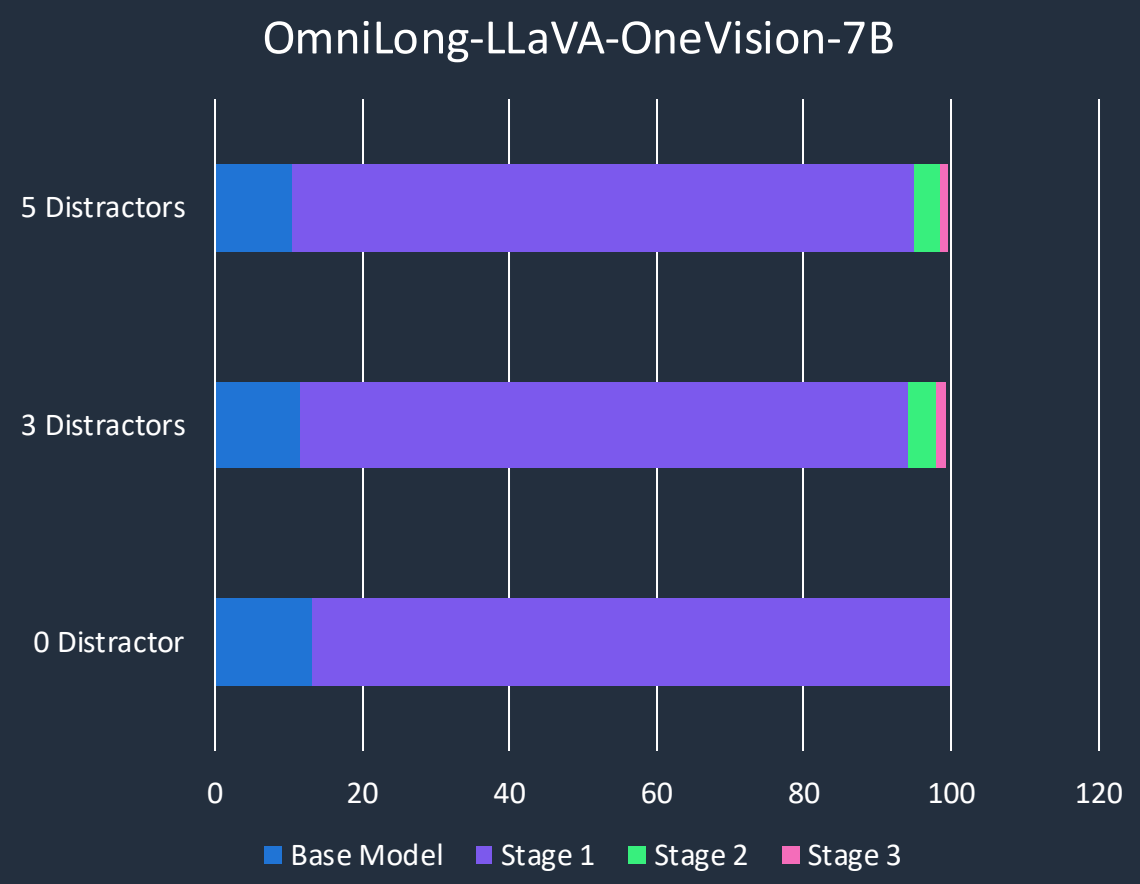
Text Needles in A Haystack (NIAH)

