ReWind: Understanding Long Videos with Instructed Learnable Memory

Anxhelo Diko[1*†] Tinghuai Wang[2*] Wassim Swaileh[2] Shiyan Sun[2] Ioannis Patras[2]
[1]La Sapienza University of Roma [2]Huawei Helsinki Research Center
diko@di.uniroma1.it {tinghuaiwang,shiyansun,wassim.swaileh,ioannis.patras}@huawei.com

setting

VQA：

dataset：MovieChat-1K dataset（1000 for 9.13min）

1. global ： requires processing the entire video and answering questions about its content

2. breakpoint ： processing the video up to a specific timestamp and answering questions about the event at that point

metric：acc；score

Temporal Grounding：

dataset：Charades-STA

metric：mIoU；R@XX

motivation

Challenges：

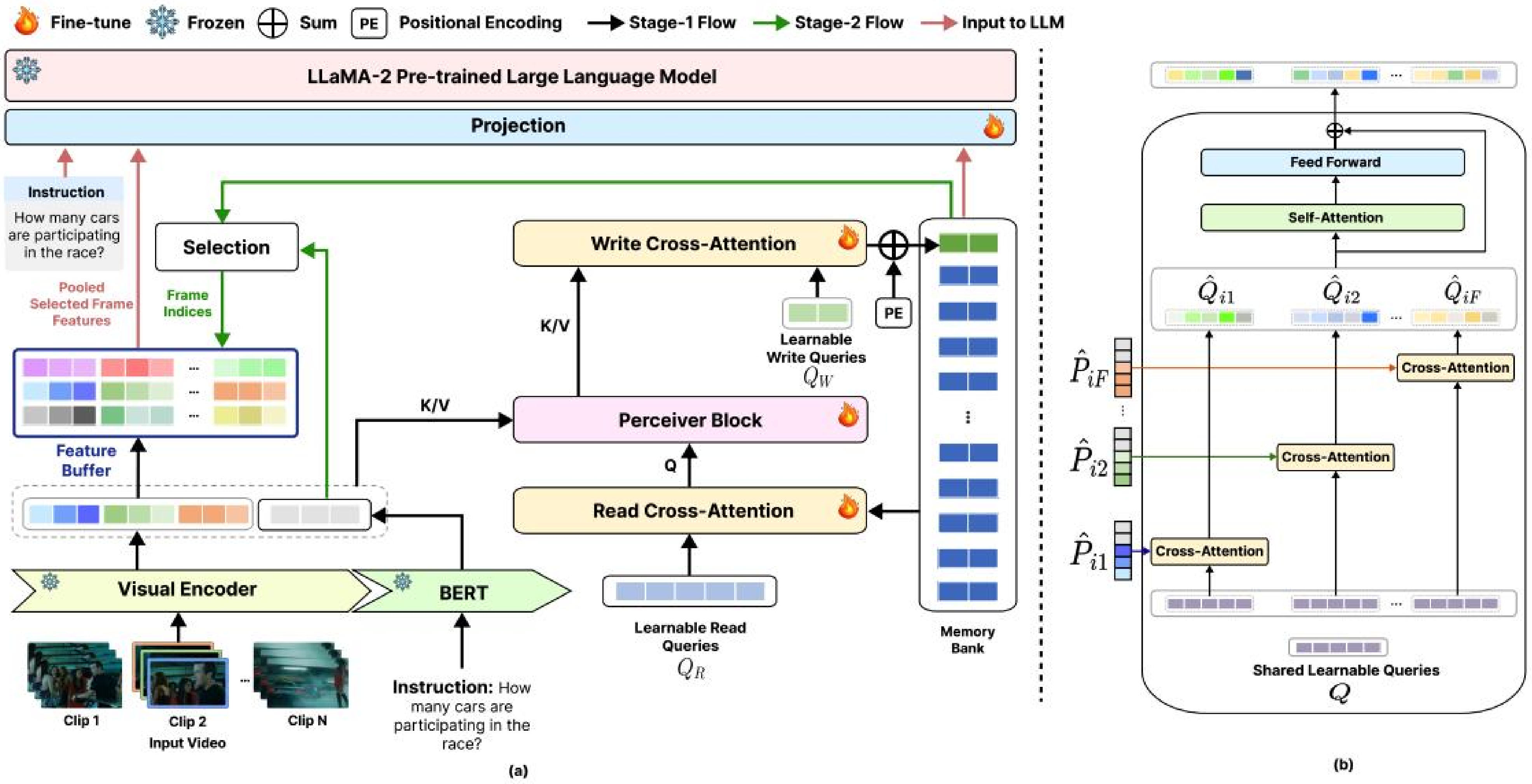
1. self-attention mechanisms require substantial memory that scales quadratically with the number of tokens, making long video processing computationally intensive

