Video-XL: Extra-Long Vision Language Model for Hour-Scale Video Understanding

CVPR (2025)

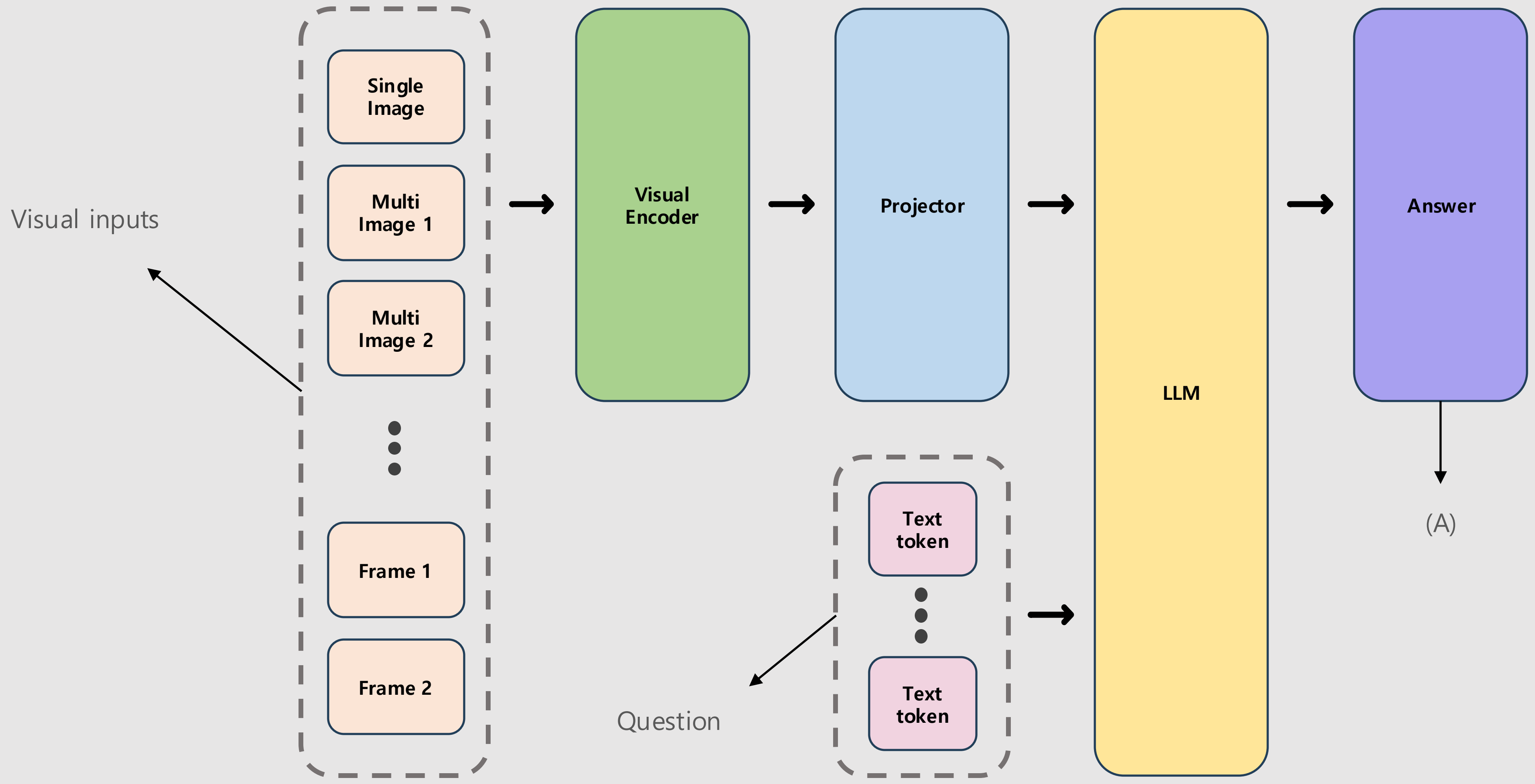
Yan Shu(1, 2) Zheng Liu(2) Peitian Zhang(3) Minghao Qin(4)
Junjie Zhou(5) Zhengyang Liang(2) Tiejun Huang(6) Bo Zhao(1, 2)

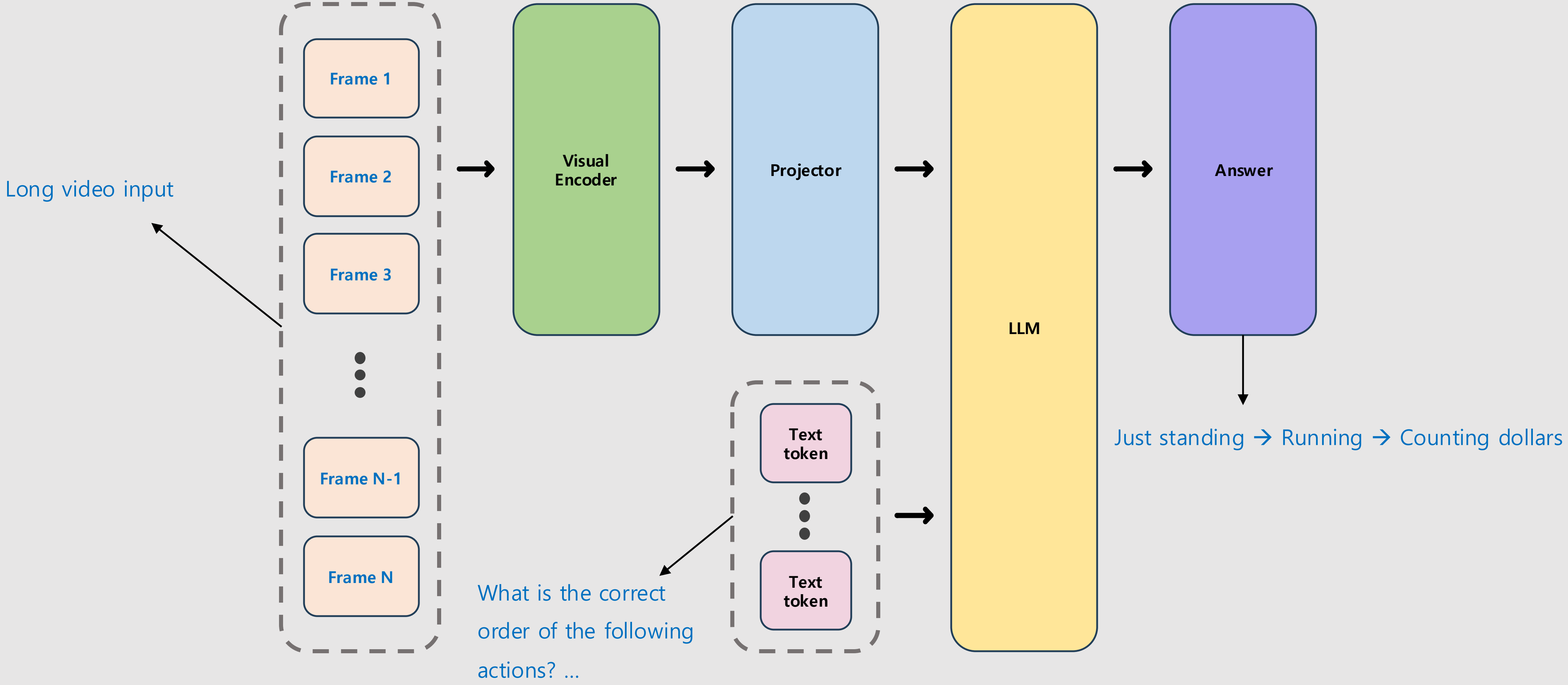
(1) Shanghai Jiaotong University (2) Beijing Academy of Artificial Intelligence (3) Renmin University of China
(4) Chinese Academy of Sciences (5) Beijing University of Posts and Telecommunications (6) Peking University

