Two Papers about Video LLM

Paper Reading by Yiwei Sun

2024.05.21

智能多媒体内容计算实验室

Intelligent Multimedia Content Computing Lab

- □ VideoLLM范式：Video LLaMA
- □ 论文一：VaQuitA
- □ 论文二：Koala
- □ 总结反思

Video LLaMA

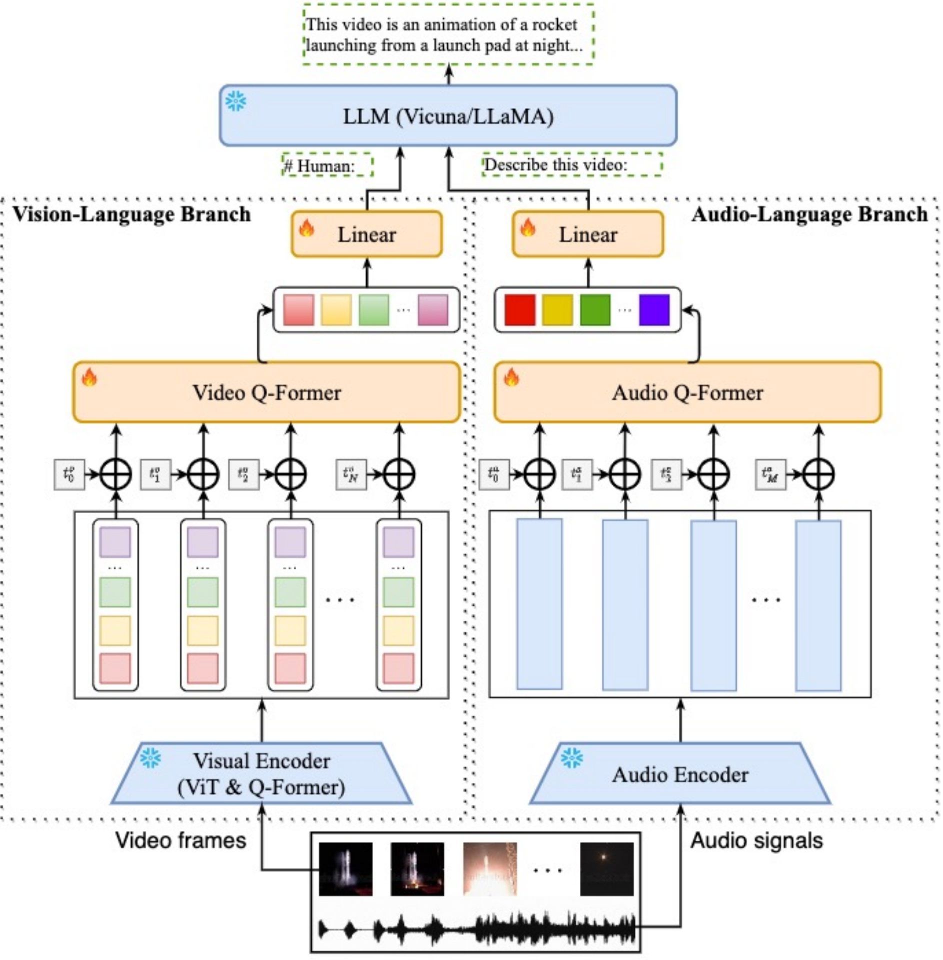


Figure 1: Overall architecture of Video-LLaMA.

视频分支：
1. 视觉编码器（冻结）：ViT + Q-Former；
2. 视频编码器（训练）：Video Q-Former;
3. LLM空间映射（训练）：Linear Layer。