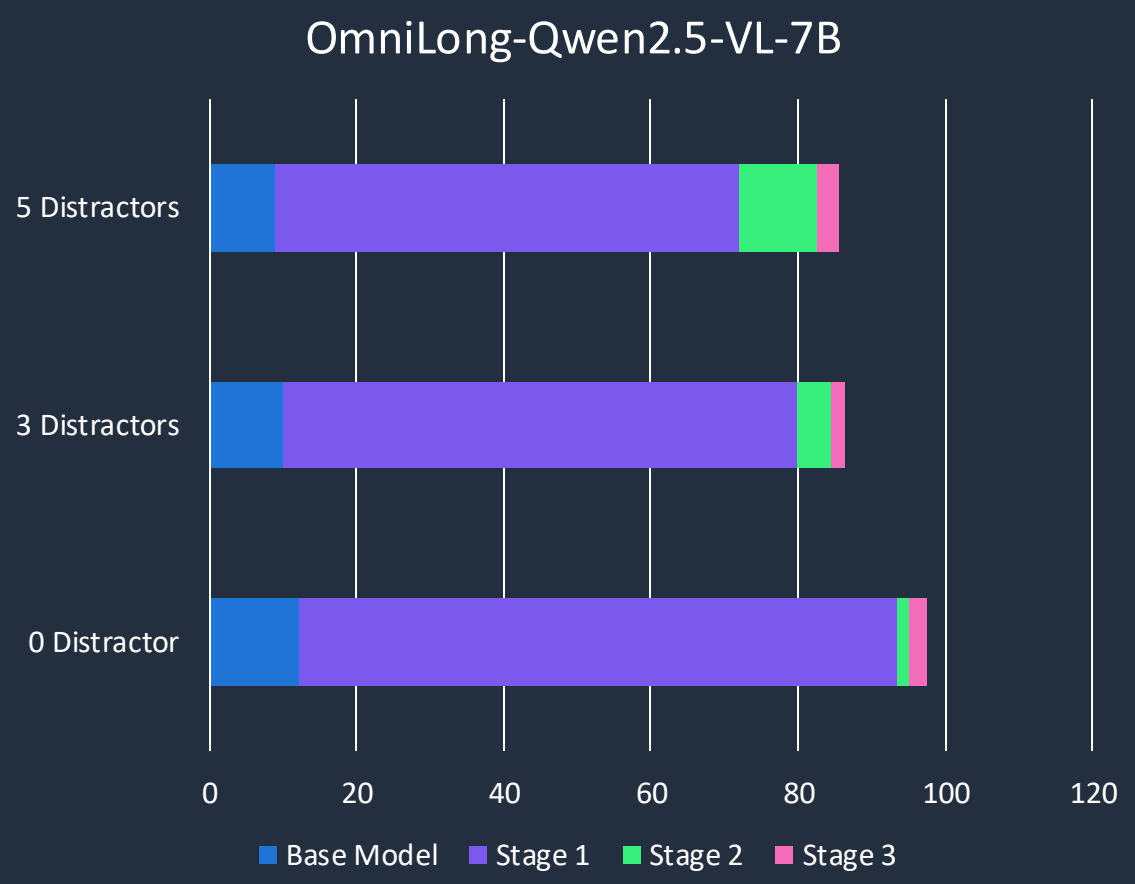
<book>
…….
…. The special magic Singapore number is: 1000 ….
…. The special magic Tokyo number is: 2333 …….
…….
</book>.
Based on the content of the book,
Question: What is the special magic Singapore number?
Answer: The special magic Singapore number is:

Needle

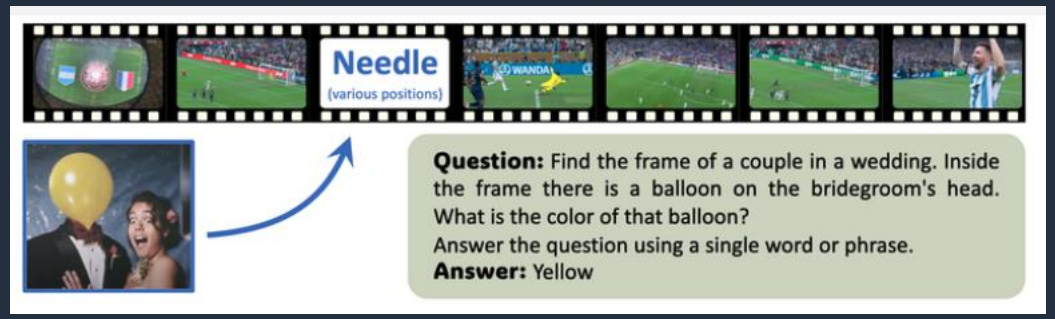
Distractor

Context Length:
32K to 512K





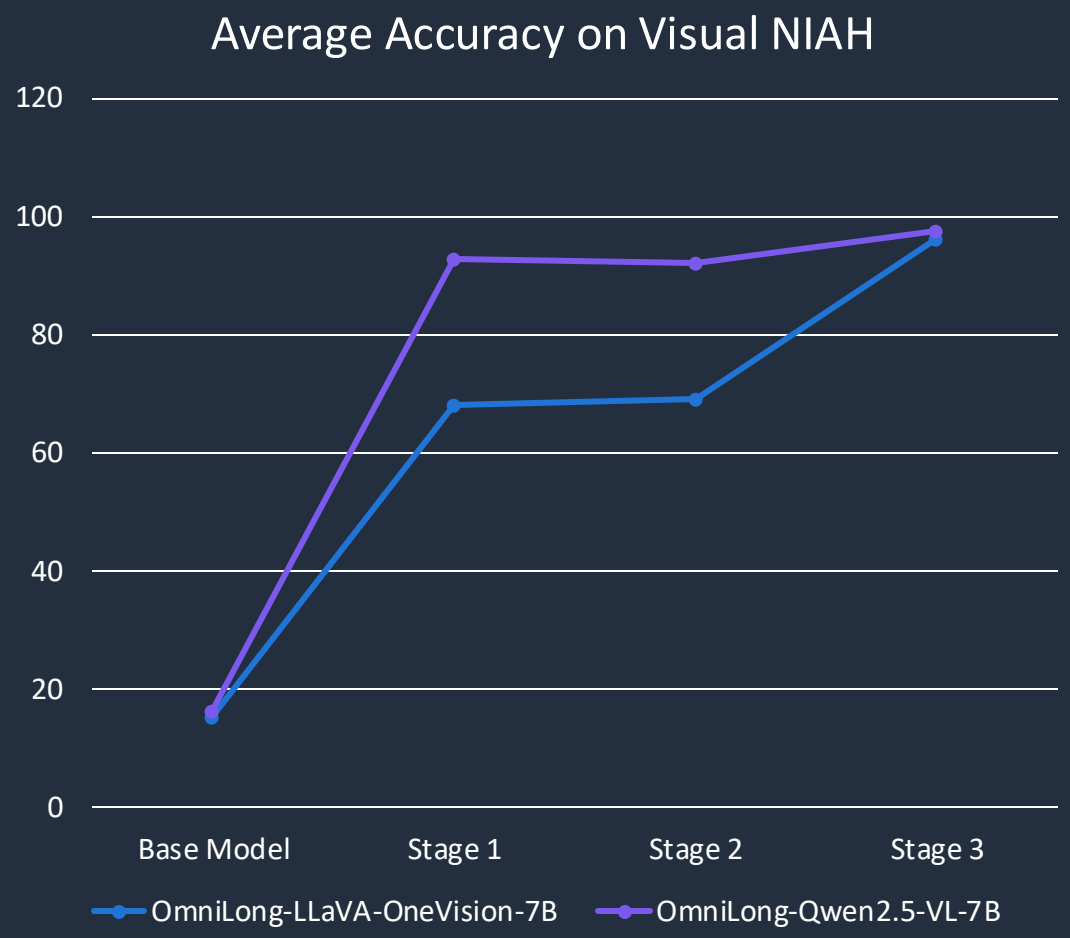
Long Context Capabilities in Video Domain



[Figure from Zhang et al. 2024]