→ use Q-Former structure to compress spatial information

2. struggle to effectively model temporal dependencies over extended sequences

→ use memory-informed method to preserve temporal fidelity

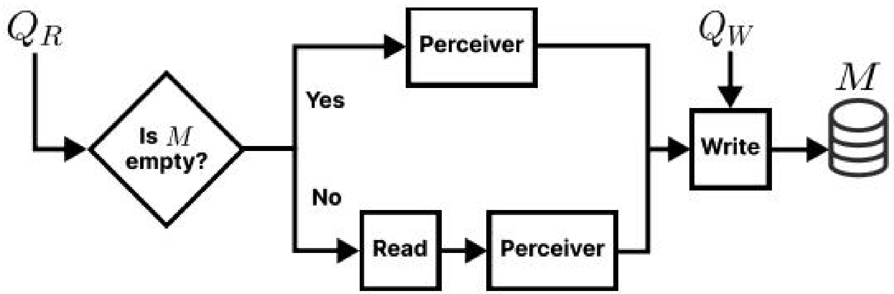
overview

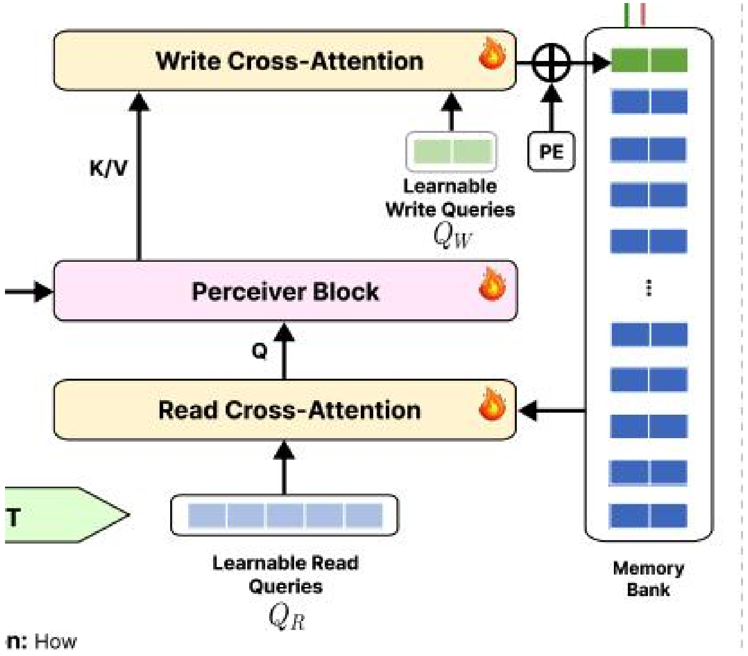


Fine-tune
Frozen
Sum
PE
Positional Encoding
Stage-1 Flow
Stage-2 Flow
Input to LLM
LLaMA-2 Pre-trained Large Language Model
Projection
Instruction
How many cars are participating in the race?
Selection
Write Cross-Attention
Pooled Selected Frame Features
Frame Indices
K/V
Learnable Write Queries
Q_W
PE
Feature Buffer
K/V
Perceiver Block
Q
Read Cross-Attention
Visual Encoder
BERT
Learnable Read Queries
Q_R
Memory Bank
Clip 1
Clip 2
Clip N
Input Video
Instruction: How many cars are participating in the race?
(a)
Feed Forward
Self-Attention
$\hat{Q}_{i1}$
$\hat{Q}_{i2}$
$\hat{Q}_{iF}$
$\hat{P}_{iF}$
Cross-Attention
$\hat{P}_{i2}$
Cross-Attention
$\hat{P}_{i1}$
Cross-Attention
Shared Learnable Queries
Q
(b)