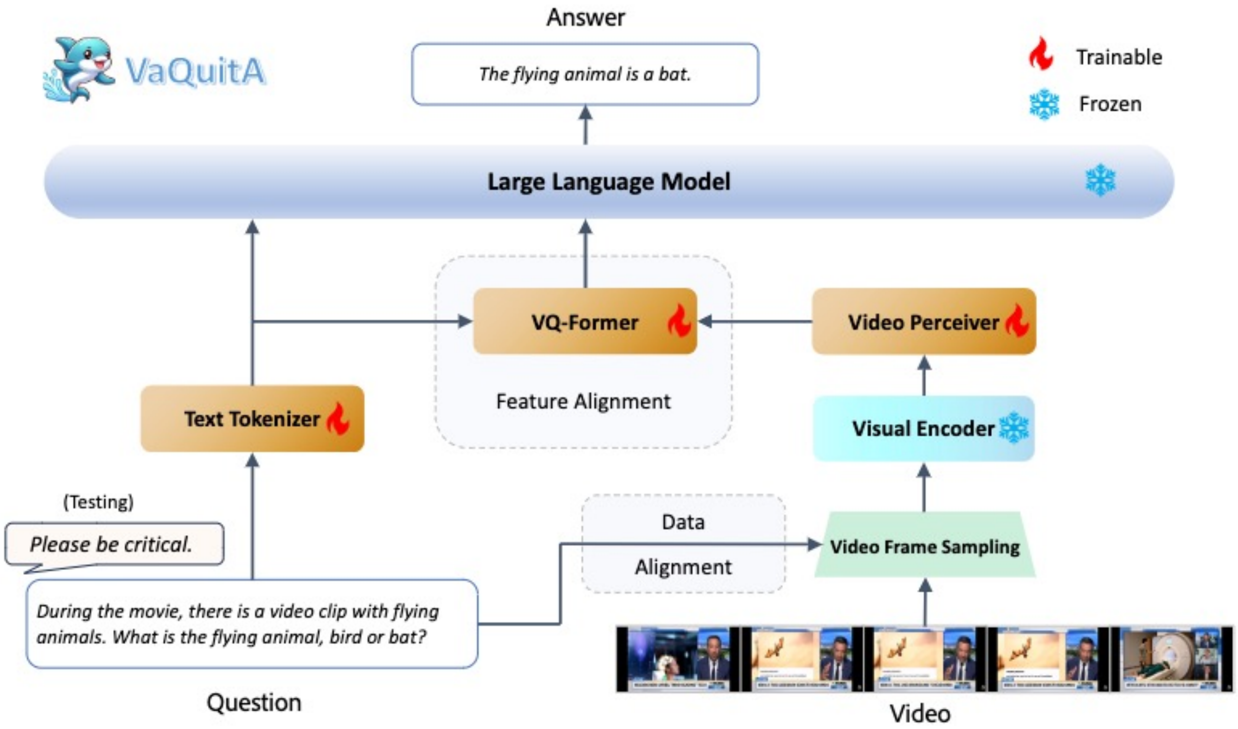
QA：
1. 视觉编码器之争；
2. Video Q-Former的局限性（时序建模）。