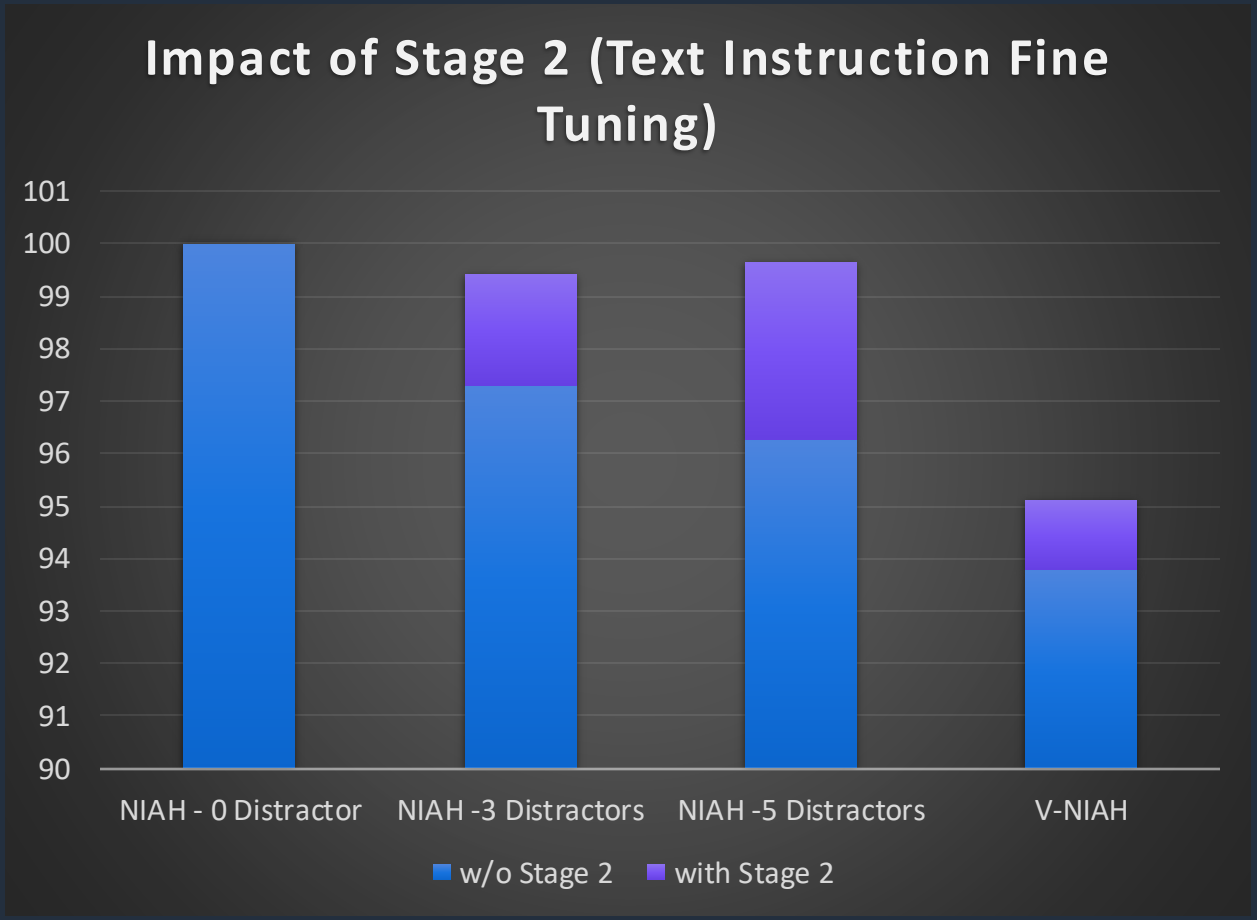
Vision Needle in the Haystack

- Purely synthetic long vision benchmark inspired by the language model's NIAH test
- 5 video question-answering problems and inserts each answer (ie. the “needle”) as a single frame into hours-long videos (ie. the “haystack”)



Ablation study

- W/O Stage 2 vs with Stage 2
- Stage2 (Text Instruction Fine-tuning) generally further enhance the long context capabilities