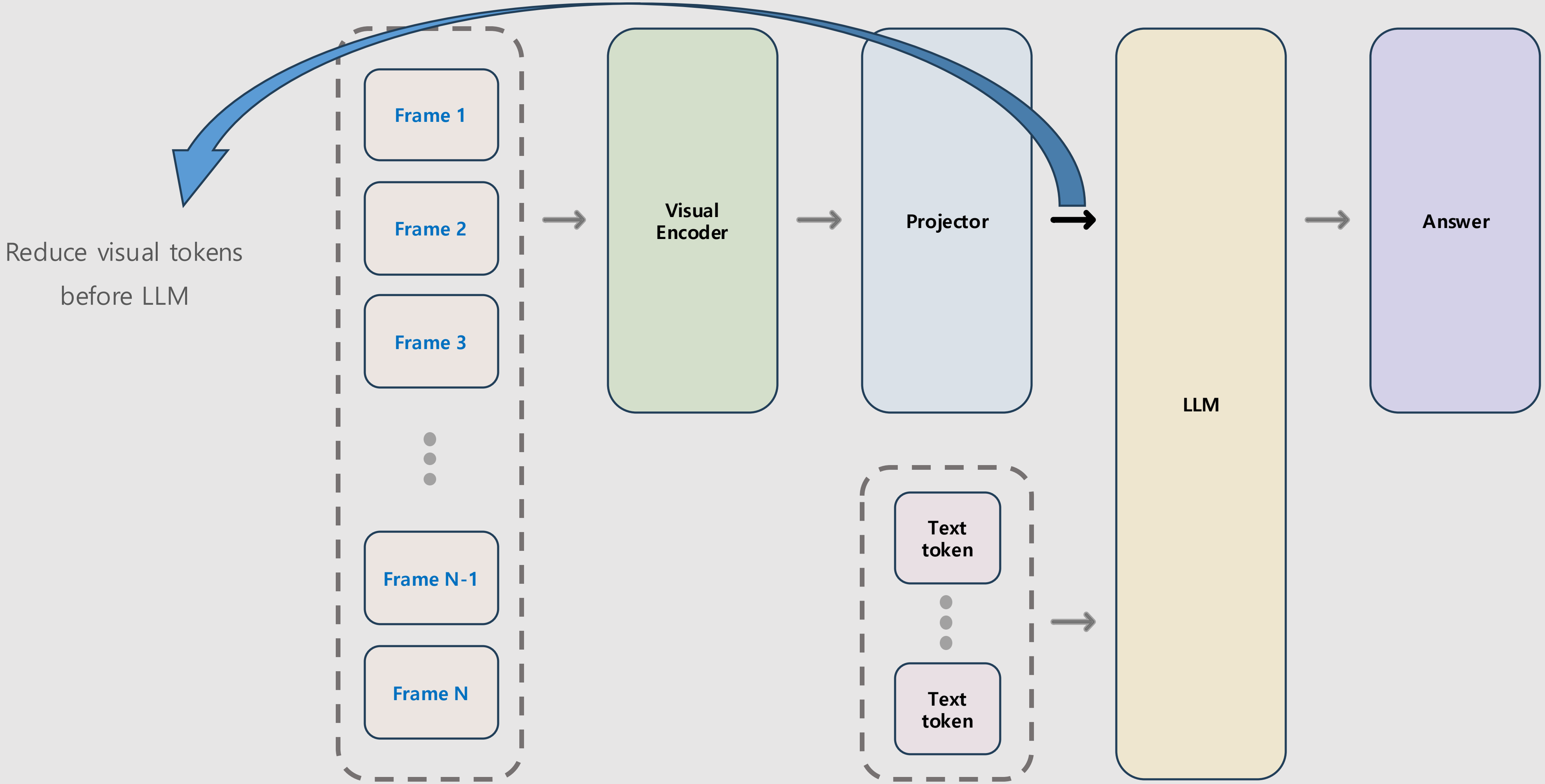
Jaemin Kim

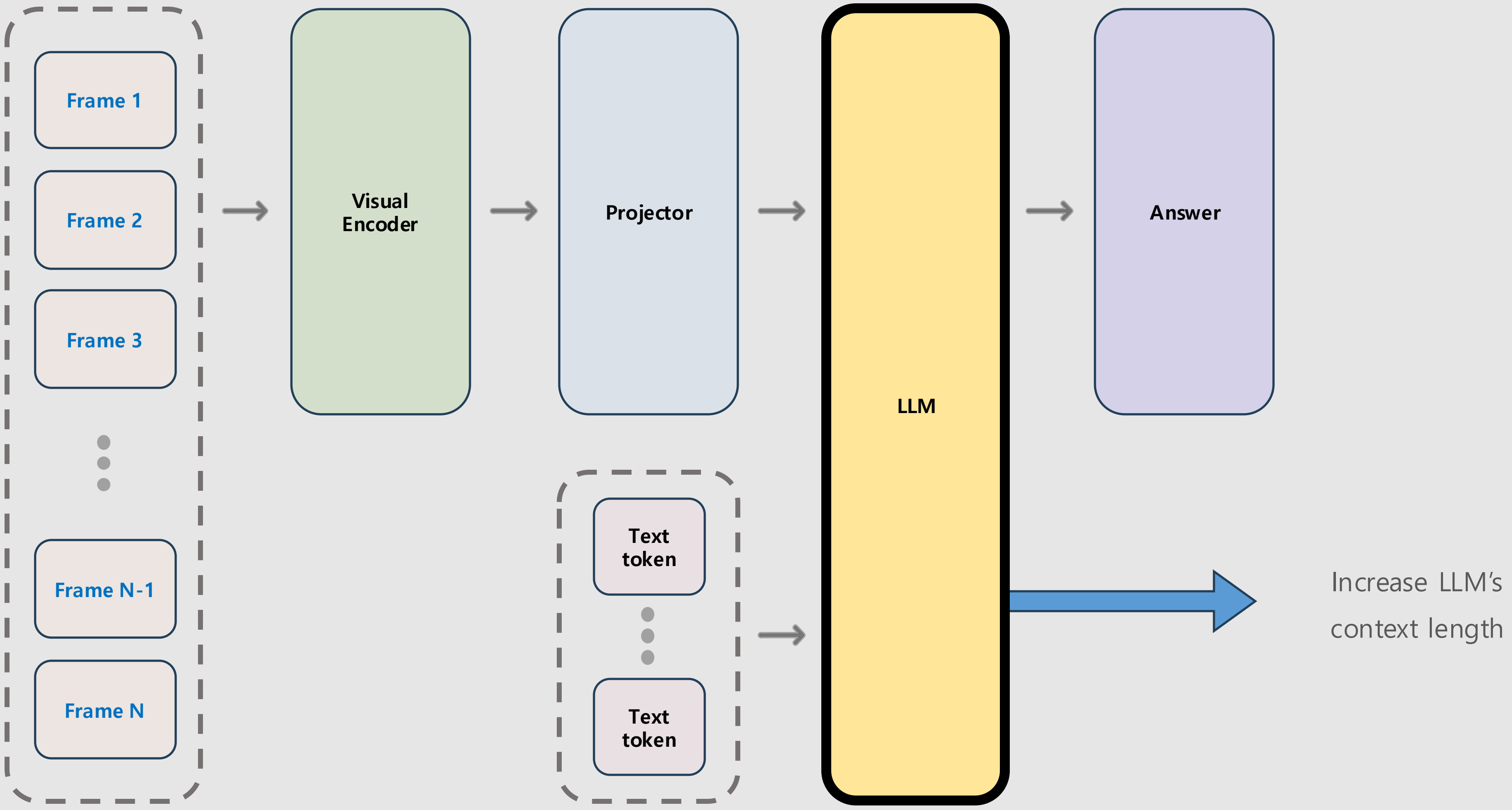
2025-12-05

Contents


Visual inputs
Single Image
Multi Image 1
Multi Image 2
Frame 1
Frame 2
Visual Encoder
Projector
LLM
Answer
(A)
Text token
Text token
Question


Long video input
Frame 1
Frame 2
Frame 3
Frame N-1
Frame N
Visual Encoder
Projector
LLM
Answer
Text token
Text token
What is the correct order of the following actions? …
Just standing → Running → Counting dollars


Reduce visual tokens
before LLM
Frame 1
Frame 2
Frame 3
Frame N-1
Frame N
Visual Encoder
Projector
Text token
Text token
LLM
Answer

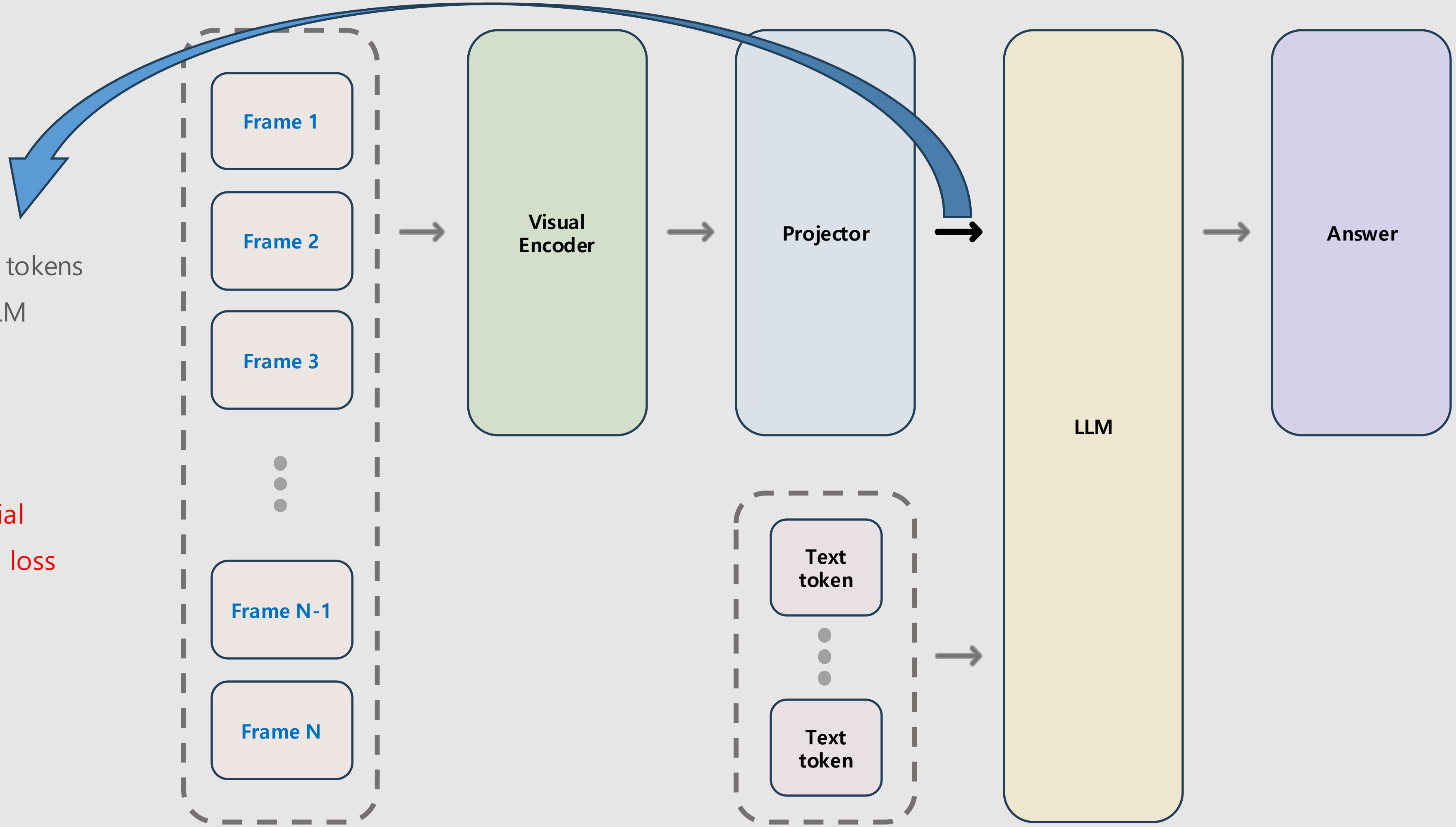

Frame 1
Frame 2
Frame 3
Frame N-1
Frame N
Visual Encoder
Projector
LLM
Answer
Text token
Text token
Increase LLM's context length