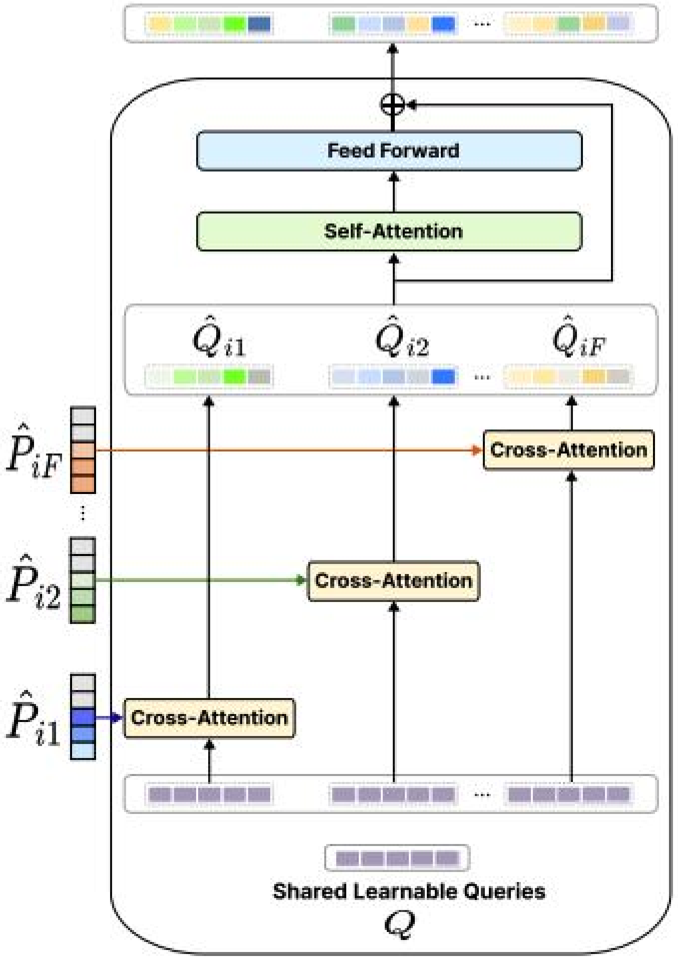
method

Instructed Memory Architecture



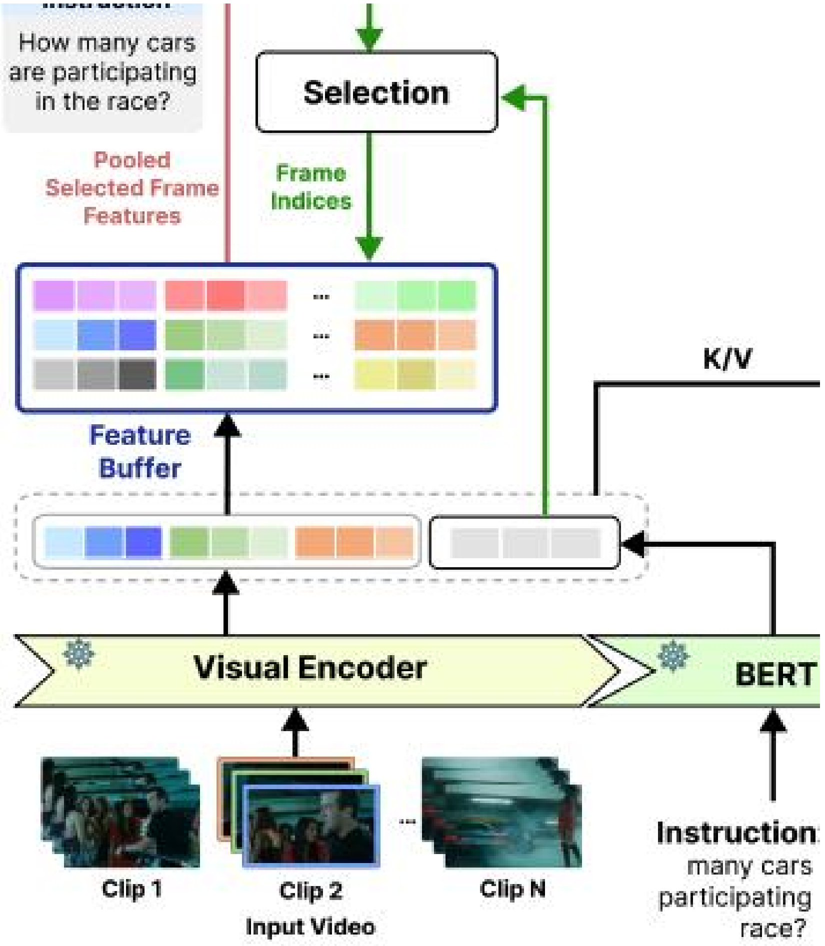


Perceiver Block



method

Dynamic Frame Selection



1. Instruction based selection
compute the attention matrix between
I and contents of M select top L frames set Z_I
2. Clustering