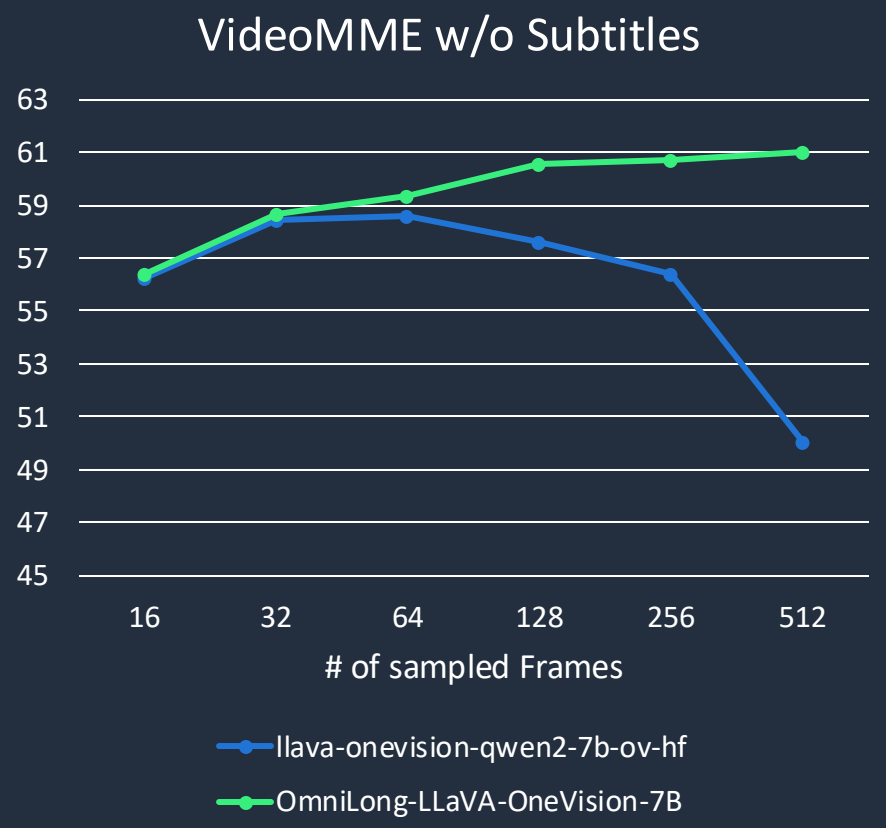
OmniLong-LLaVA-OneVision-7B

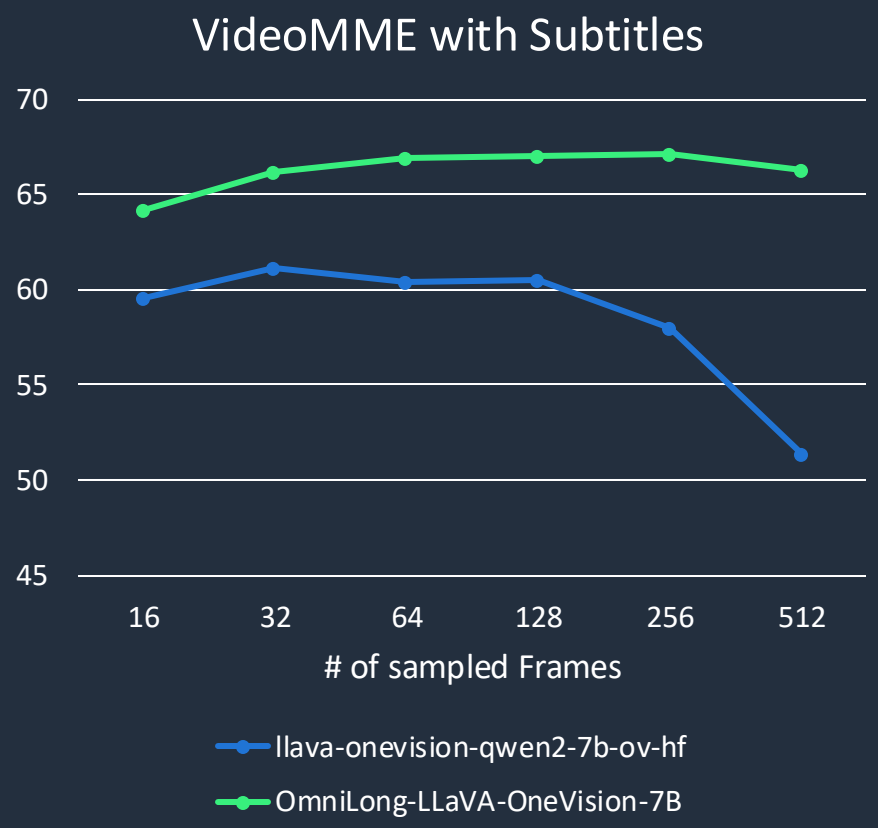
VideoMME: Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Understanding

