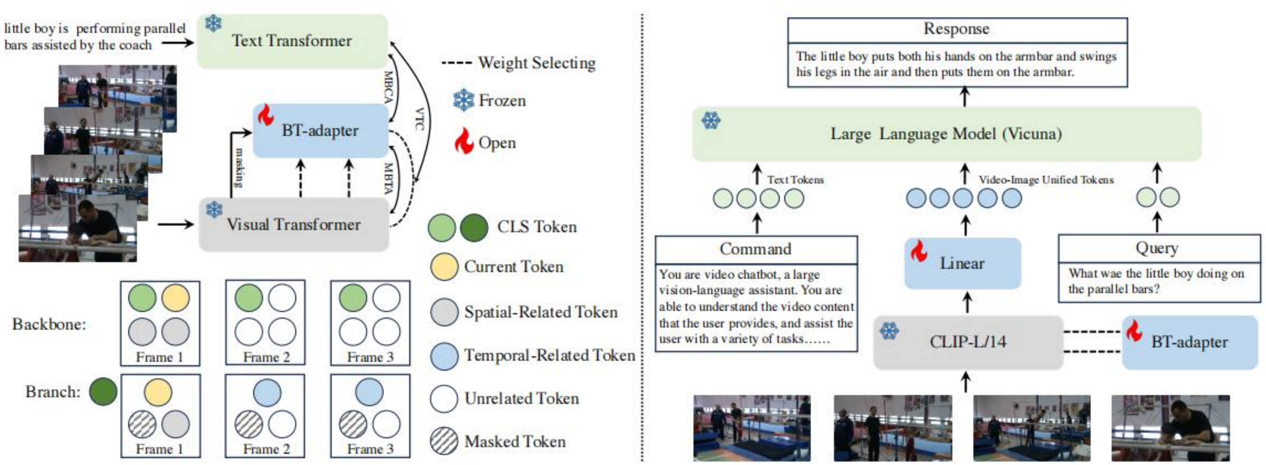
两篇视频理解LLM论文分享

Bowei Pu
2024.06.18

ONE FOR ALL: VIDEO CONVERSATION IS FEASIBLE WITHOUT VIDEO INSTRUCTION TUNING

Ruyang Liu* [1] **Chen Li** [2] **YiXiao Ge** [2] **Ying Shan** [2] **Thomas H. Li** [1] **Ge Li** ✉ [1]
[1] School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University
[2] Applied Research Center (ARC), Tencent PCG
{ruyang@stu,geli@ece,thomas@}.pku.edu.cn
{palchenli,yixiaoge,yingsshan}@tencent.com



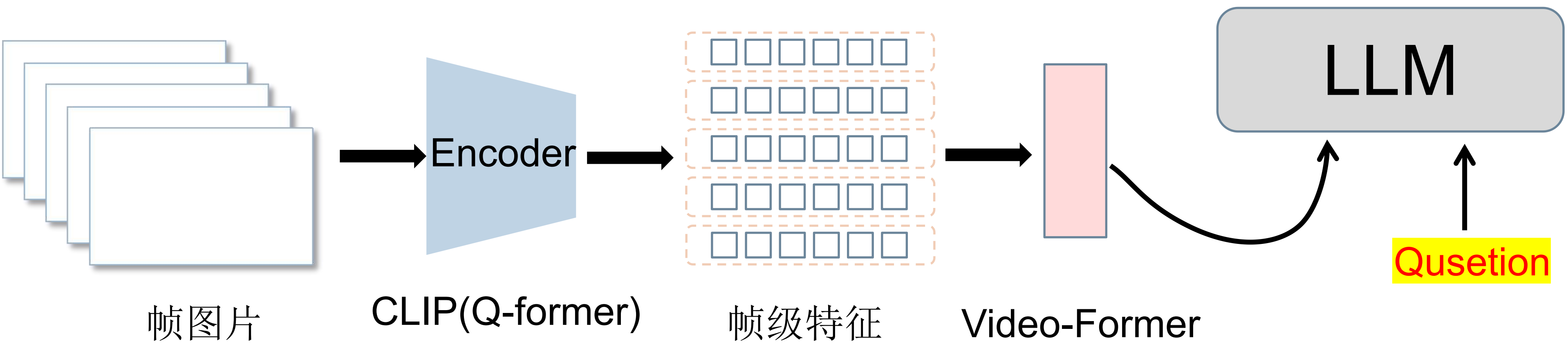
8 V100(32G) GPUs
in a mere 3 hours

能多媒体内容计算实验室
lligent Multimedia Content Computing Lab

One For All: Video Conversation is Feasible Without Video Instruction Tuning

- 研究背景
- 相关工作
- 主要方法
- 实验结果

One For All: Video Conversation is Feasible Without Video Instruction Tuning



Video-LLaMA为代表的视频理解LLM
Video-LLaVA 删去各种Former,仅保留线性层
总体结构： 解码器->连接器->LLM （posterior structure）