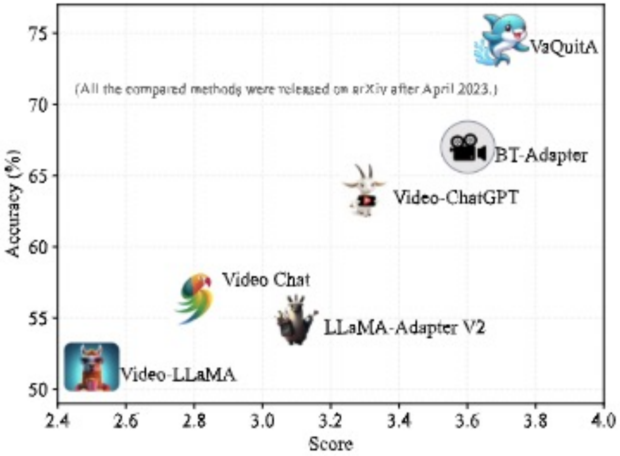
音频分支：
相同的结构。训练数据采用视频文本数据。

- □ VideoLLM范式：Video LLaMA
- □ 论文一：VaQuitA
- □ 论文二：Koala
- □ 总结反思

论文一：VaQuitA

VaQuitA: Enhancing Alignment in LLM-Assisted Video Understanding



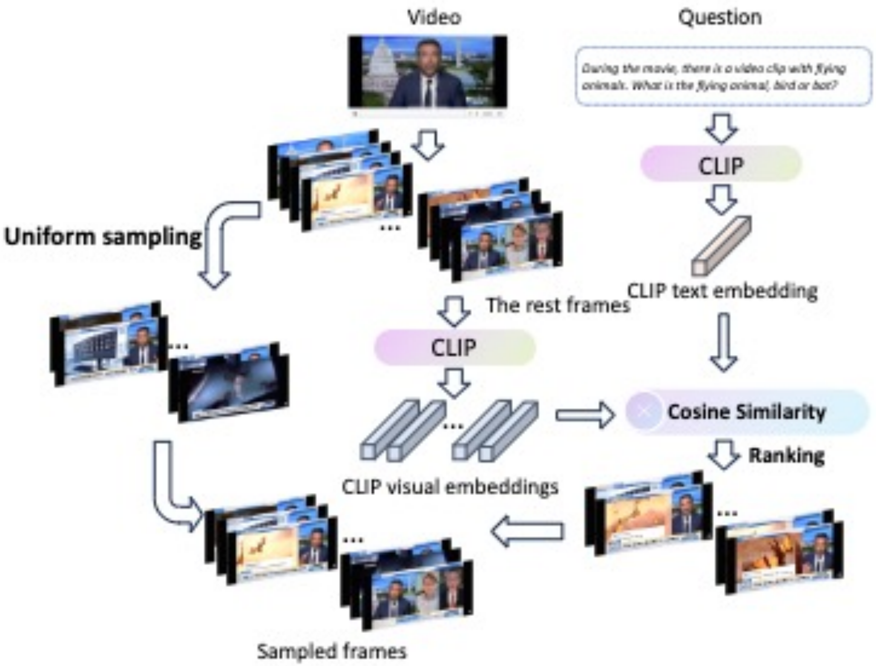


创新点：

1. 基于clip score的采样；
2. 基于文本引导的特征映射；
3. Magic Prompt。

本文专注于利用文本去引导视觉特征的提取

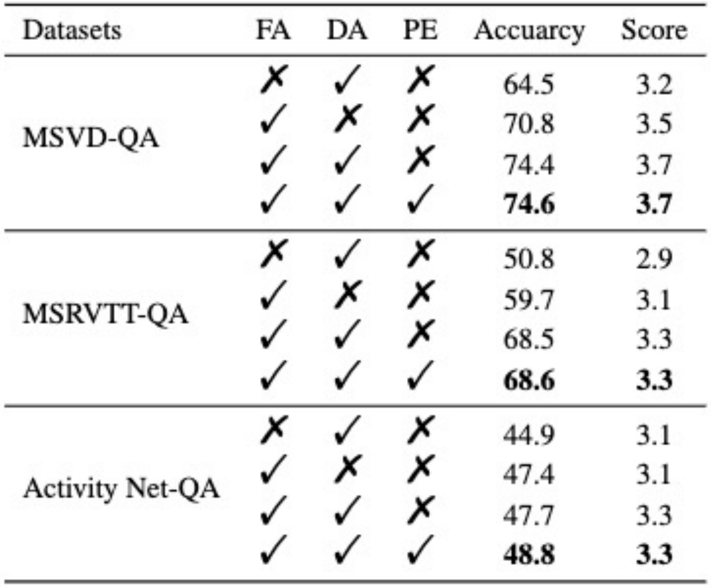
论文一：VaQuitA



实现：
考虑到时间开销，本方案仅在训练阶段执行。实验结果证明，这是有效的。我认为这或许减少了噪声视频本文对在训练过程中带来的消极作用。