- short (<2 min.)
- medium (4–15min.)
- long (30–60min.)
- A total of 900 videos (254 hours)
- 2,700 question-answer pairs.
- Subtitles

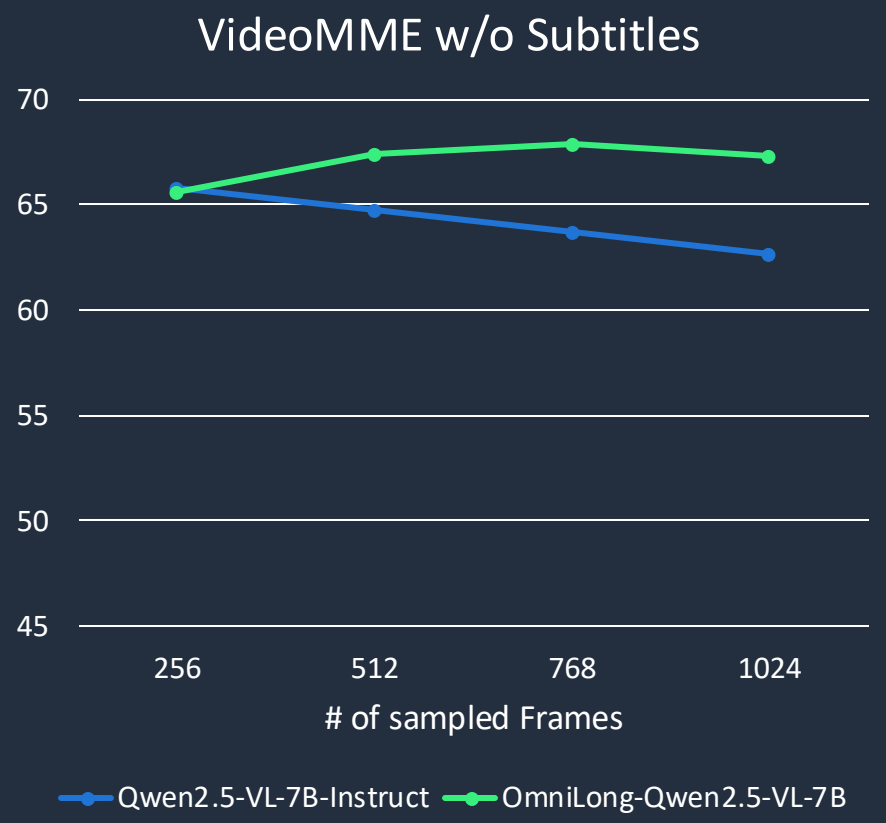
Screen shot from VideoMME website

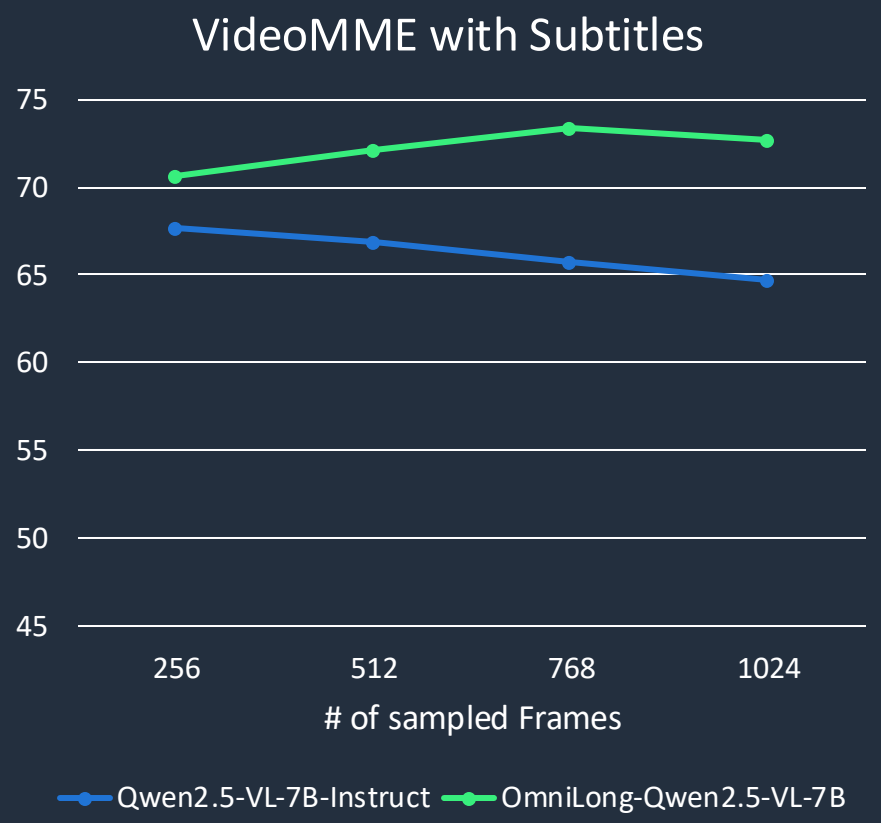
VideoMME Benchmark - OmniLong-LLaVA-OneVision-7B





VideoMME Benchmark - OmniLong-Qwen2.5-VL-7B

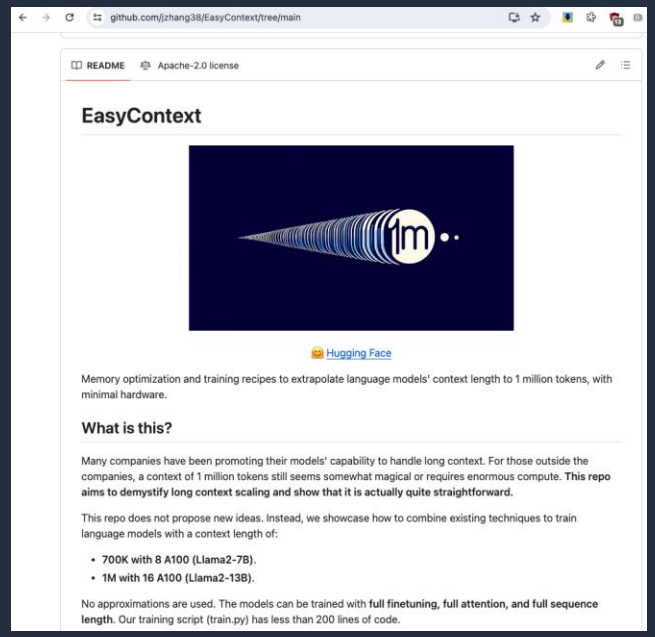




VideoMME Leadboard

Model Name	# of Params	Overall (%) - w/o subs	Overall (%) - w subs
OmniLong-Qwen2.5-VL-7B	7B	**67.9**	**73.4**
Qwen2.5-VL-7B-Instruct	7B	65.1	71.6
OmniLong-LLaVA-OneVision-7B	7B	**60.7**	**67.1**
llava-onevision-qwen2-7b-ov-hf	7B	58.2	61.5
LongVA	7B	52.6	54.3

Validated Long Context Finetuning Solutions on AWS



https://github.com/jzhang38/EasyContext/

AWS Sagemaker AI

PyTorch

Try OmniLong-Qwen2.5-VL-7B and stay tuned!

https://huggingface.co/aws-prototyping/long-llava-qwen2-7b

Thank you!

Questions and Discussions