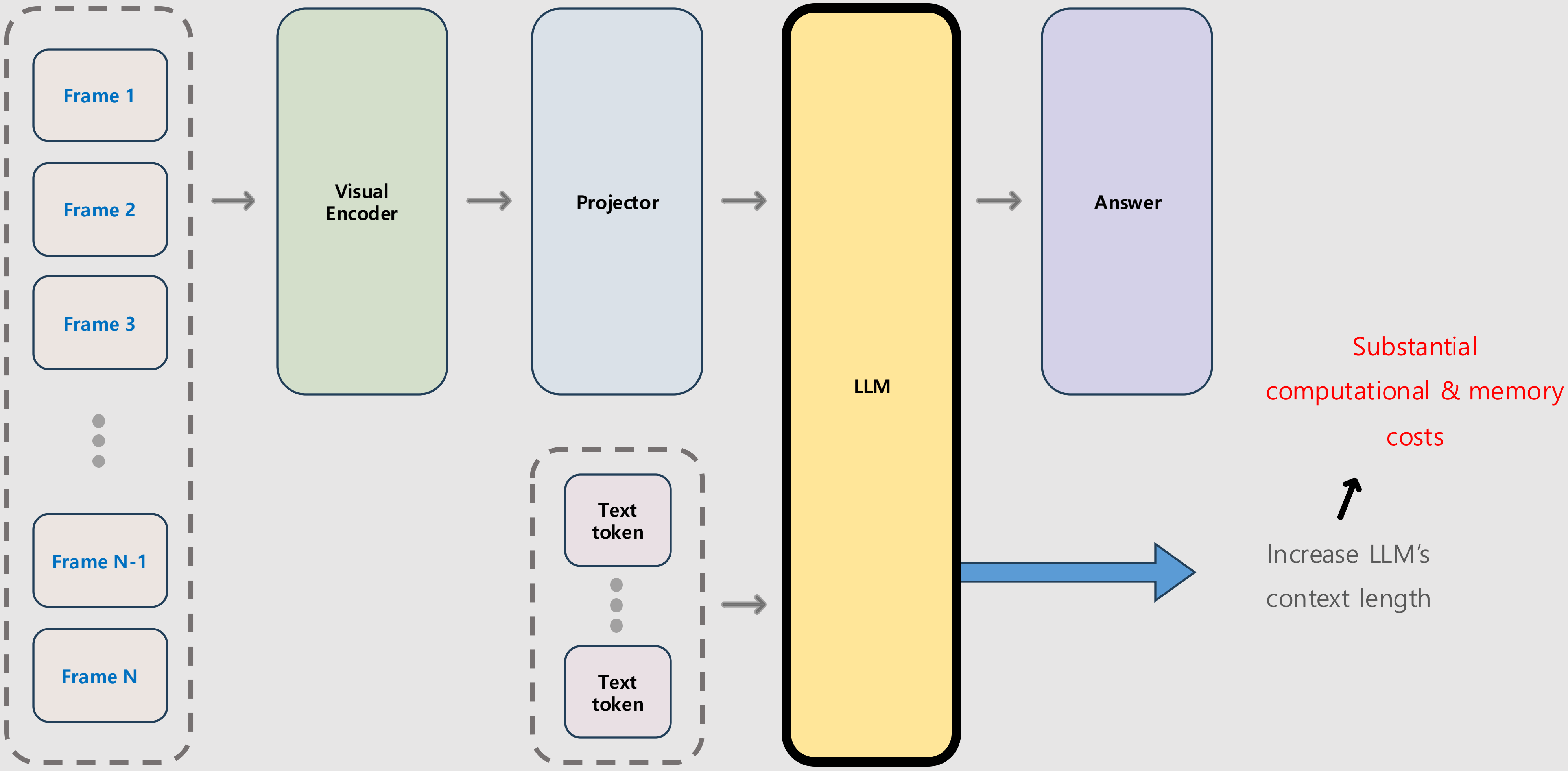
GSPS

Reduce visual tokens before LLM

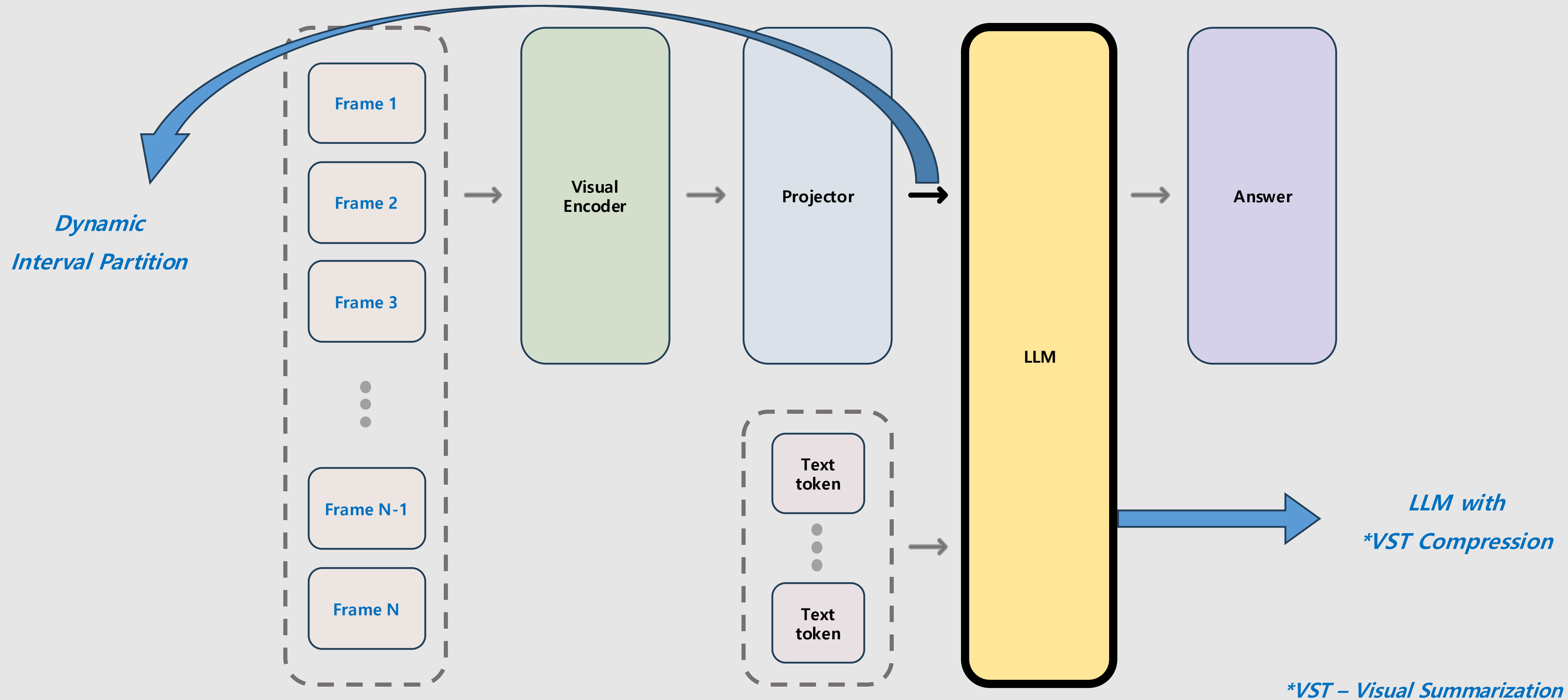
↓

Substantial information loss



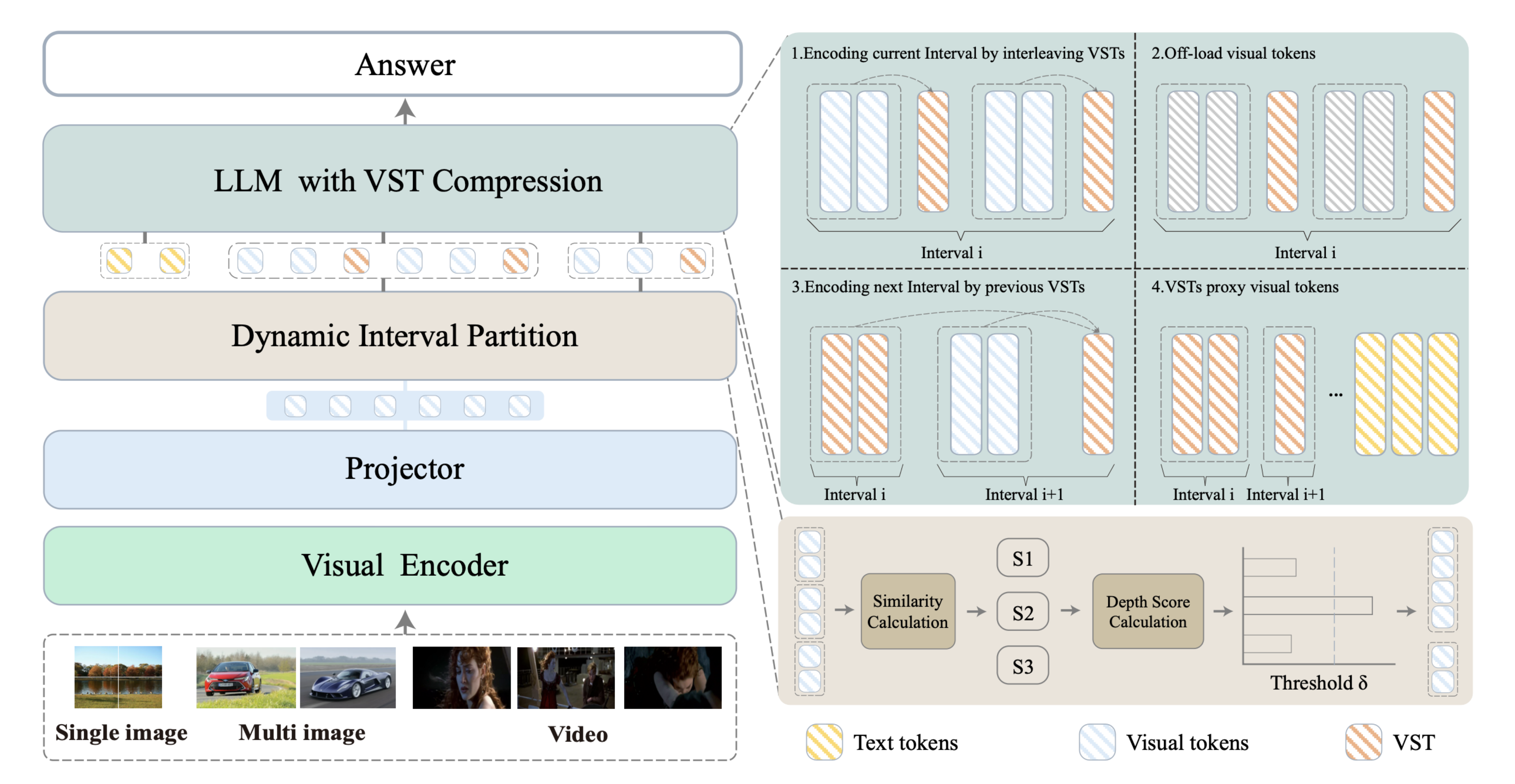

Frame 1
Frame 2
Frame 3
Frame N-1
Frame N
Visual Encoder
Projector
LLM
Answer
Text token
Text token
Increase LLM's context length
Substantial computational & memory costs