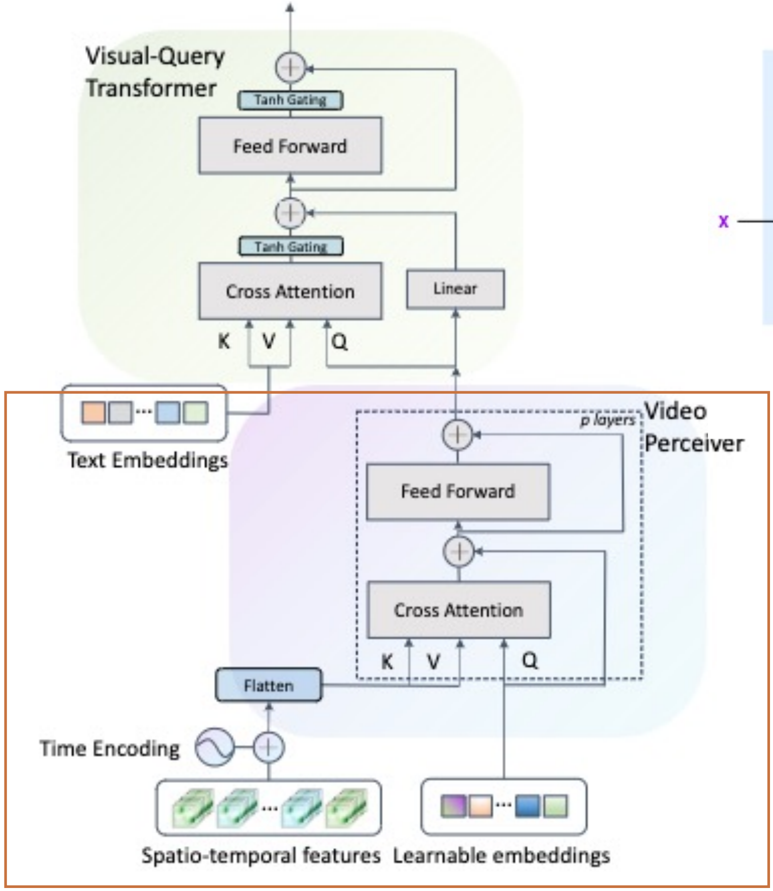
动机：均匀采样造成关键信息丢失。

方案：采样T帧
1. 均匀采样T/2帧；
2. 采样与文本的clip score较高的T/2帧；
3. 重新排序。

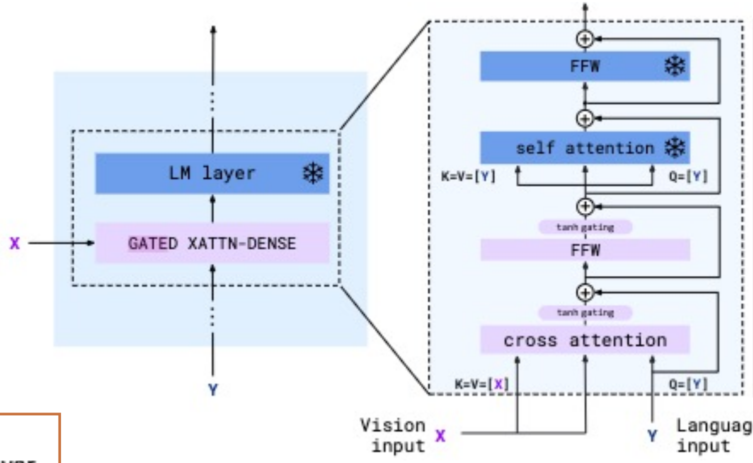


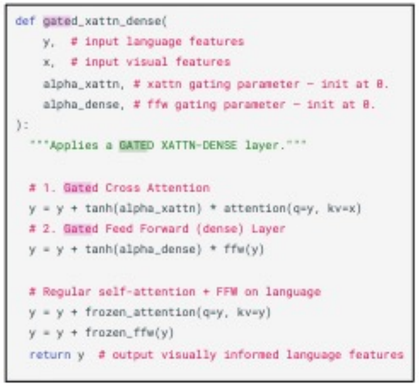
Datasets	FA	DA	PE	Accuarcy	Score
MSVD-QA	✗	✓	✗	64.5	3.2
	✓	✗	✗	70.8	3.5
	✓	✓	✗	74.4	3.7
	✓	✓	✓	**74.6**	**3.7**
MSRVTT-QA	✗	✓	✗	50.8	2.9
	✓	✗	✗	59.7	3.1
	✓	✓	✗	68.5	3.3
	✓	✓	✓	**68.6**	**3.3**
Activity Net-QA	✗	✓	✗	44.9	3.1
	✓	✗	✗	47.4	3.1
	✓	✓	✗	47.7	3.3
	✓	✓	✓	**48.8**	**3.3**