$$\sigma_l = exp(-\frac{1}{K} \sum_{z_k \in KNN(z_l, Z)} ||z_k - z_l||^2)$$

$$\rho_l = \begin{cases} \min\limits_{j:\sigma_j > \sigma_l} ||z_j - z_l||^2 & \textit{if} \, \exists j \;\; s.t. \; \sigma_j > \sigma_l, \\ \max\limits_{j} ||z_j - z_l||^2 & otherwise. \end{cases}$$

select top K (σl × ρl) frames features Z

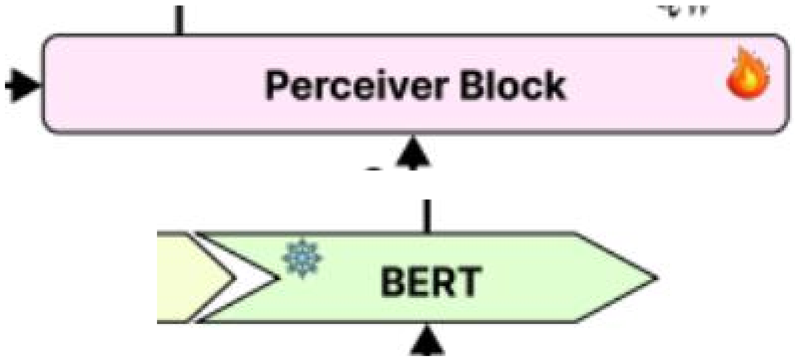
final input： <M, spec, Z>

method

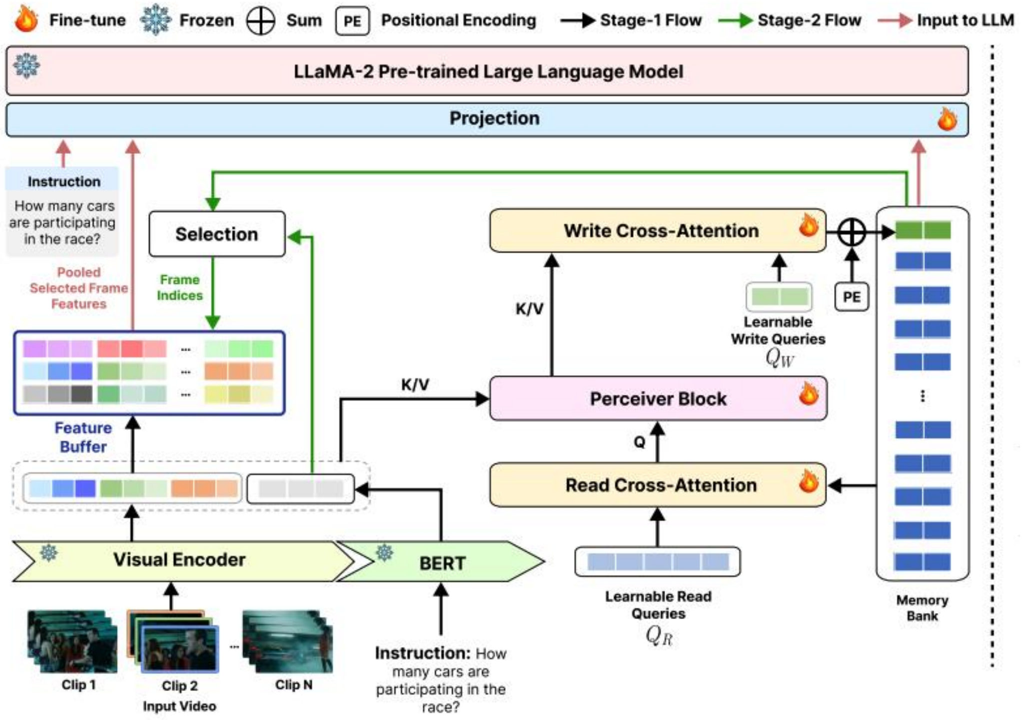
Training

Multimodal Pretraining Stage: contrastive learn

$$\mathcal{L}_{SigLIP} = -\frac{1}{||\mathcal{B}||}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\underbrace{\log\left(\frac{1}{1+\exp(z_{ij}(-t\mathbf{x}_i\cdot\mathbf{y}_j+b))}\right)}_{\mathcal{L}_{ij}}$$