存在问题： 1.多帧输入对于GPU的需求更大
2.微调编码器消耗也很多

目标解决方案：用增量微调的方式微调CLIP编码器，实现图片转化到视频领域（BT-Adapter）

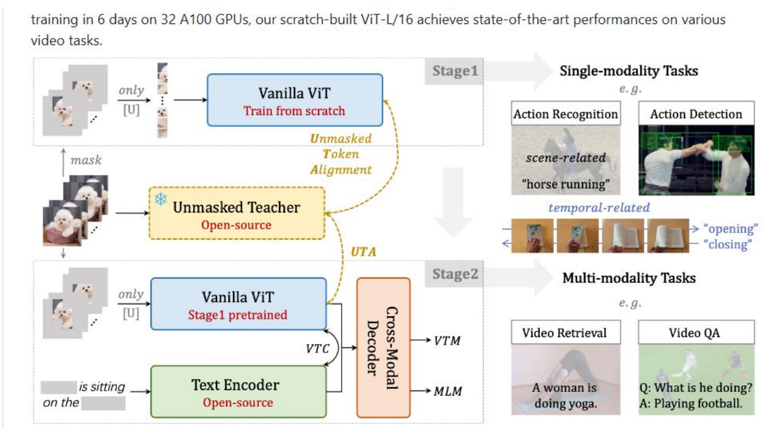
One For All: Video Conversation is Feasible Without Video Instruction Tuning

- 研究背景
- 相关工作
- 主要方法
- 实验结果

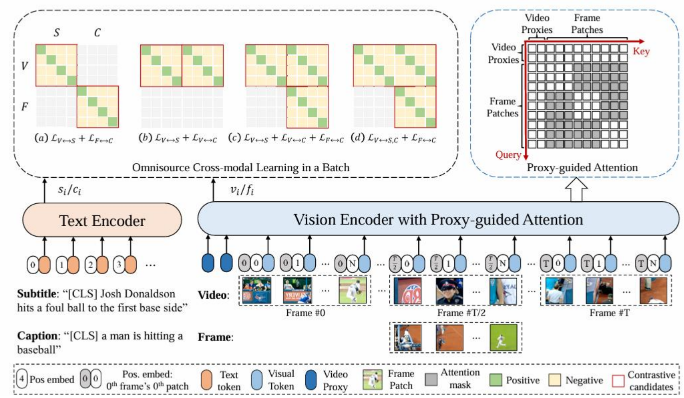
One For All: Video Conversation is Feasible Without Video Instruction Tuning

joint-spatial-temporal modeling

[1]



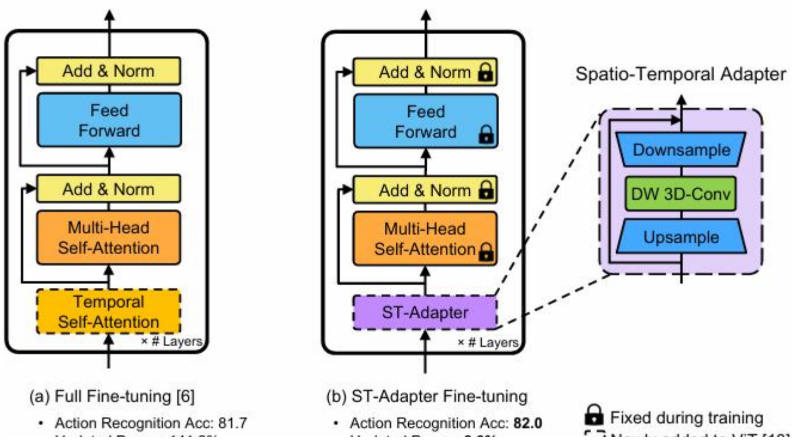
[2]



[1]Unmasked Teacher: Towards Training-Efficient Video Foundation Models
[2]CLIP-ViP: Adapting Pre-trained Image-Text Model to Video-Language Representation Alignment