论文一：VaQuitA

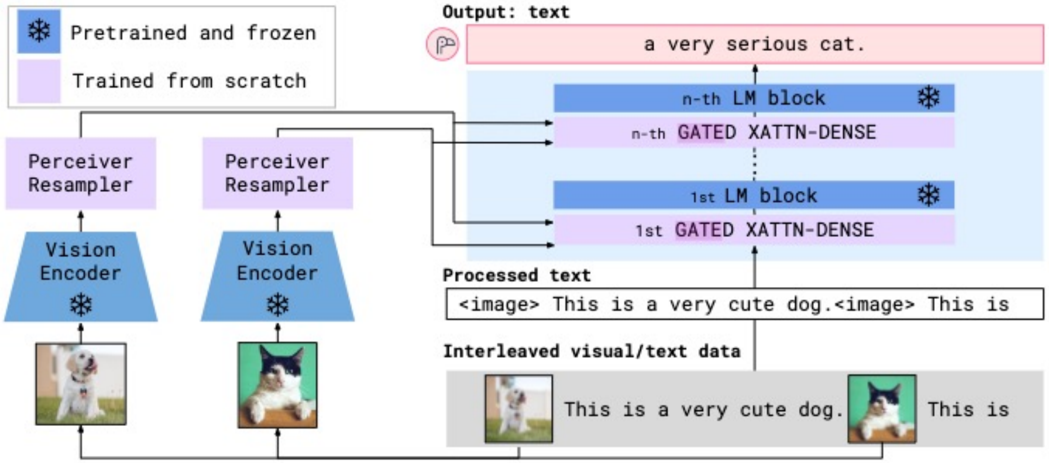


与Video LLaMA一样



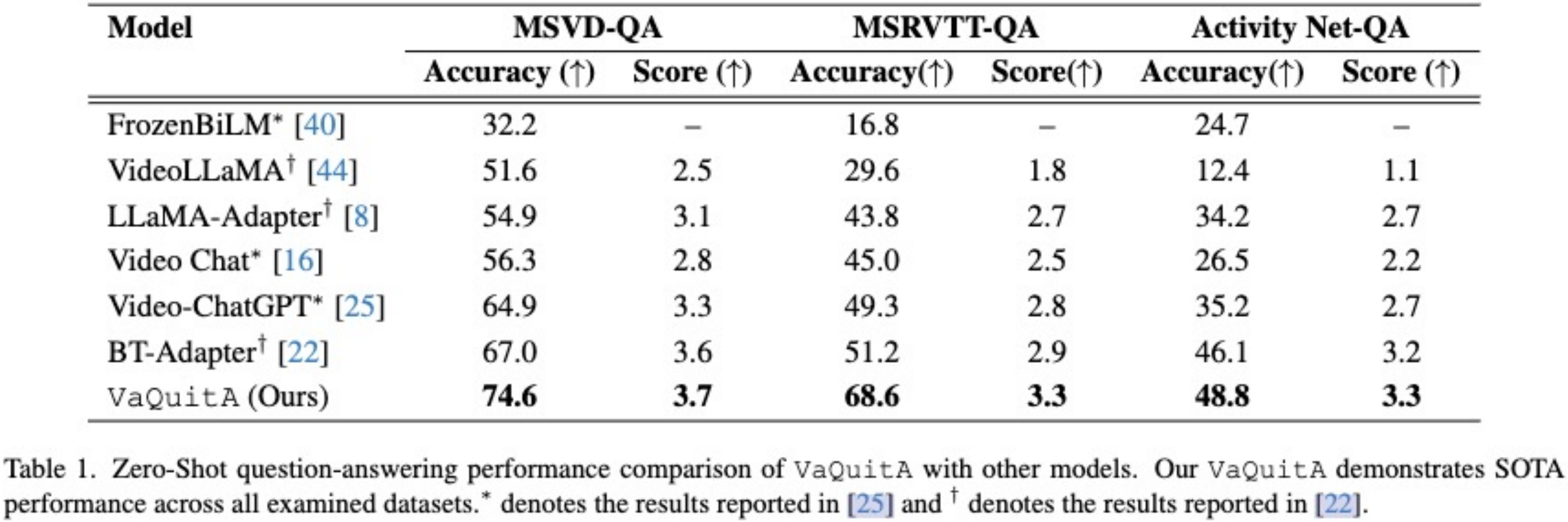


```
def gated_xattn_dense(
    y,  # input language features
    x,  # input visual features
    alpha_xattn, # xattn gating parameter – init at 0.
    alpha_dense, # ffw gating parameter – init at 0.
):
  """Applies a GATED XATTN-DENSE layer."""

  # 1. Gated Cross Attention
  y = y + tanh(alpha_xattn) * attention(q=y, kv=x)
  # 2. Gated Feed Forward (dense) Layer
  y = y + tanh(alpha_dense) * ffw(y)

  # Regular self-attention + FFW on language
  y = y + frozen_attention(q=y, kv=y)
  y = y + frozen_ffw(y)
  return y  # output visually informed language features
```