Instruction Tuning Stage:



(a)

experiment

long video QA

Model	Num Frames	Num Tokens	Global VQA		Breakpoint VQA		Global Generation					Breakpoint Generation				
			Accuracy	Score	Accuracy	Score	CI	DO	CU	TU	CO	CI	DO	CU	TU	CO
Video LLaMA [27]	32	32	51.4	3.10	38.2	2.31	3.30	2.53	3.28	2.77	3.42	2.42	2.85	2.87	2.00	2.87
Video-ChatGPT [18]	100	356	44.2	2.71	49.8	2.71	2.48	2.78	3.03	2.48	2.99	3.11	3.32	3.29	2.62	3.29
Video Chat [14]	32	3072	61.0	3.34	48.3	2.43	3.26	3.20	3.38	2.97	3.47	2.96	3.09	3.24	2.46	3.22
MovieChat [21]	2048	8192	67.8	3.81	50.4	2.96	3.32	3.28	3.44	3.06	3.48	3.07	3.24	3.31	2.70	3.45
ReWind (Ours)	548*	1184*	**80.6**	**4.46**	**57.2**	**3.4**	**4.18**	**4.00**	**4.24**	**4.02**	**3.54**	**3.41**	**3.37**	**3.64**	**2.97**	**3.61**

short video QA

Model	LLM	Backbone	CI	DO	CU	TU	CO	AVG
Video Chat [14]	Vicuna-7B	ViT-G	2.23	2.50	2.53	1.94	2.24	2.29
Video LLaMA [27]	Vicuna-7B	ViT-G	1.96	2.18	2.16	1.82	1.79	1.98
Video-ChatGPT [18]	Vicuna-7B	ViT-L	2.40	2.52	2.62	1.98	2.37	2.38
LLaMA Adapter [28]	LLaMA-7B	ViT-L	2.03	2.32	2.30	1.98	2.15	2.16
Chat-UniVi [12]	Vicuna1.5-7B	ViT-L	2.89	2.91	3.46	2.39	**2.81**	2.89
VTimeLLM [11]	Vicuna1.5-7B	ViT-L	2.78	**3.10**	3.40	2.49	2.47	2.85
MovieChat [21]	LLaMA2-7B	ViT-G	2.76	2.93	3.01	2.24	2.42	2.67
LLaMA-VID [15]	Vicuna-7B	ViT-G	**2.96**	3.00	**3.53**	2.46	2.51	2.89
ReWind (Ours)	LLaMA2-7B	ViT-G	2.91	2.85	3.42	**2.71**	2.68	**2.91**

Temporal grounding

Model	Charades-STA			
	R@0.3	R@0.5	R@0.7	mIoU
Video Chat [14]	9.0	3.3	1.3	6.5
Video LLaMA [27]	10.4	3.8	0.9	7.1
Video-ChatGPT [18]	20.0	7.7	1.7	13.7
GroundingGPT [16]	-	29.6	11.9	-
TimeChat [20]	-	32.2	13.4	-
VTimeLLM [11]	51.0	27.5	11.4	31.2
ReWind (Ours)	**59.0**	**41.6**	**20.53**	**39.3**

AdaCM^2: On Understanding Extremely Long-Term Video with Adaptive Cross-Modality Memory Reduction

Yuanbin Man[1], Ying Huang[1], Chengming Zhang[2], Bingzhe Li[3], Wei Niu[4], Miao Yin[1†]
[1]Department of CSE, University of Texas at Arlington, [2]Department of CS, University of Houston,
[3]Department of CS, University of Texas at Dallas, [4]School of Computing, University of Georgia
{yuanbin.man, ying.huang}@uta.edu, czhang48@uh.edu,
bingzhe.li@utdallas.edu, wniu@uga.edu, miao.yin@uta.edu