One For All: Video Conversation is Feasible Without Video Instruction Tuning

separated Spatial-Temporal modeling



ST-Adapter: Parameter-Efficient Image-to-Video Transfer Learning

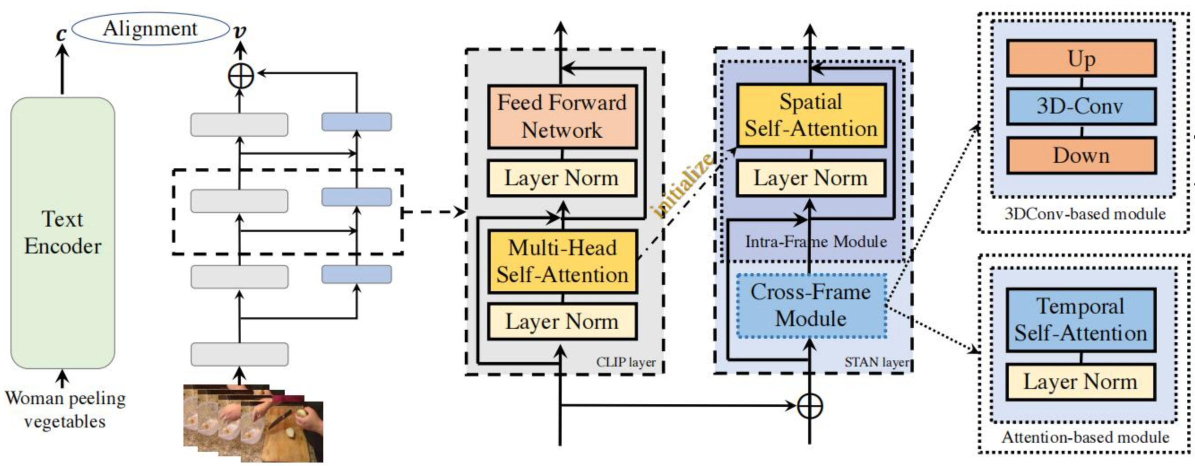


Figure 2. The overview of our proposed STAN architecture, including the global overview of our backbone (left), details of the internal structure of our spatial-temporal module (middle), and implementations of the cross-frame module (right).

Revisiting Temporal Modeling for CLIP-based Image-to-Video Knowledge Transferring

(与本工作最相近的工作）

One For All: Video Conversation is Feasible Without Video Instruction Tuning

- 研究背景
- 相关工作
- 主要方法
- 实验结果

One For All: Video Conversation is Feasible Without Video Instruction Tuning

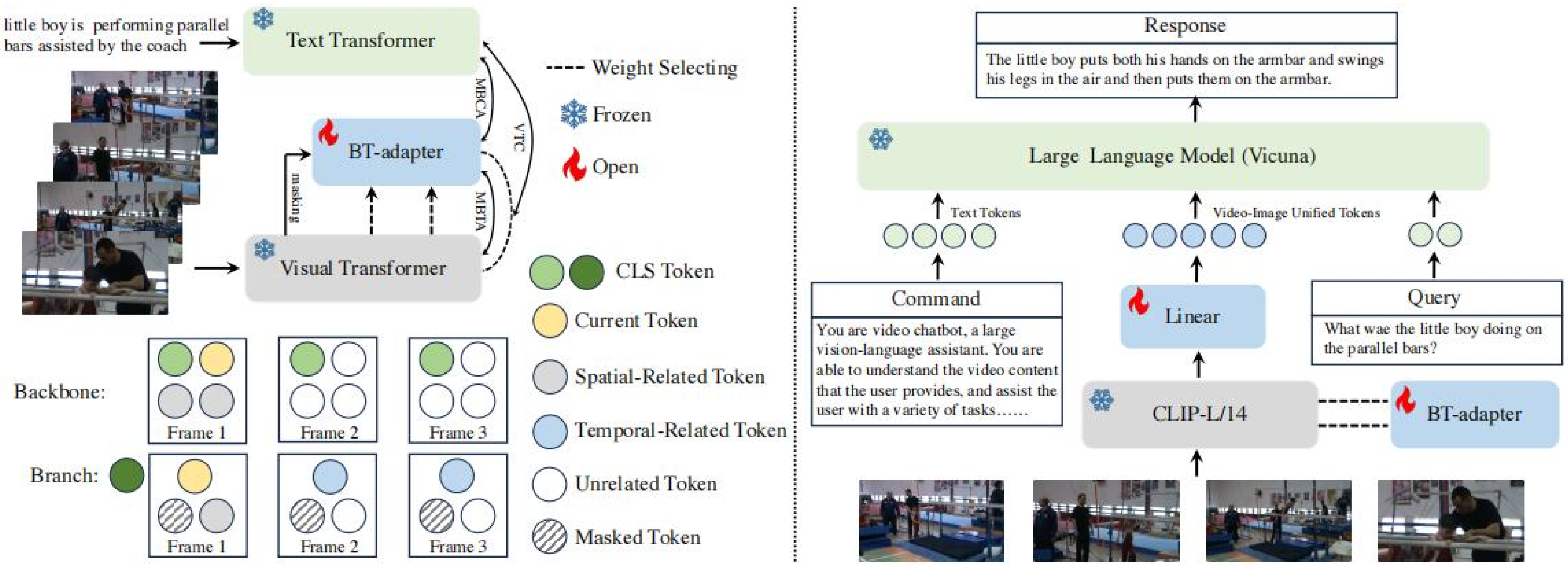


Figure 2: The overview of our model. The left side shows the model architecture and the data flow during pretraining. The right side depicts the pipeline of video conversation.