GSPS



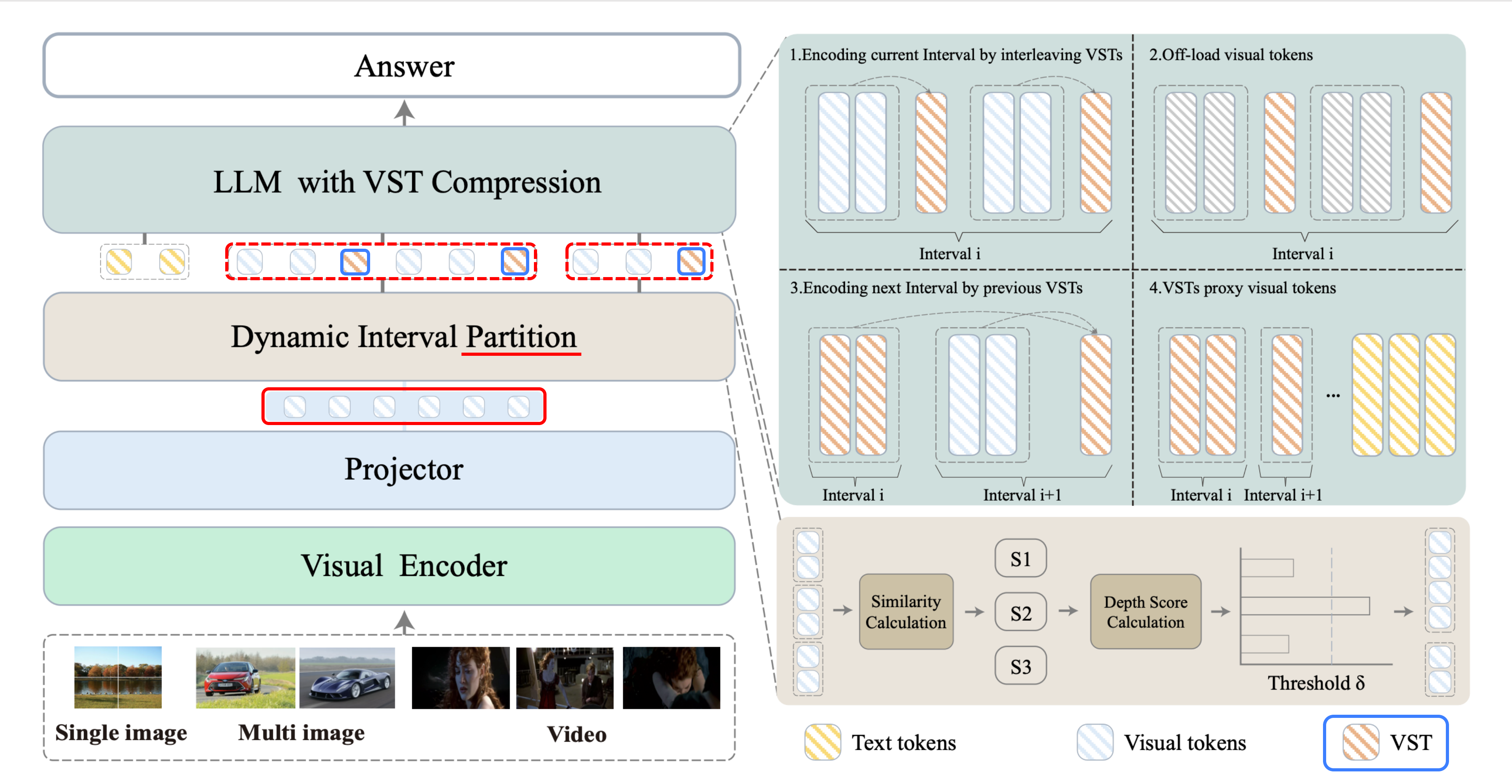
*VST – Visual Summarization Token

Overview of Video-XL

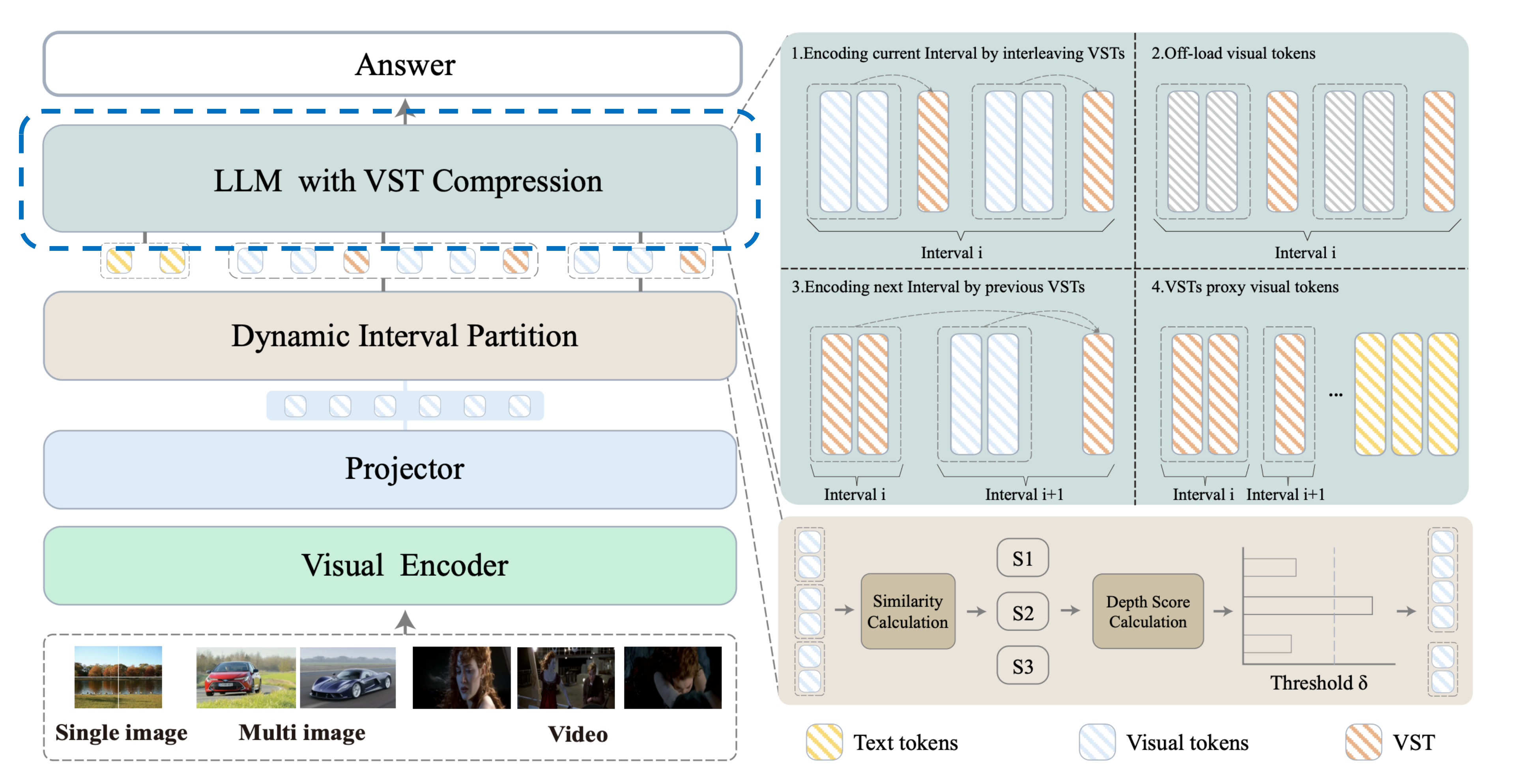


GSPS

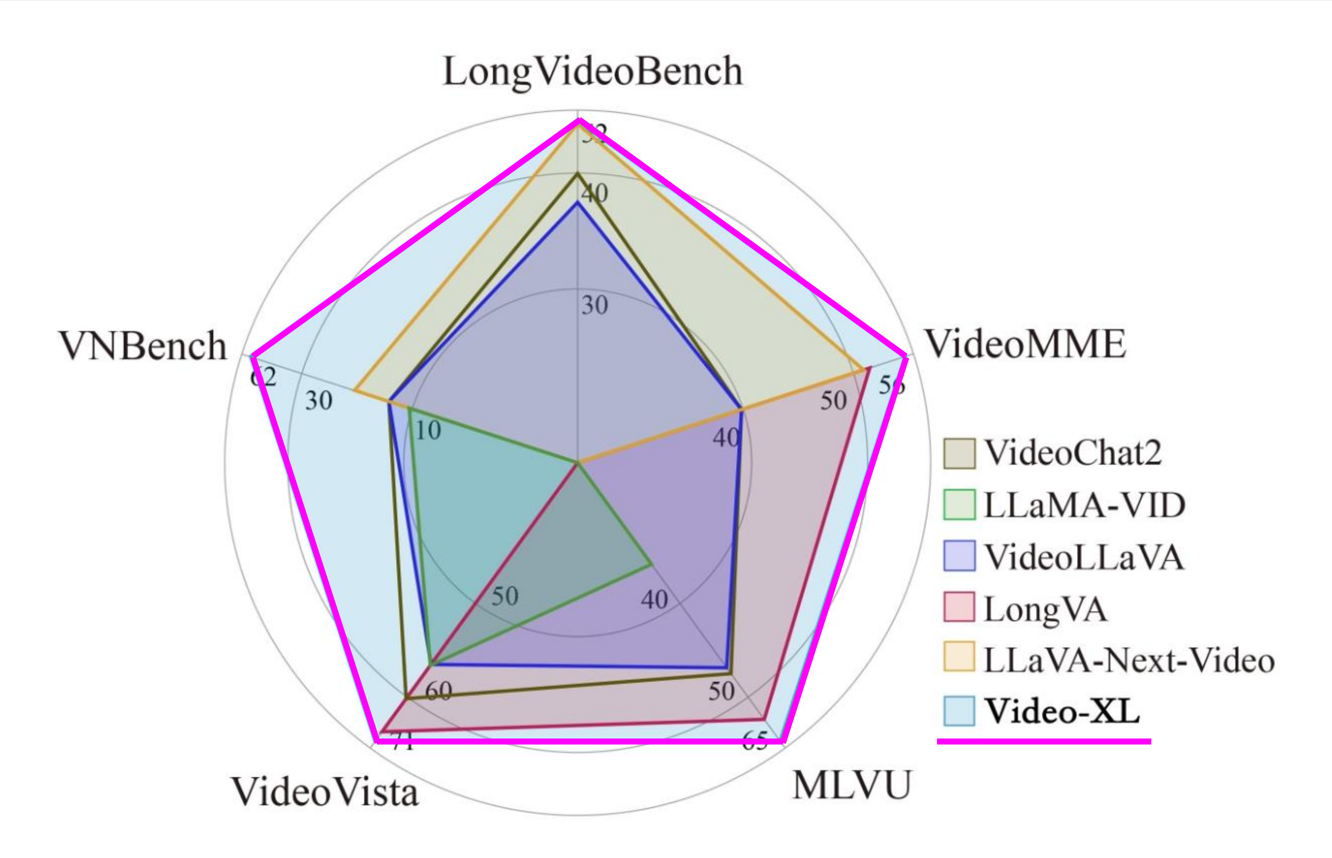
Overview of Video-XL



Overview of Video-XL



Video-XL shows state-of-the-art



Other studies

- VideoChat2
- LLaMA-VID
- VideoLLaVA
- LongVA
- LLaVA-Next-Video

Benchmarks

- LongVideoBench
- VNBench
- Video Vista
- MLVU
- VideoMME

Coming soon..

GSPS

Related Work

MovieChat, MA-LMM

Use long-term memory banks

LLaMA-VID

Reduces each frame to 2 tokens (context token & content token)

LongVLM, Video-CCAM

Token merging & causal cross-attention

→ Suffer from serious information loss

LWM (RingAttention)

Increases LLM's context length by RingAttention

LongVA

Increases LLM's context length by long context fine-tuning

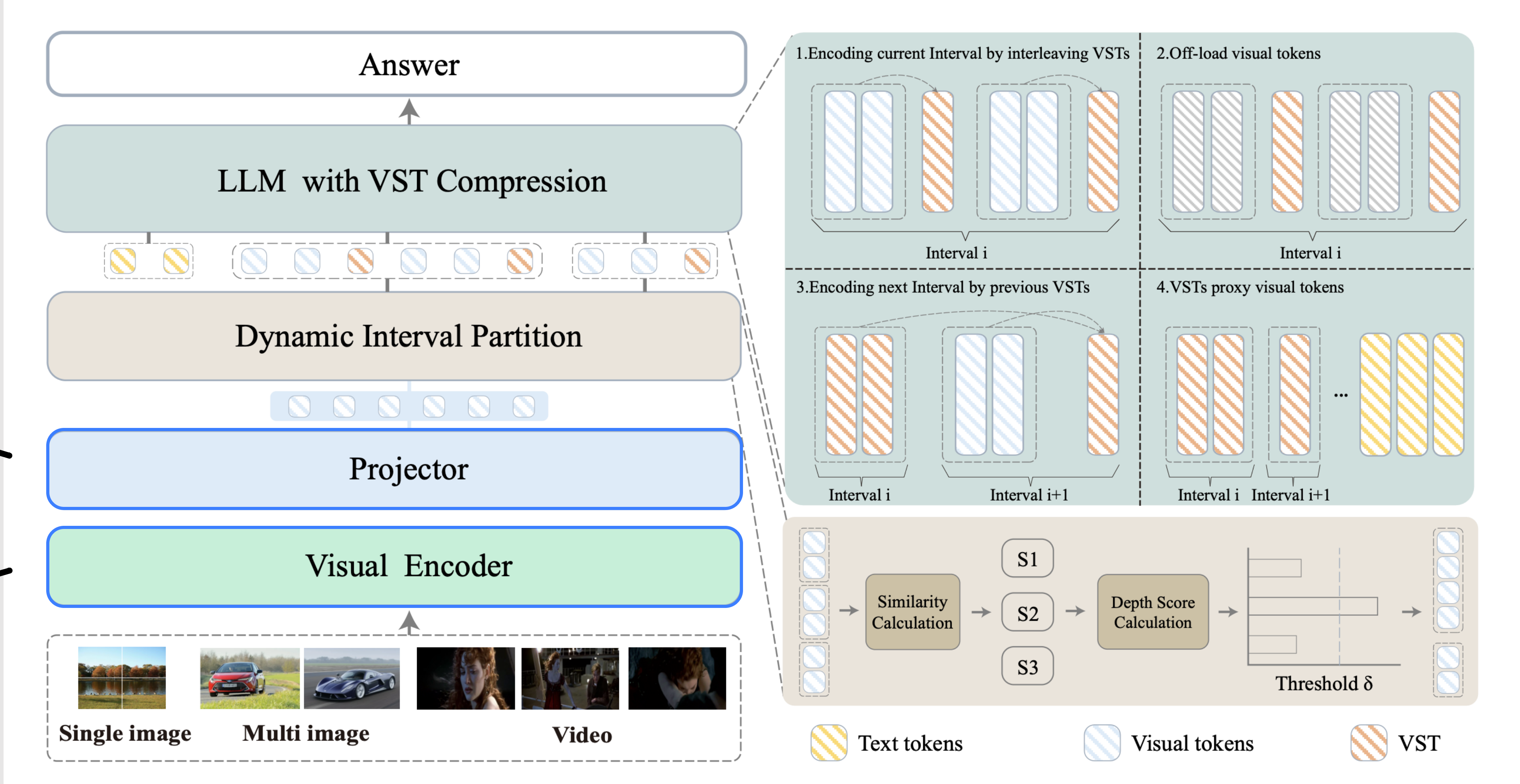
→ Substantial computational and memory costs

Video-XL (Overview)