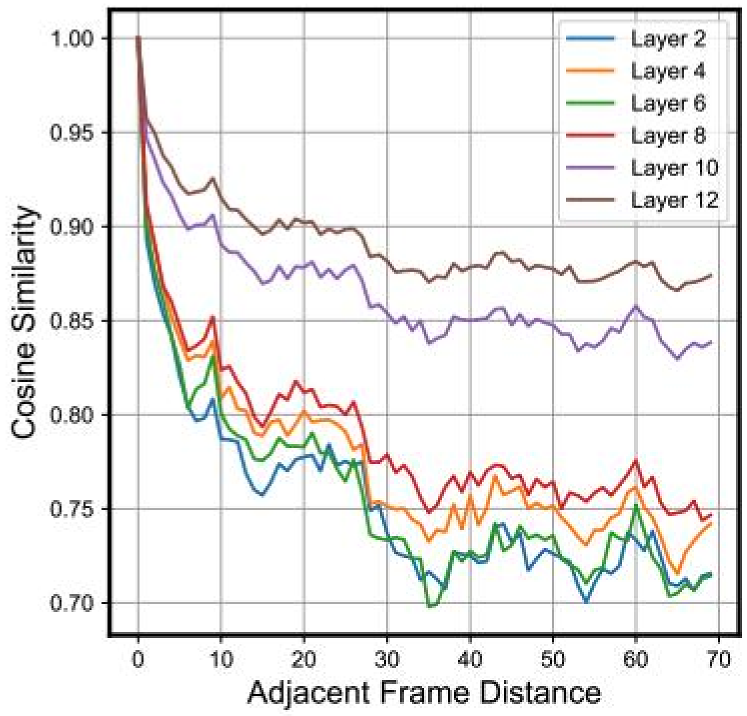
setting

Long-Trem Video Understanding
dataset：LVU
content understanding and metadata prediction tasks

Video Captioning：
dataset：MSRVTT；MSVD；YouCook2
metric：METEOR“语义正确性” / CIDEr“表述符合共识”