全局图
1.CLIP适配视频数据
2.step2加入LLM进行对话

One For All: Video Conversation is Feasible Without Video Instruction Tuning

本文的Backbone-Branch Interaction
Vi,j 表示第i帧的第j个token，右上角代表Block编号
若有N个分支层，在第K层添加分支，第0个Block的分支的新的CLS token利用原CLS token初始化如下：

$$\hat{v}_{i,j}^{(0)} = v_{i,j}^{(k)} + \mathrm{P}_i^t + \mathrm{P}_j^s,$$

其他层是:

$$\hat{v}_{i,j}^{(l)} = \mathrm{Sigmoid}(\mathrm{W_b}) \cdot \hat{v}_{i,j}^{(l-1)} + (1 - \mathrm{Sigmoid}(\mathrm{W_b})) \cdot v_{i,j}^{(k+l-1)},$$

值得注意的是v和ṽ相差k，换而言之v^0 和$\tilde{v}^k$是同一层的新[CLS]和原[CLS]

最后一层输出如下，添加一个线性层

$$v = \mathrm{W_{v\_proj}}(\mathrm{LN}(\mathrm{Sigmoid}(\mathrm{W_b}) \cdot \hat{v}_{0,0}^{(-1)} + (1 - \mathrm{Sigmoid}(\mathrm{W_b})) \cdot \frac{1}{T}\sum_{i=1}^{T} v_{i,0}^{-1})),$$

One For All: Video Conversation is Feasible Without Video Instruction Tuning

针对视频对话的时间建模方法

Model	parameter-efficient	multimodal-friendly	temporal-sensitive	Correctness of Information	Temporal Understanding
Baseline*				2.38	1.93
ST Pooling*	✓	✓		2.40	1.98
Joint-ST			✓	1.92	2.11
Separate-ST			✓	2.10	2.13
Separate-ST*	✓			2.29	2.01
BT-Adapter*	✓	✓	✓	**2.55**	**2.26**

MASK方法预训练

在分支网络中遮掩70%以上的token，减少一半以上计算资源消耗

1.对比损失

$$\mathcal{L}_{nce}(x,y) = -\frac{1}{B}\sum_{m=1}^{B}\log\frac{\exp(\tau x_m \cdot y_n)}{\sum_{n=1}^{B}\exp(\tau x_m \cdot y_n)}, \quad \mathcal{L}_{VTC} = \mathcal{L}_{nce}(v,t) + \mathcal{L}_{nce}(t,v),$$

2.NCE 损失 匹配对应的分支CLS和文本CLS

$$\hat{v} = \frac{1}{(1-\rho)NT}\sum_{i=1}^{T}\sum_{j=1}^{N}\mathrm{W}_{\mathrm{v\_proj}} \cdot \hat{v}_{i,j}^{-1}, \quad \mathcal{L}_{MBCA} = \mathcal{L}_{nce}(\hat{v},t) + \mathcal{L}_{nce}(t,\hat{v}), \quad (i,j) \notin M. \quad ($$

One For All: Video Conversation is Feasible Without Video Instruction Tuning

- 研究背景
- 相关工作
- 主要方法
- 实验结果

One For All: Video Conversation is Feasible Without Video Instruction Tuning

Table 2: The zero-shot results of text-to-video retrieval on MSR-VTT, DiDemo, LSMDC, and ActivityNet. Source denotes the scale of pretraining data.