Overview of Video-XL

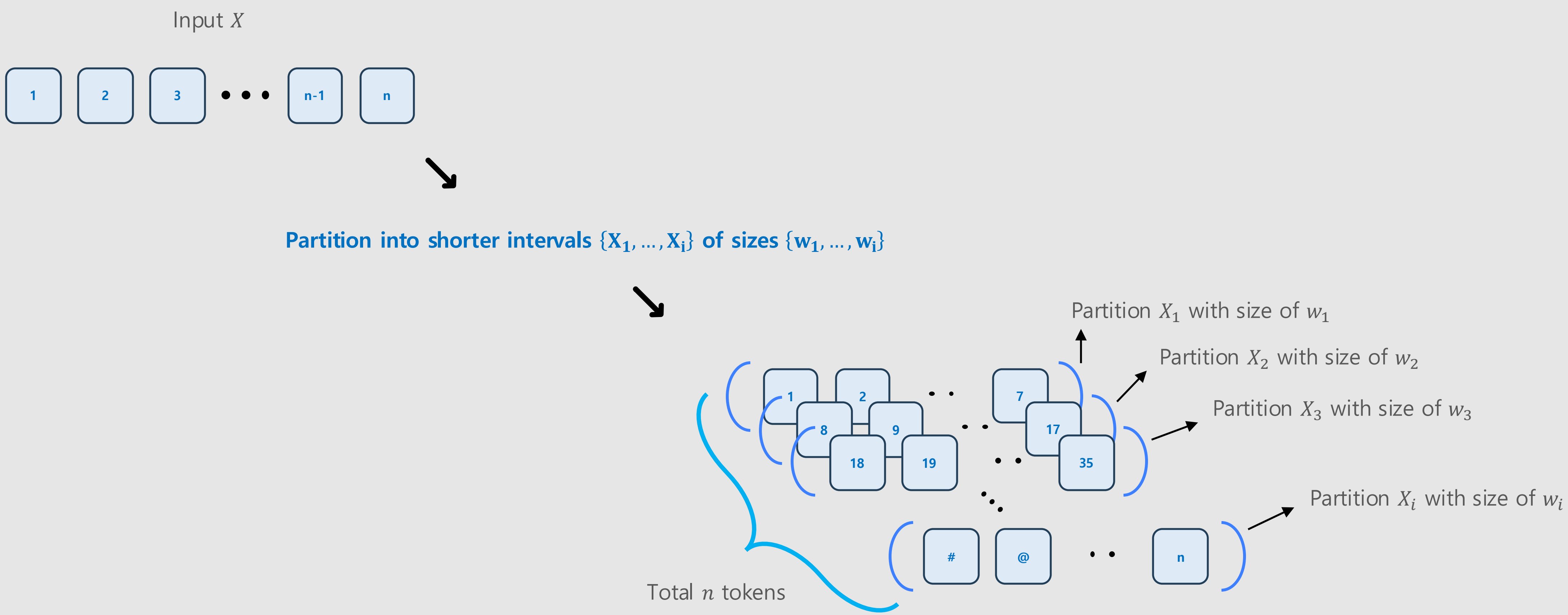
Two-layer MLP
+
GELU

CLIP-ViT-L



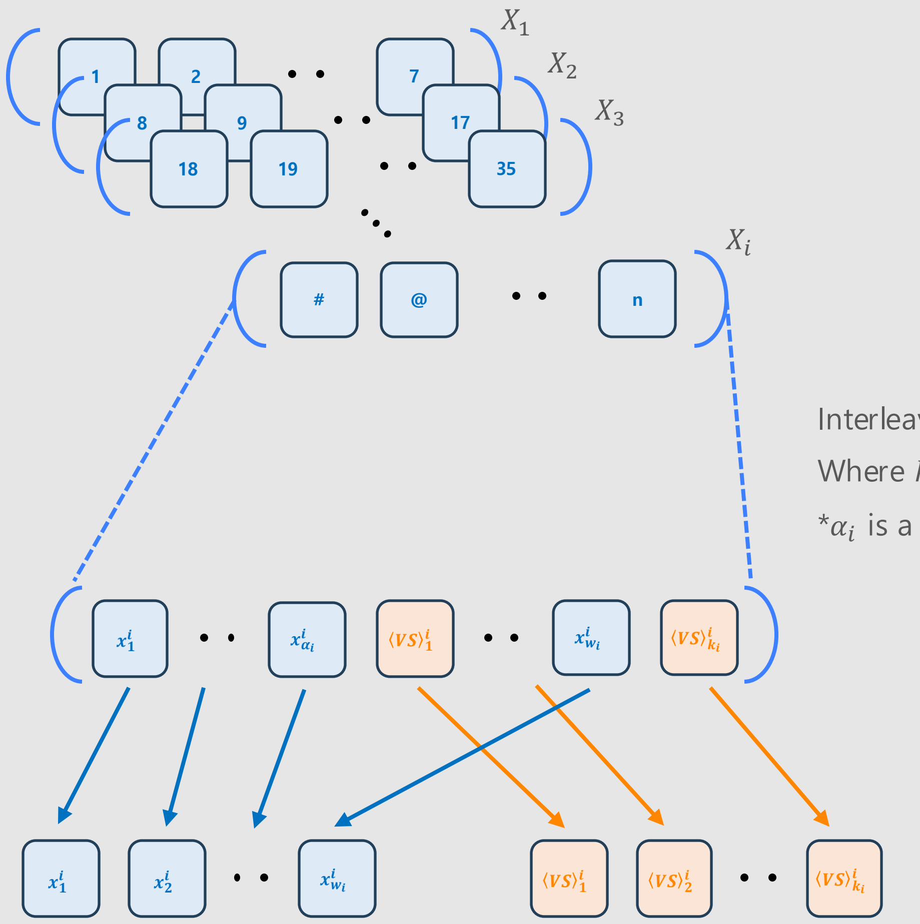
Video-XL: Extra-Long Vision Language Model for Hour-Scale Video Understanding

Compression mechanism



GSPS

Compression mechanism



Total w_i tokens for interval X_i

Total k_i VST tokens for interval X_i

Interleave k_i VSTs into interval

Where $k_i = w_i \,/\, \alpha_i$

*α_i is a compression ratio

[Example]

Interval X_7 with 20 tokens (w_i = 20)

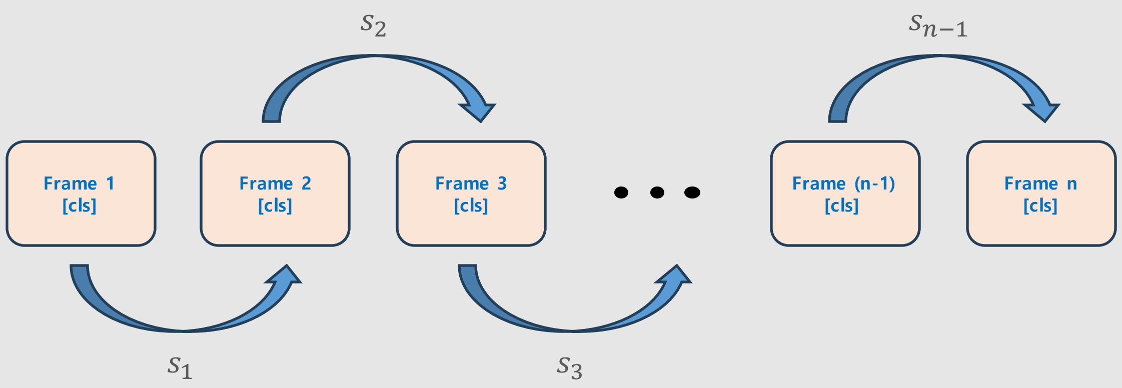
If compression ratio α_7 = 5

total $k_i = (w_i \,/\, \alpha_i) = (20/\,5) = 4$ VST tokens are interleaved into interval X_7

Dynamic compression strategy

Information density is variant for different parts of the video

→ Each interval's size should be different based on its density



s_i is the similarity score between neighboring frames

Based on similarity scores (s_i), we can estimate the consistency of visual semantic using the depth score d_i

$$d_i = \max(s_i \cdots s_{i-1}) + \max(s_{i+1} \cdots s_n) - 2 \times s_i$$

With threshold δ, where the peak scores satisfying $d_i > \delta$ are chosen as the boundaries of video intervals

Video-XL (VST Compression)

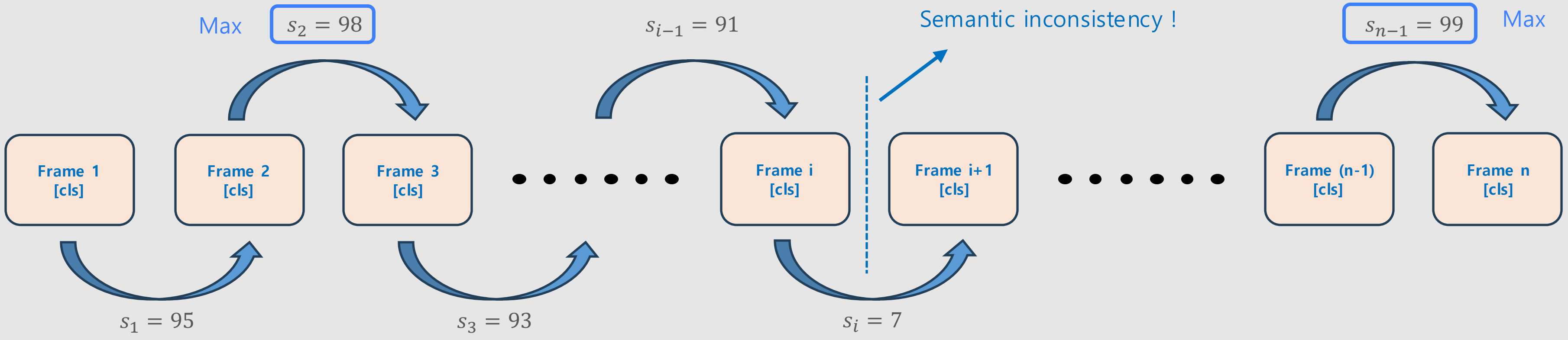
Dynamic compression strategy

Depth score d_i

$d_i = \max(s_i \cdots s_{i-1}) + \max(s_{i+1} \cdots s_n) - 2 \times s_i$

Information density is variant for different parts of the video

→ Each interval's size should be different based on its density



$$d_3 = \max(s_1, s_2) + \max(s_3 \cdots s_n) - 2 \times s_3$$
$$= 98 + 99 - 2 \times 93 = 11$$

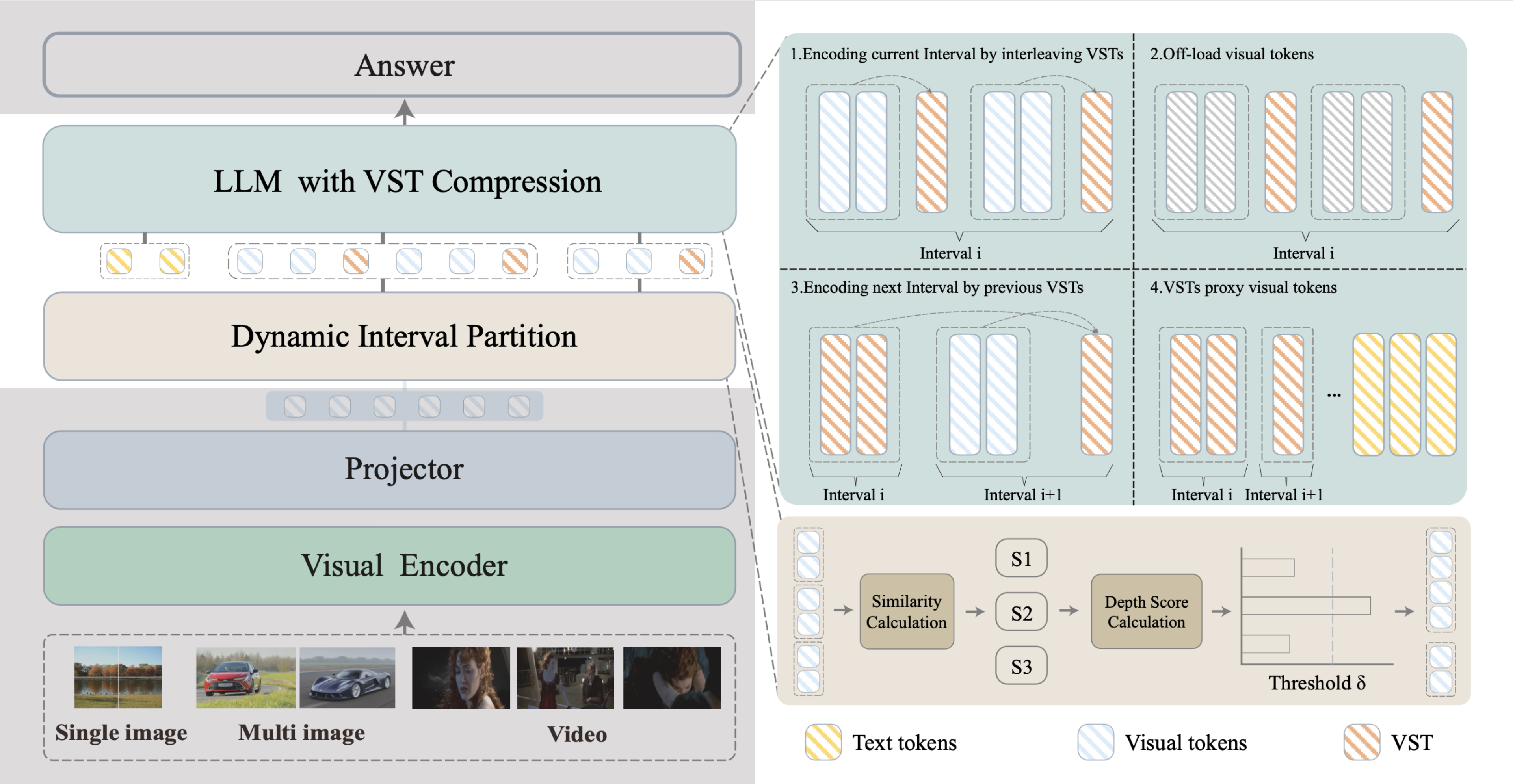
$$d_i = \max(s_i \cdots s_{i-1}) + \max(s_{i+1} \cdots s_n) - 2 \times s_i$$
$$= 98 + 99 - 2 \times 7 = \boxed{183}$$

Huge depth score

between frame i & frame i+1

s_i is the similarity score between neighboring frames

Overview of Video-XL



GSPS

Objective function

Video-XL is trained by instruction tuning

Generation probability of the next token t_{i+1}

$$\Pr(t_{i+1} \mid \underbrace{\langle VS \rangle_1^1, \cdots, \langle VS \rangle_{k_j}^j}_{\text{compressed KVs}}, \overbrace{s_i, \cdots, s_M}^{\text{instruction}}, \underbrace{t_1, \cdots, t_i}_{\text{ground-truth}}; \Theta)$$

Θ is learnable parameters of the MLLM and VST module

Curriculum learning

VST module is expected to support a wide range of compression ratios (α_i)