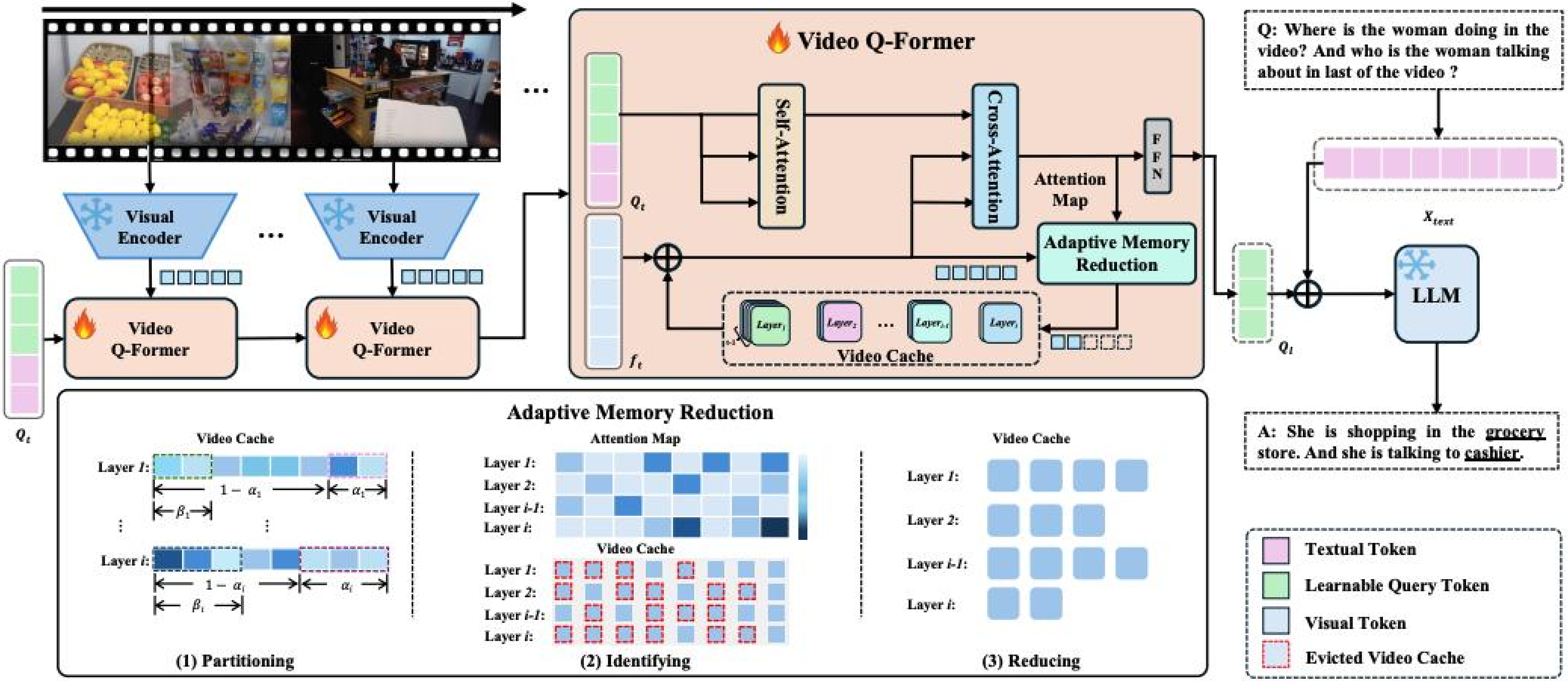
motivation

Observation
1. Only a subset of visual tokens exhibits high correlations to the text query within a frame
2. Correlation varies across different layers



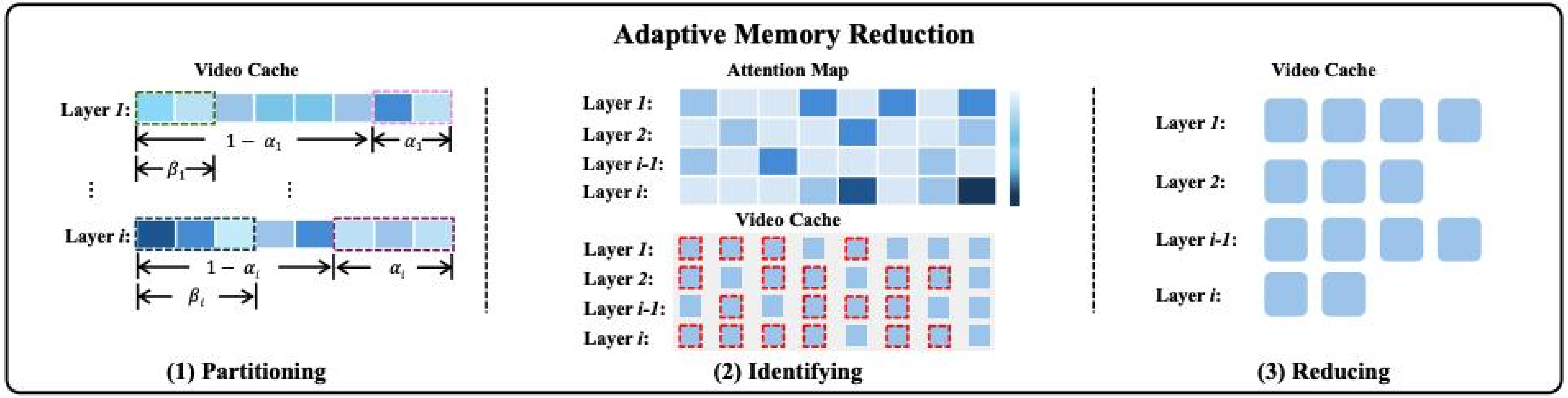
Adaptive Memory Reduction with Cross-Modality Attention

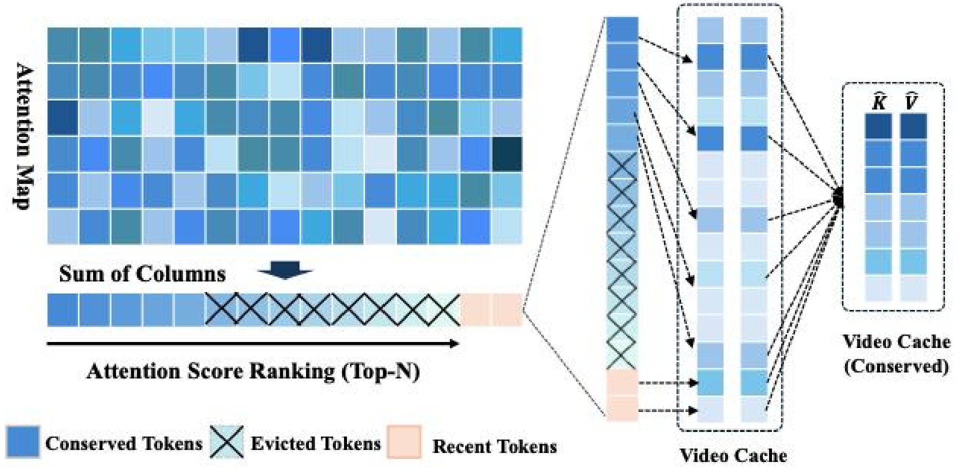
overview


Video Q-Former
Self-Attention
Cross-Attention
F F N
Attention Map
Adaptive Memory Reduction
Video Cache
Visual Encoder
Video Q-Former
Q: Where is the woman doing in the video? And who is the woman talking about in last of the video ?
LLM
A: She is shopping in the grocery store. And she is talking to cashier.
Adaptive Memory Reduction
Video Cache
Attention Map
(1) Partitioning
(2) Identifying
(3) Reducing
Textual Token
Learnable Query Token
Visual Token
Evicted Video Cache