Method	Source	MSR-VTT			DiDeMo			LSMDC			ActivityNet		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Non-CLIP models													
Frozen (Bain et al., 2021)	5M	24.7	46.9	57.2	21.1	46.0	56.2	-	-	-	-	-	-
Clover (Huang et al., 2023)	5M	26.4	49.5	60.0	29.5	55.2	66.3	17.4	29.2	38.2	-	-	-
OmniVL (Wang et al., 2022a)	14M	34.6	58.4	66.6	33.3	58.7	68.5	-	-	-	-	-	-
HiTeA (Ye et al., 2022)	5M	29.9	54.2	62.9	36.1	60.1	70.3	15.5	31.1	39.8	-	-	-
Singularity (Lei et al., 2022)	17M	34.0	56.7	66.7	**37.1**	61.7	69.9	-	-	-	30.6	55.6	66.9
VideoCoCa (Yan et al., 2022)	100M	34.3	57.8	67.0	-	-	-	-	-	-	34.5	63.2	76.6
CLIP-L/14													
CLIP (Radford et al., 2021)	-	35.4	58.8	68.1	30.3	54.9	65.4	17.0	31.8	40.3	28.8	57.6	71.8
ImageBind (Girdhar et al., 2023)	-	36.8	61.8	70.0	-	-	-	-	-	-	-	-	-
InternVideo (Wang et al., 2022b)	12.8M	40.7	-	-	31.5	-	-	17.6	-	-	30.7	-	-
TVTSv2 (Zeng et al., 2023)	8.5M	38.2	62.4	73.2	34.6	**61.9**	71.5	17.3	32.5	41.4	-	-	-
UMT-L (Li et al., 2023c)	5M	33.3	58.1	66.7	34.0	60.4	68.7	**20.0**	**37.2**	43.7	31.9	60.2	72.0
BT-Adapter	2M	**40.9**	**64.7**	**73.5**	35.6	**61.9**	**72.6**	19.5	35.9	**45.0**	**37.0**	**66.7**	**78.9**

Table 3: The results of video conversation on video-based generative performance benchmarking. FT and ZS mean with and without video instruction tuning respectively.

Evaluation Aspect	VideoLLaMA	LLaMA-Adapter	VideoChat	VideoChatGPT	Ours (ZS)	Ours (FT)
Temporal Understanding	1.82	1.98	1.94	1.98	2.13	**2.34**
Correctness of Information	1.96	2.03	2.23	2.40	2.16	**2.68**
Detail Orientation	2.18	2.32	2.50	2.52	2.46	**2.69**
Contextual Understanding	2.16	2.30	2.53	2.62	2.89	**3.27**
Consistency	1.79	2.15	2.24	2.37	2.20	**2.46**
Mean	1.98	2.16	2.29	2.38	2.46	**2.69**

One For All: Video Conversation is Feasible Without Video Instruction Tuning

可以集成到Video LLM上

Table 4: The results of video conversation zero-shot question-answering. FT and ZS mean with and without instruction tuning respectively.

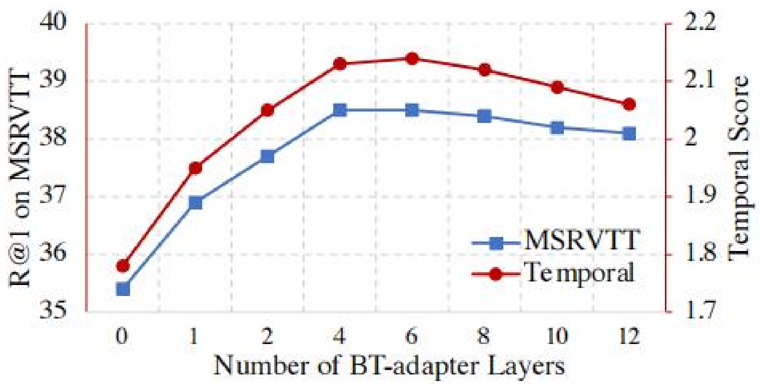
Method	MSVD-QA		MSRVTT-QA		ActivityNet-QA	
	Acc	Score	Acc	Score	Acc	Score
VideoLLaMA	51.6	2.5	29.6	1.8	12.4	1.1
LLaMA-Adapter	54.9	3.1	43.8	2.7	34.2	2.7
VideoChat	56.3	2.8	45.0	2.5	26.5	2.2
VideoChatGPT	64.9	3.3	49.3	2.8	35.2	2.7
Ours (ZS)	67.0	3.6	51.2	2.9	**46.1**	**3.2**
Ours (FT)	**67.5**	**3.7**	**57.0**	**3.2**	45.7	**3.2**

Table 5: The results of zero-shot video conversation on different image-centric dialogue models.

Method	Temporal	Correctness
LLaVA-Vicuna	1.78	2.06
+BT-Adapter	2.13	2.16
MiniGPT4-Vicuna	1.88	2.48
+BT-Adapter	2.56	2.71
MiniGPT4-LLama2	1.56	1.44
+BT-Adapter	2.15	1.81

One For All: Video Conversation is Feasible Without Video Instruction Tuning

消融实验



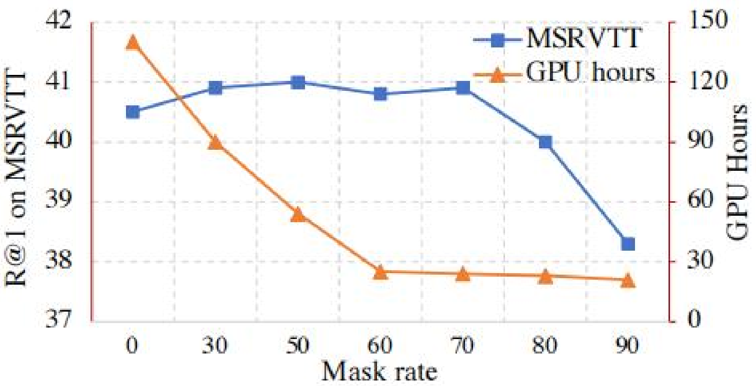


Table 6: Ablation study on the structures of BT-Adapter. We report the results on zero-shot R@1 of MSRVTT and DiDemo retrieval and zero-shot video conversation.