→ Video's different length and different information densities

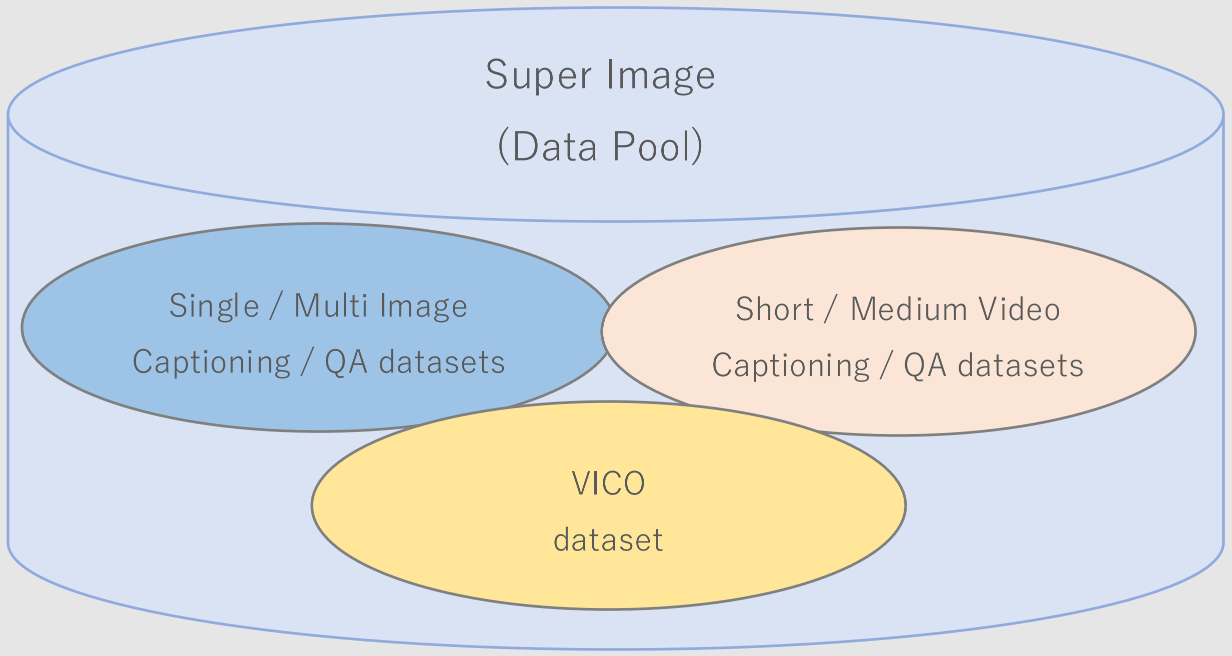
Two stages of curriculum learning

1) Random sample small compression ratios from (2, 4)

2) Gradually improve the candidate compression ratios to 8, 12, and 16 → (2, 16)

Composite Data curation

Due to the scarcity of long-video instruction tuning data..



[Single Image]

_Bunny

_Sharegpt-4o (57k)

_MMDU (20k)

[Multi Image]

_NExT-QA (32k)

_Sharegpt-4o (2k)

_CinePile (10k)

_VCG (25k)

_in-house video captions with GPT-4V (11k)

[VICO (Visual Clue Order]

_Synthetic dataset (20k QA pairs)

Composite Data curation

VICO (Visual Clue Order)

- 20k QA pairs
- Video of 3minutes on average (sourced from CinePile)

Generated by VILA-1.5

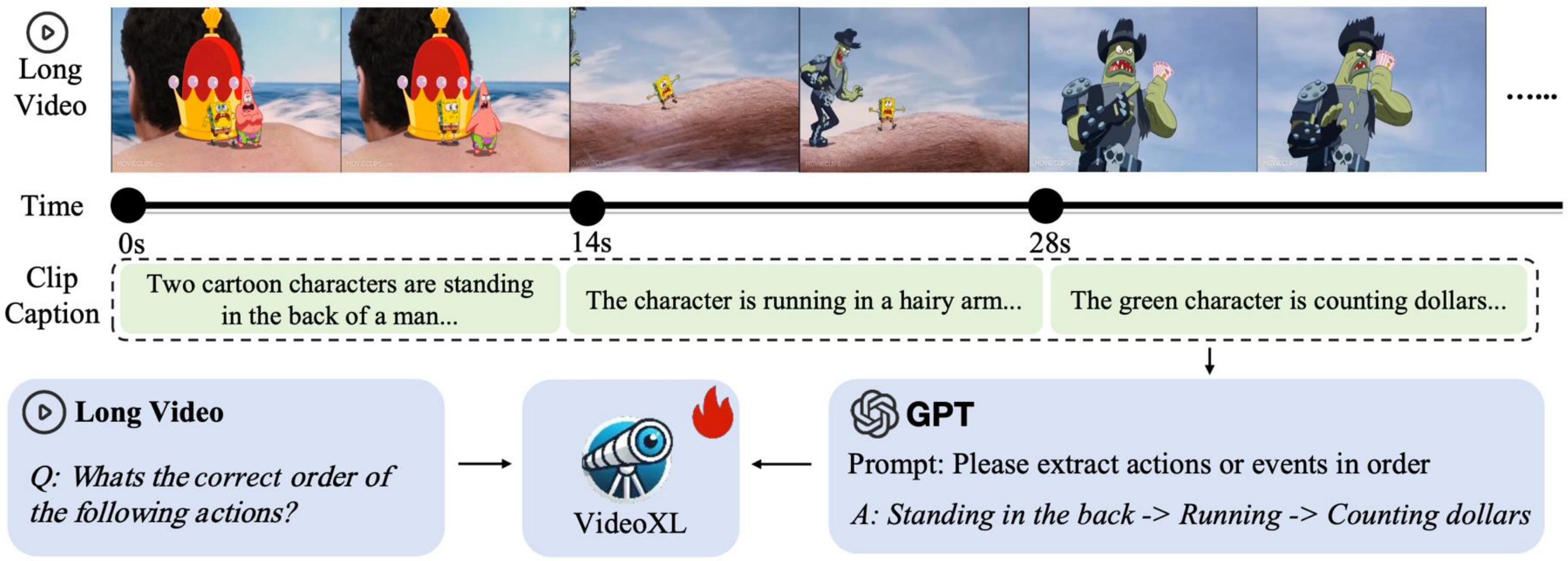
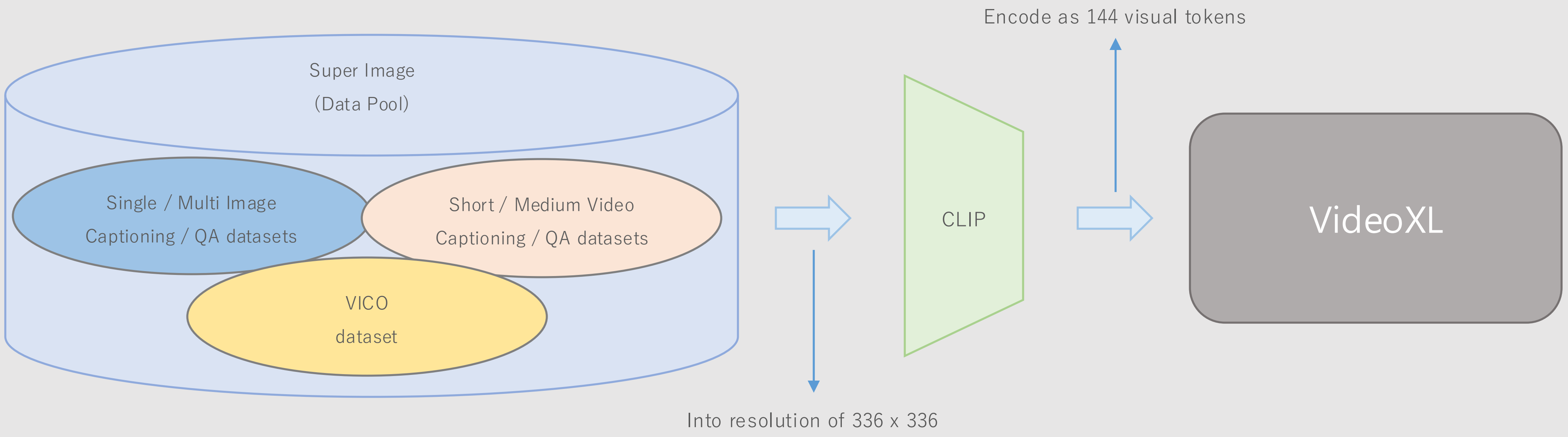


Figure 3. The pipeline of VICO generation. First, we generate short-clip captions for each few-second segments of the video. Then, we use GPT to extract key actions and events following the temporal order. Finally, the long video and QA pair are presented for model training.

Composite Data curation

Due to the scarcity of long-video instruction tuning data..



Implementation

Trained on Qwen-2-7B

Pre-training

Optimizing projector with Laion-2M dataset

(Batch size 8, Learning rate 5e-5)

Finetuning

Optimizing parameters of [Vision Encoder], [Projector], [LLM] with visual instruction tuning

(Bach size 1, Learning rate 1e-5)

- A800-80G x 8

GSPS

Benchmarks

	Benchmarks	Remark
Long Video Evaluation	MLVU	Comprehensive benchmark with both multiple choice and generation tasks
	Video-MME	Extensive benchmark covering videos of diverse genres and lengths
	VNBench	Synthetic benchmark focused on long video handling tasks
	LongVideoBench	Require precise retrieval and reasoning over detailed multi-modal information
	Video-Vista	Aims to evaluate model's long-context reasoning ability
Short Video Evaluation	VideoChatGPT	Short video question answering benchmark
	MVBench	Short video question answering benchmark

GSPS

Main Results

Multiple-choice tasks average

Development set

Test set

Subtitle

Generation tasks average

Model	Size	MLVU Dev M-avg	MLVU Dev G-avg	MLVU Test M-avg	MLVU Test G-avg	VideoMME W/o sub	VideoMME W sub	VNBench	VideoVista	LongVideo.	VideoChat.	MVBench
Proprietary Models												
GPT-4V [34]	-	49.2	5.35	43.3	4.67	59.5	63.3	48.9	-	59.1	**4.06**	**43.5**
GPT-4o [35]	-	**64.6**	**5.80**	**54.9**	**5.87**	71.9	71.2	64.4	**78.3**	**66.7**	-	-
Gemini-1.5-Pro [38]	-	-	-	-	-	**75.0**	**81.3**	**66.7**	-	64.0	-	-
Open-source MLLMs												
VideoChat2 [17]	7B	47.9	3.99	35.1	3.99	39.5	43.8	12.4	61.6	39.3	2.98	62.3
LLaMA-VID [18]	7B	33.2	4.22	17.2	3.43	-	-	10.8	56.9	-	2.89	41.4
VideoLLaVA [20]	7B	47.3	3.84	30.7	3.68	39.9	41.6	12.4	56.6	39.1	2.84	43.0
ST-LLM [26]	7B	-	-	-	-	37.9	42.3	22.7	49.3	-	3.15	54.9
Shargpt4Video [3]	7B	46.4	3.77	33.8	3.63	39.9	43.6	-	53.6	39.7	-	51.2
LLaVA-Next-Video [52]	34B	-	-	-	-	52.0	54.9	20.1	56.7	50.5	**3.26**	-
PLLaVA [46]	7B	-	-	-	-	-	-	-	60.4	40.2	3.12	46.6
LongVA† [51]	7B	56.3	4.33	41.1	3.91	52.6	54.3	41.5	67.4	47.8	-	-
VideoLLaMA2† [4]	8x7B	-	-	-	-	47.9	49.7	24.9	60.5	36.0	**3.26**	53.9
Video-CCAM† [6]	9B	58.5	3.98	**42.9**	3.57	50.3	52.4	35.6	69.0	43.1	-	**64.6**
Long-LLaVA [41]	13B	-	-	-	-	51.9	-	52.1	-	-	-	-
Video-XL	7B	**64.9**	**4.50**	**45.5**	**4.21**	**55.5**	**61.0**	**61.6**	**70.6**	**50.7**	3.17	55.3

Table 1. Experimental results on mainstream video benchmarks. “LongVideo.” and “VideoChat.” refer to LongVideoBench and VideoChat-GPT Bench, respectively. † indicates that the results on VNBench and LongVideoBench were reproduced using their official weights.