method

Adaptive Memory Reduction with CrossModality Attention





Algorithm 1 Layer-wise Adaptive Video Reduction.

Input: Video frames $\{\boldsymbol{f}_t\}_{t=1}^{T}$, hyper-parameters α, β for each layer;
Output: Reduced video KV cache $\boldsymbol{K}_t, \boldsymbol{V}_t$.
for $t = 1$ to T **do**
  $\boldsymbol{K}_t \leftarrow [\boldsymbol{K}_{t-1}, \boldsymbol{f}_t\boldsymbol{W}_K]; \boldsymbol{V}_t \leftarrow [\boldsymbol{V}_{t-1}, \boldsymbol{f}_t\boldsymbol{W}_V]$;
  Partitioning: $\boldsymbol{K}_t \rightarrow [\hat{\boldsymbol{K}}_t, \tilde{\boldsymbol{K}}_t]; \boldsymbol{V}_t \rightarrow [\hat{\boldsymbol{V}}_t, \tilde{\boldsymbol{V}}_t]$;
  Identifying: Determine $\boldsymbol{S}_t^c$ using Eq.5;
  Reducing: Obtain $\overline{\boldsymbol{K}}_t$ using Eq.7;
  $\boldsymbol{K}_t \leftarrow [\overline{\boldsymbol{K}}_t, \tilde{\boldsymbol{K}}_t]; \boldsymbol{V}_t \leftarrow [\overline{\boldsymbol{V}}_t, \tilde{\boldsymbol{V}}_t]$;
end for

experiment

Table 1. Comparison with state-of-the-art methods on the LVU dataset. The underlined number means the second best.

Model	Content			Metadata				Avg
	Relation	Speak	Scene	Director	Genre	Writer	Year	
Obj_T4mer [45]	54.8	33.2	52.9	47.7	52.7	36.3	37.8	45.1
Performer [9]	50.0	38.8	60.5	58.9	49.5	48.2	41.3	49.6
Orthoformer [32]	50.0	38.3	66.3	55.1	55.8	47.0	43.4	50.8
VideoBERT [36]	52.8	37.9	54.9	47.3	51.9	38.5	36.1	45.6
LST [19]	52.5	37.3	62.8	56.1	52.7	42.3	39.2	49.0
VIS4mer [19]	57.1	40.8	67.4	62.6	54.7	48.8	44.8	53.7
S5 [43]	**67.1**	42.1	73.5	67.3	65.4	51.3	48.0	59.2
MA-LMM [16]	58.2	**44.8**	80.3	74.6	61.0	70.4	51.9	63.0
Ours	63.1	40.2	**86.2**	**75.4**	**68.0**	**77.0**	**62.5**	**67.5**

Model	Breakfast	COIN
TSN [44]	-	73.4
VideoGraph [18]	69.5	-
Timeception [17]	71.3	-
GHRM [13]	75.5	-
D-Sprv. [25]	89.9	90.0
ViS4mer [19]	88.2	88.4
S5 [43]	90.7	90.8
MA-LMM [16]	93.0	93.2
Ours	**94.4**	**93.3**

Model	MSRVTT		MSVD		YouCook2	
	M	C	M	C	M	C
UniVL [28]	28.2	49.9	29.3	52.8	-	127.0
SwinBERT [24]	29.9	53.8	41.3	120.6	15.6	109.0
GIT [41]	32.9	73.9	51.1	180.2	17.3	129.8
mPLUG-2 [47]	**34.9**	**80.3**	48.4	165.8	-	-
VideoCoca [49]	-	73.2	-	-	-	128.0
VideoLLaMA [52]	32.9	71.6	49.8	175.3	16.5	123.7
MA-LMM [16]	33.4	74.6	51.0	179.1	**17.6**	**131.2**
Ours	33.0	73.1	**51.4**	**189.4**	**17.6**	125.6