Model	MSRVTT	DiDemo	Temporal
CLIP (baseline)	35.4	30.3	1.78
+4 layer separate-ST	35.7	31.0	1.81
+branch modeling	37.4	32.9	1.97
+backbone-branch interaction	**38.5**	**33.9**	**2.06**

Table 7: Ablation study on the training objectives. We report the zero-shot R@1 of retrieval.

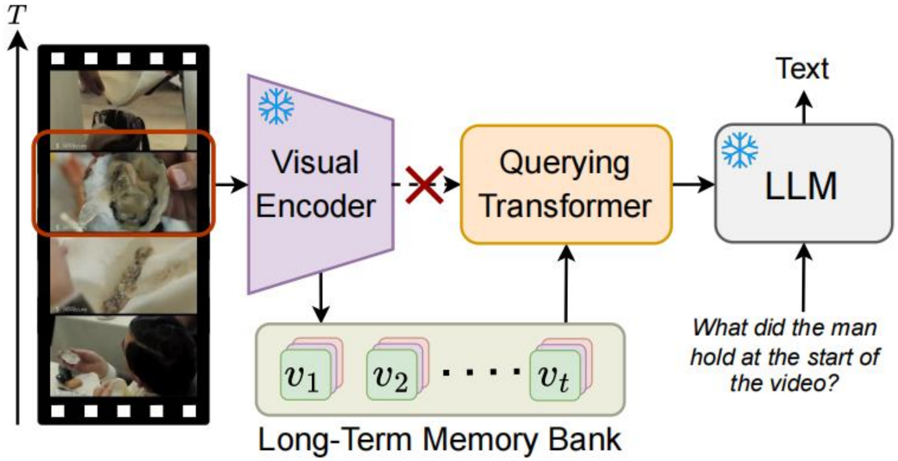
MBTA	MBCA	MSRVTT	DiDemo
		38.5	33.9
✓		39.3	34.3
	✓	40.1	34.9
✓	✓	**40.9**	**35.6**

MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding

Bo He[1,2*] Hengduo Li[2] Young Kyun Jang[2] Menglin Jia[2] Xuefei Cao[2]
Ashish Shah[2] Abhinav Shrivastava[1] Ser-Nam Lim[3]