Extra-Long Evaluation & Inference Efficiency

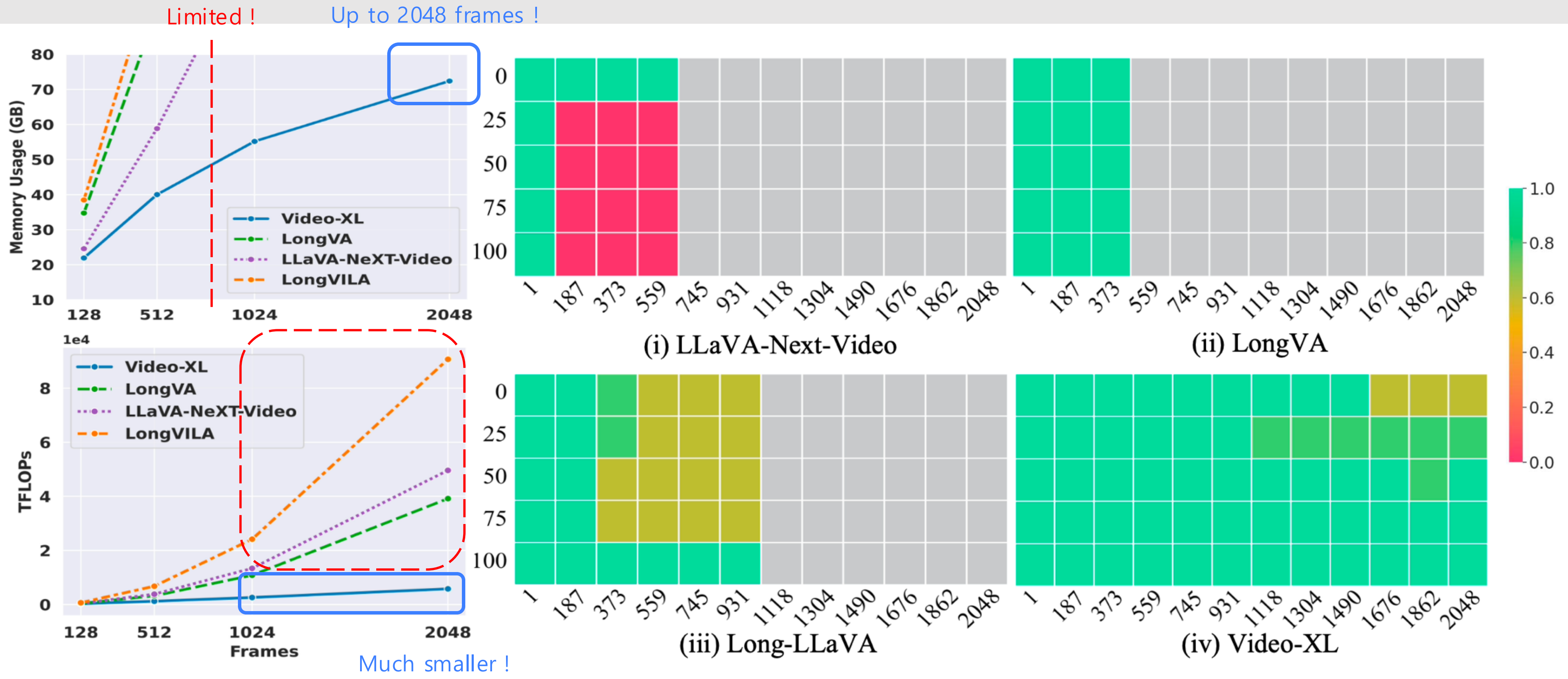


Figure 4. (Left) Comparison of the memory usage and the forward FLOPs of different models. (Right) Results on the Needle-in-a-haystack evaluation within a single A100 80GB GPU. The x-axis represents the total number of frames in the video haystack. The y-axis shows the position where the needle image is located. Gray grids mean "OOM'.

Extra-Long Evaluation & Inference Efficiency

Needle-in-a haystack

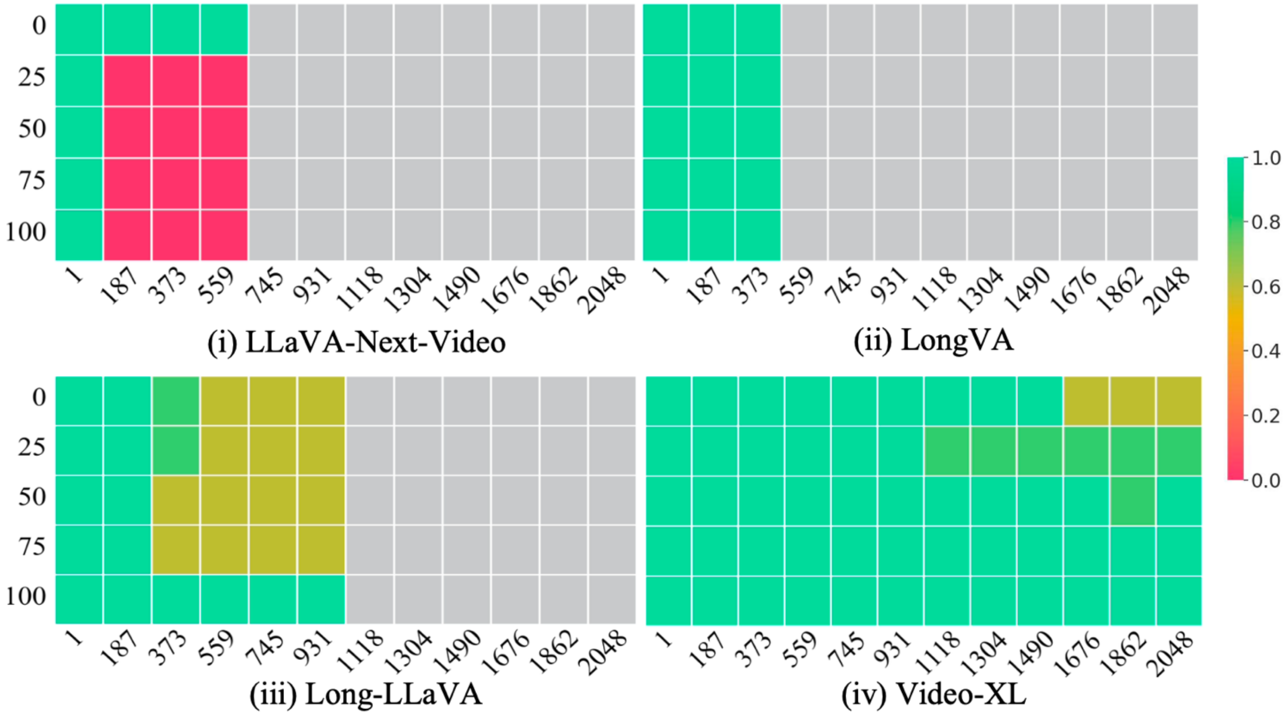


Figure 4. (Left) Comparison of the memory usage and the forward FLOPs of different models. (Right) Results on the Needle-in-a-haystack evaluation within a single A100 80GB GPU. The x-axis represents the total number of frames in the video haystack. The y-axis shows the position where the needle image is located. Gray grids mean "OOM".

Ablation Studies

Compression mechanism

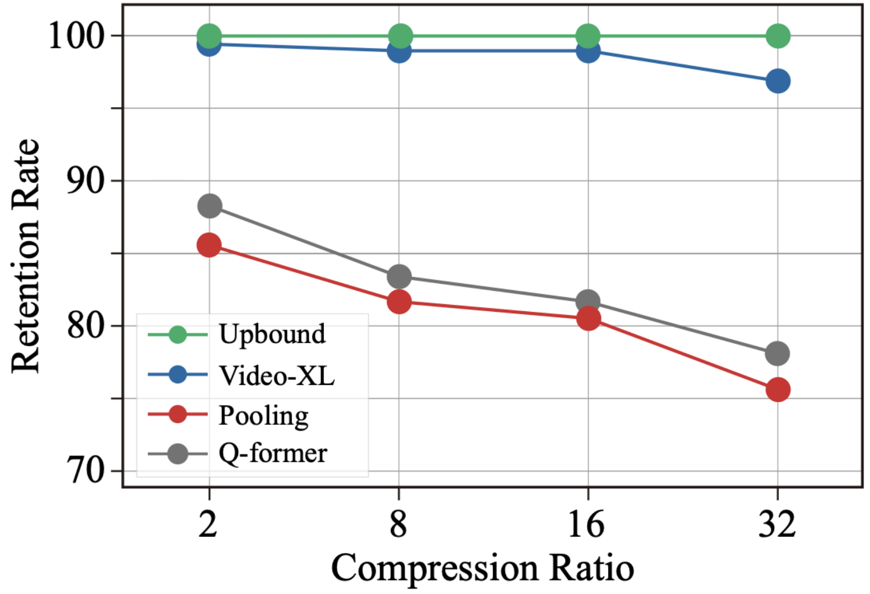


Figure 5. MLVU performance with variant compression ratios. The retention rate is calculated as the ratio to the upper-bound.

Model	MLVU	VideoMME	MME	MMB
Pooling	33.7	41.0	1405.5	62.3
Q-Former	35.1	42.1	1410.2	61.9
LLaMA-VID	35.5	45.7	1421.2	64.3
LLaMA-Adapter	35.3	42.2	1418.3	65.5
C-Abstractor	37.1	46.3	1440.2	65.1
Video-XL	41.4	52.0	1510.2	70.9
Upper-bound	41.8	52.6	1533.7	71.6

Table 2. Comparison of compression techniques. All methods are implemented in the same setting and conducted with 16× compression.

Pooling
- Average pooling

Q-Former
- Fixed Query embedding for visual inputs

LLaMA-VID
- Dual token(content & context) compression per frame

LLaMA-Adapter
- Adding learnable Adapter at LLM's transformer layer

C-Abstractor
- Similarity based token merging

GSPS

Ablation Studies

Dynamic compression strategy & Curriculum learning

Fixed compression based on an interval of 1440 tokens

Train	Test	MLVU	VideoMME	MME	MMB
✗	✗	39.8	50.9	1460.6	70.9
✗	✓	39.6	50.8	1455.0	70.8
✓	✗	41.5	52.0	1515.5	71.2
✓	✓	**41.6**	**52.3**	**1520.0**	**71.3**

Table 3. Evaluation of dynamic compression strategy.

No improvements are obtained
if it's only enabled for testing stage

Settings	MLVU	VideoMME	MME	MMB
w/o random compre.	40.5	51.0	1500.4	70.3
w/o curriculum learn.	41.1	51.6	1512.4	71.0
Ours	**41.6**	**52.3**	**1520.0**	**71.3**

Table 4. Evaluation of curriculum learning.

Fixed 16x compression ratio

Random (2, 16)x
compression ratio

GSPS

Ablation Studies

Composite data curation

Marginal

Significant improve by multi-image

Video	Single Image	Multi Image	TR	NQA	AO	Avg
100k	-	-	73.4	64.5	53.6	63.8
100k	350k	-	77.5	66.9	54.0	66.1
100k	700k	-	80.6	70.0	54.1	68.2
100k	1M ↑	-	81.3 ↑	69.8 ↓	53.8 ↓	68.3 ↑
100k	700k	20k	**82.0** ↑	**70.3** ↑	55.3 ↑	**69.5** ↑
100k	700k	40k	82.1	70.1	**55.4**	69.2

Table 5. Analysis of training effect from different data.

TR – Topic Reasoning

for holistic understanding capability

NQA – Needle QA

for single-detail understanding capability

AO – Action Order

for multi-detail understanding capability

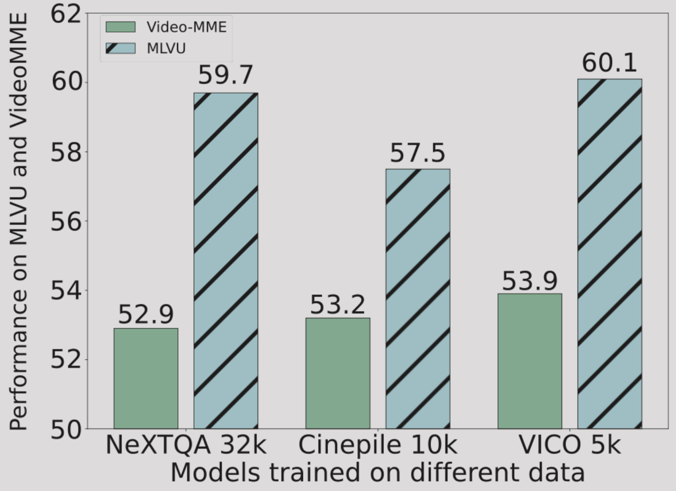


Figure 6. Analysis of training effect from VICO.

GSPS

Ablation Studies

Composite data curation

Re-trained the model using

three video instruction-tuning datasets

VICO was the smallest one (5k),

but outperforms the other two datasets (32k, 10k)

Video	Single Image	Multi Image	TR	NQA	AO	Avg
100k	-	-	73.4	64.5	53.6	63.8
100k	350k	-	77.5	66.9	54.0	66.1
100k	700k	-	80.6	70.0	54.1	68.2
100k	1M	-	81.3	69.8	53.8	68.3
100k	700k	20k	**82.0**	**70.3**	55.3	**69.5**
100k	700k	40k	82.1	70.1	**55.4**	69.2

Table 5. Analysis of training effect from different data.