[1]University of Maryland, College Park [2]Meta [3]University of Central Florida

https://boheumd.github.io/MA-LMM/



MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding

- 研究背景
- ~~相关工作~~
- 主要方法
- 实验结果

MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding

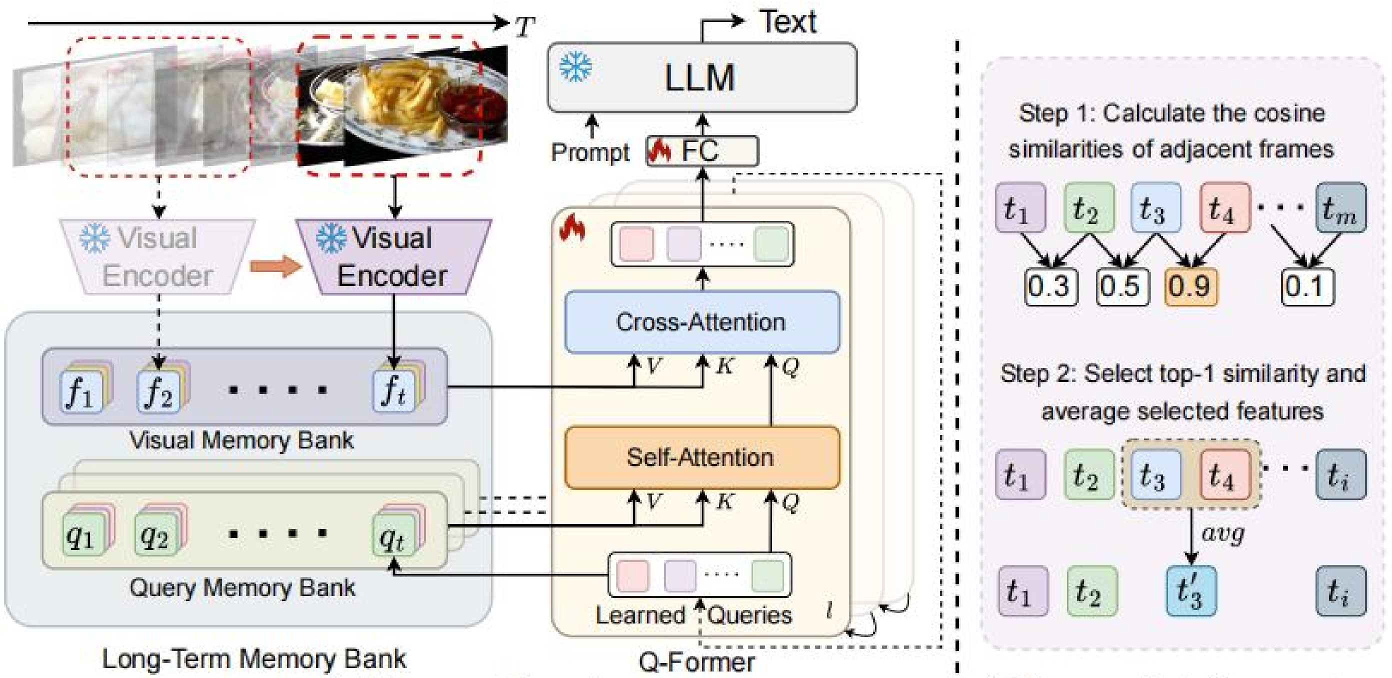
长视频理解的痛点：
1. 信息冗余
2. 资源受限

本文解决方法：
1. 建立内存，参与Q-former的解码，并直接输入到LLM中

MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding

- 研究背景
- ~~相关工作~~
- 主要方法
- 实验结果

MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding

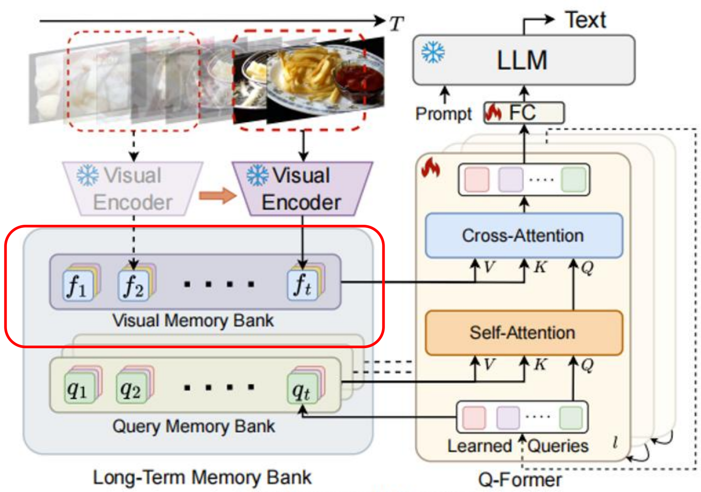


(a) Framework Overview (b) Memory Bank Compression

查询和视觉内存机制实现不同前置信息下不同的关注内容感知

MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding

视觉内存机制



(a) Framework Overview

PE是时间信息的嵌入层，CLIP解码后添加时间信息

$$f_t = v_t + PE(t), f_t \in \mathbb{R}^{P\times C}.$$

拼接全部的内存特征

$$F_t = \text{Concat}[f_1, f_2, .., f_t], l$$

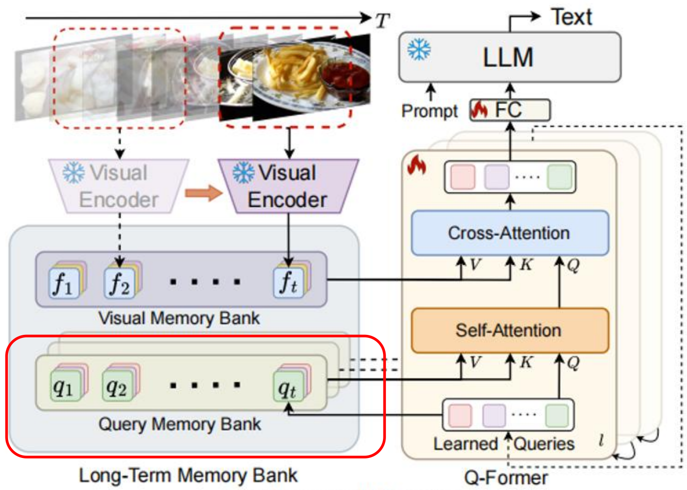
Q-Former层中的第一层self-attn 与长期记忆融合，关注当前帧的不同信息