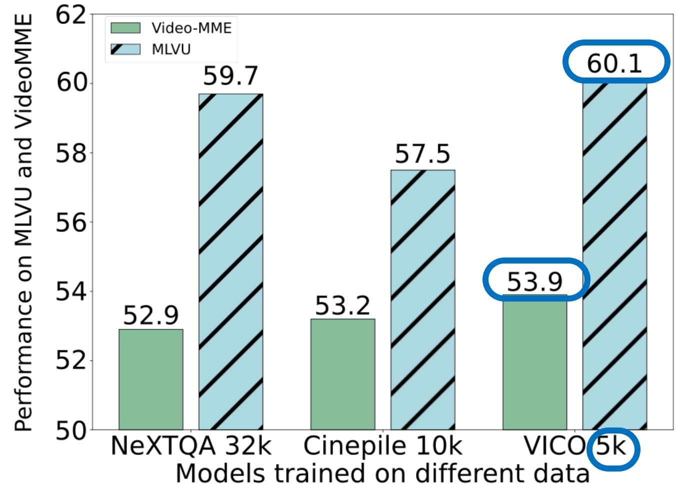
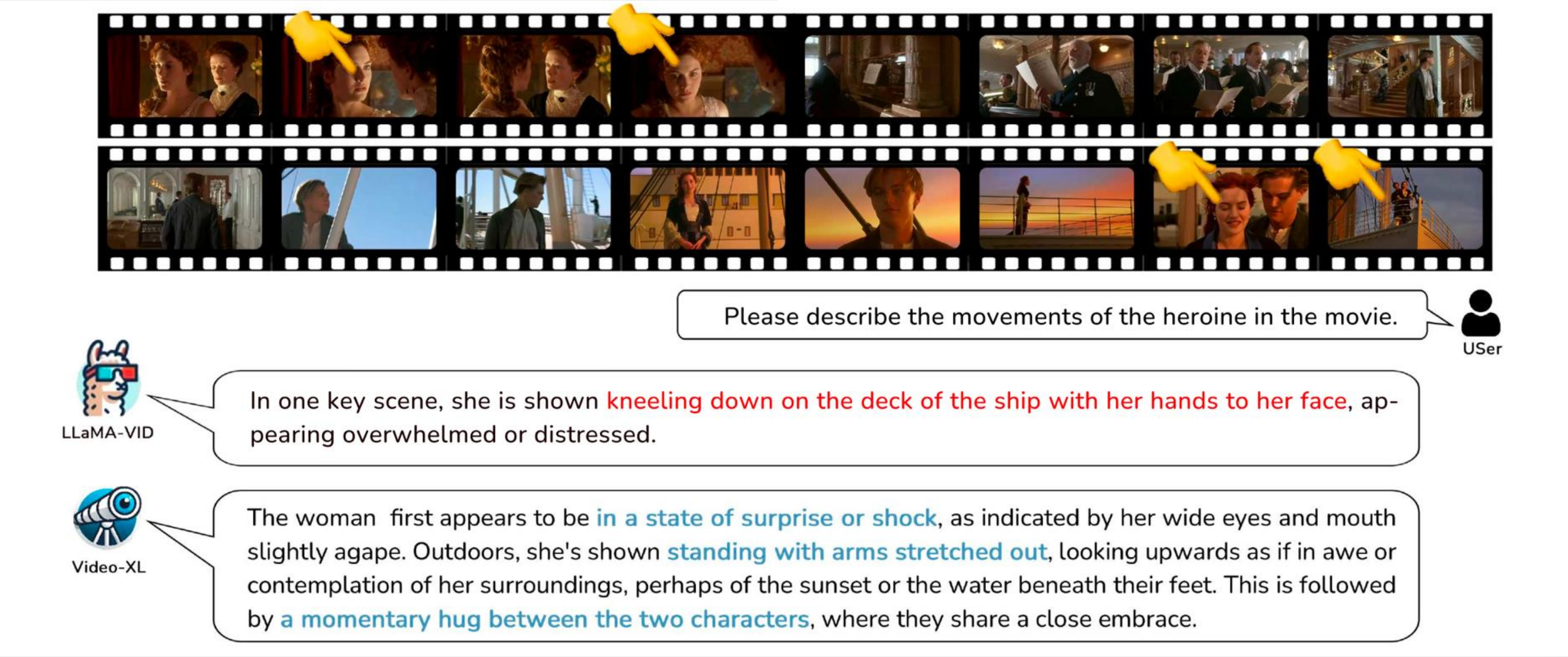


Figure 6. Analysis of training effect from VICO.

Experiment & Result

Qualitative Evaluation

Based on extra-long videos
(over 30 minutes)

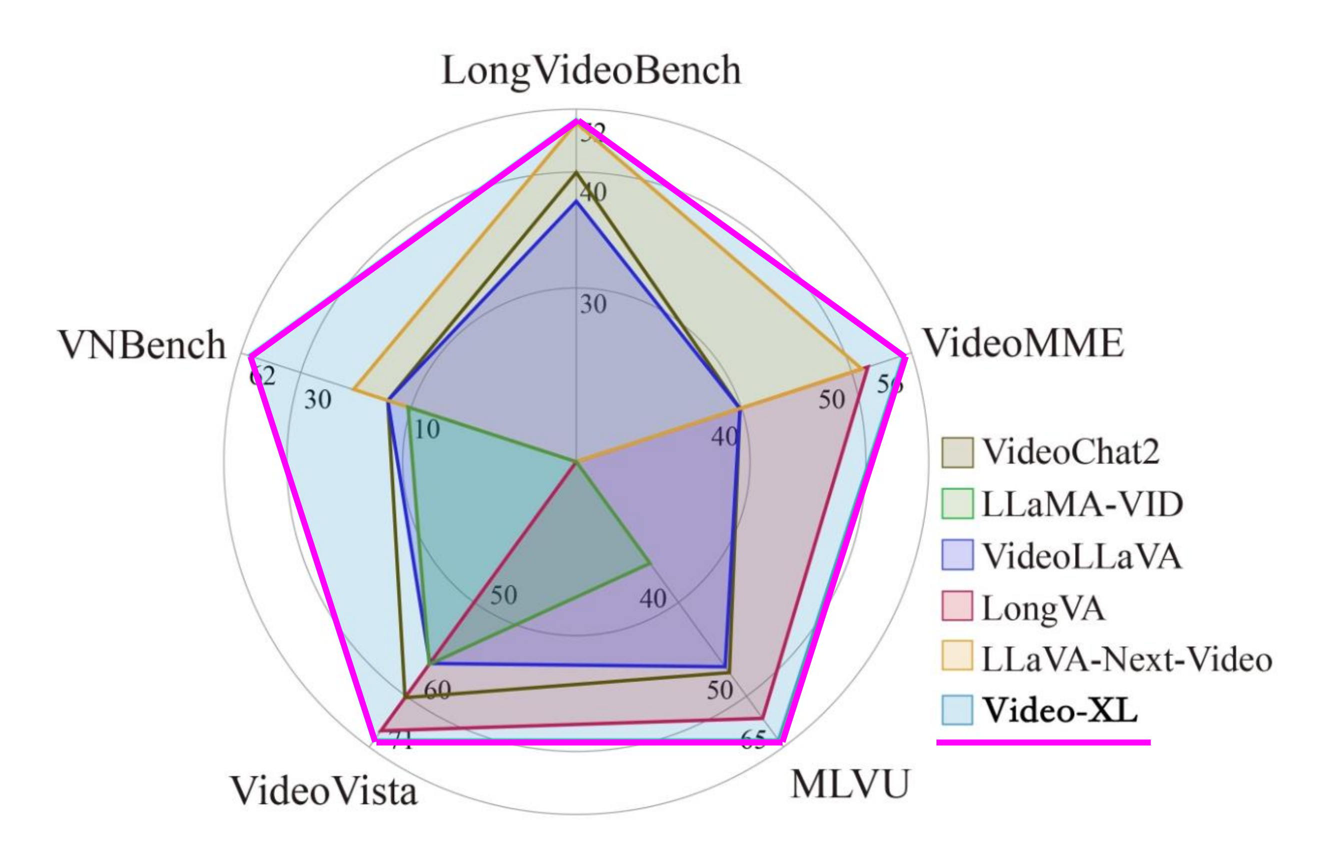


GSPS

Video-XL

Long video understanding ✓

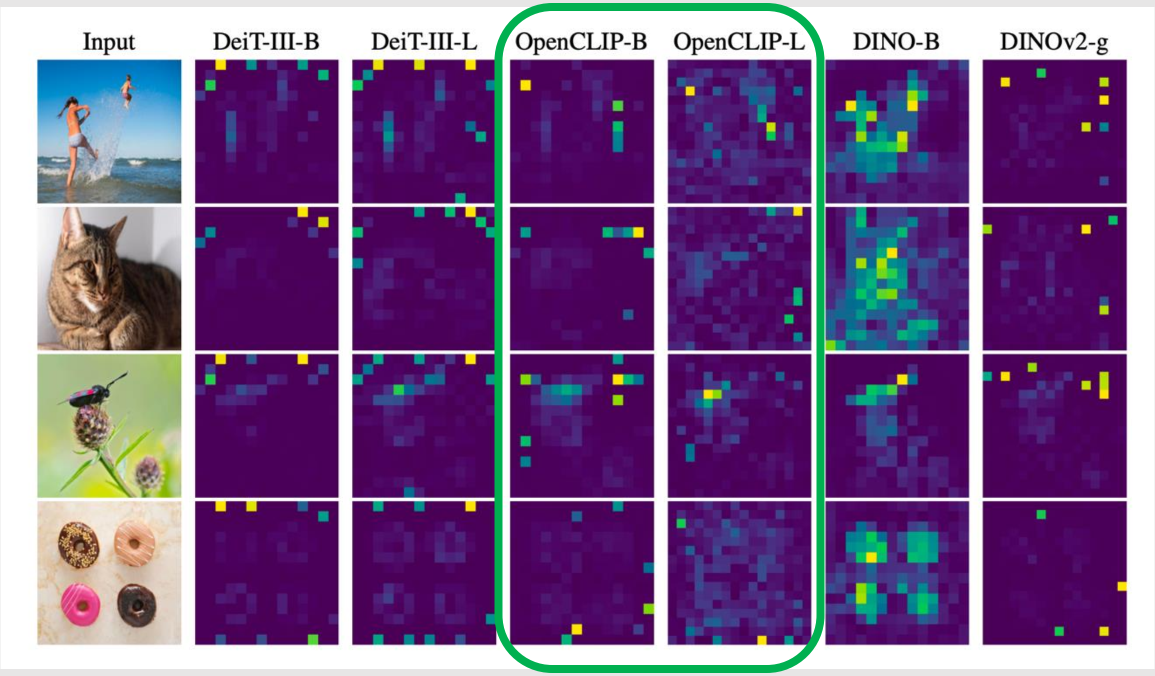
More efficient ✓



Conclusion & Discussion

Vision encoder outputs visual sink tokens

Visual sink tokens hold global information



	IN1k	CF10	CF100
CLS token	85.6	99.4	93.4
central token	73.3	98.0	88.1
outlier token	84.5	99.2	92.8
trained register	83.1	99.2	93.0
test-time register	84.5	99.1	93.0

Table 1: **Linear probing classification results (DINOv2 ViT-L/14).** Test-time registers achieve higher performance on linear probing than non-outlier tokens, suggesting that they hold global information similarly to trained registers. They match the performance of outlier tokens, indicating that they have absorbed the role of outliers.

	IN1k	P205	Airc.	CF10	CF100	CUB	Cal101	Cars	DTD	Flow.	Food	Pets	SUN	VOC
[CLS]	**86.0**	**66.4**	**87.3**	**99.4**	**94.5**	**91.3**	96.9	**91.5**	**85.2**	**99.7**	**94.7**	**96.9**	**78.6**	89.1
normal	65.8	53.1	17.1	97.1	81.3	18.6	73.2	10.8	63.1	59.5	74.2	47.8	37.7	70.8
outlier	69.0	55.1	79.1	99.3	93.7	84.9	**97.6**	85.2	84.9	99.6	93.5	94.1	78.5	**89.7**

Table 1: Image classification via linear probing on normal and outlier patch tokens. We also report the accuracy of classifiers learnt on the class token. We see that outlier tokens have a much higher accuracy than regular ones, suggesting they are effectively storing global image information.

Thank You