$$Q = z_t W_Q,\ K = F_t W_K,\ V = F_t W_V.$$

we apply the cross-attention operation as:

$$O = Attn(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{C}}\right)V.$$

MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding



(a) Framework Overview

查询内存机制，类似视觉内存

$$Q = z_t W_Q,\ K = Z_t W_K,\ V = Z_t W_V.$$

内存压缩

相似度最高的两帧融合

$$s_t^i = \cos(f_t^i, f_{t+1}^i), t \in [1, M], i \in [1, P].$$

$$\hat{f}_k^i = (f_k^i + f_{k+1}^i)/2.$$

MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding

- 研究背景
- ~~相关工作~~
- 主要方法
- 实验结果

MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding

Table 3. Comparison with state-of-the-art methods on the video question answering task. Top-1 accuracy is reported.

Model	**MSRVTT**	**MSVD**	**ActivityNet**
JustAsk [74]	41.8	47.5	38.9
FrozenBiLM [75]	47.0	54.8	43.2
SINGULARITY [76]	43.5	–	44.1
VIOLETv2 [77]	44.5	54.7	–
GiT [78]	43.2	56.8	–
mPLUG-2 [79]	48.0	58.1	–
UMT-L [80]	47.1	55.2	47.9
VideoCoCa [81]	46.3	56.9	**56.1**
Video-LLaMA [12]	46.5	58.3	45.5
Ours	**48.5**	**60.6**	49.8

Table 5. Action anticipation results on EpicKitchens-100.

Model	**Accuracy@Top-5**			**Recall@Top-5**		
	Verb	Noun	Act.	Verb	Noun	Act.
Video-LLaMA	73.9	47.5	29.7	**26.3**	27.3	11.7
Ours	**74.5**	**50.7**	**32.7**	25.9	**29.9**	**12.2**

MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding

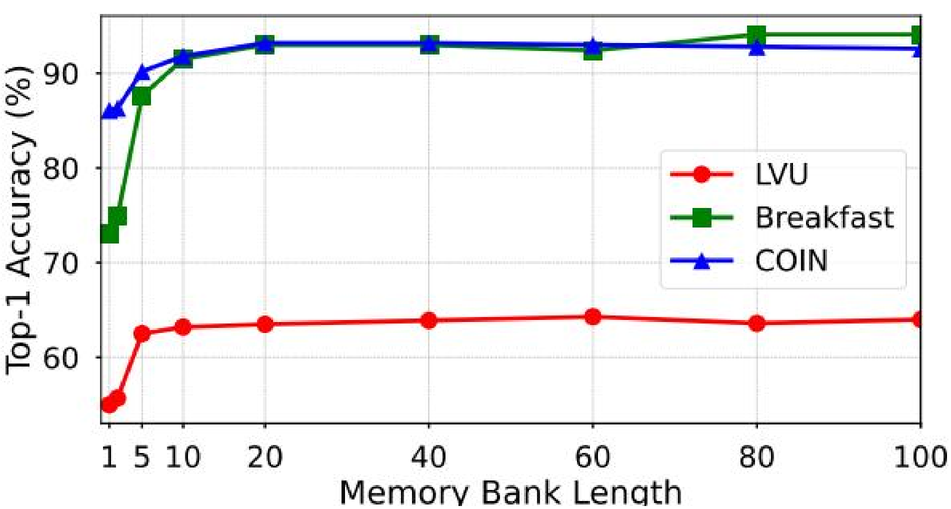
消融实验

Visual	Query	LVU	Breakfast	COIN
✗	✗	48.3	74.6	72.3
✓	✗	61.5	91.8	92.4
✗	✓	58.0	81.4	88.5
✓	✓	**63.0**	**93.0**	**93.2**

视觉内存提升最大

Method	#Frame	#Token	GPU	LVU	Breakfast	COIN
Concat	60	1920	49.2	62.6	90.4	93.0
Avg Pool	100	32	21.2	57.6	80.6	87.6
ToMe	100	200	22.2	61.5	91.3	91.5
FIFO	100	32	19.1	61.3	88.5	90.4
MBC	100	32	19.1	**63.0**	**93.0**	**93.2**

内存的压缩方法差异



内存长度差异

总结：

1.One for all 针对的是短视频，分支方法CLIP转化到视频领域，新CLStoken在不同帧之间交互

2.MA-LMM 转化BLIP2的Q-former实现图片到视频的转化，使用内存机制来实现长视频理解

不足：

One for all ： 1.一个cls token作为每帧的代理进行交互，数量可能过少

2.长视频适配程度待定

MA-LMM： 1.一种posterior structure,效果应该不如分支结构微调CLIP

2.缓存机制粗暴，细粒度信息丢失

3. 难以查询最前的信息