Tanh(alpha): 可学习标量，门结构有助于增强训练的稳定性

论文一：VaQuitA

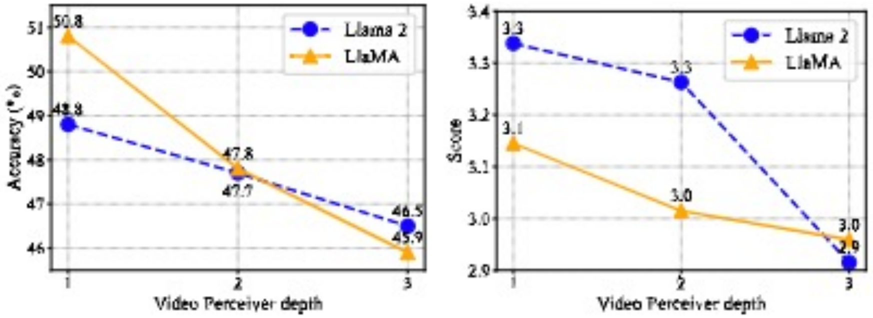


Model	MSVD-QA		MSRVTT-QA		Activity Net-QA	
	Accuracy (↑)	Score (↑)	Accuracy(↑)	Score(↑)	Accuracy(↑)	Score (↑)
FrozenBiLM* [40]	32.2	–	16.8	–	24.7	–
VideoLLaMA† [44]	51.6	2.5	29.6	1.8	12.4	1.1
LLaMA-Adapter† [8]	54.9	3.1	43.8	2.7	34.2	2.7
Video Chat* [16]	56.3	2.8	45.0	2.5	26.5	2.2
Video-ChatGPT* [25]	64.9	3.3	49.3	2.8	35.2	2.7
BT-Adapter† [22]	67.0	3.6	51.2	2.9	46.1	3.2
VaQuitA (Ours)	**74.6**	**3.7**	**68.6**	**3.3**	**48.8**	**3.3**

Table 1. Zero-Shot question-answering performance comparison of VaQuitA with other models. Our VaQuitA demonstrates SOTA performance across all examined datasets.* denotes the results reported in [25] and † denotes the results reported in [22].

数据集	来源	视频数量	问答数量
MSVD-QA	YouTube，日常相关	1970	50505
MSRVTT-QA	YouTube，10个类别	10000	243680
ActivityNet-QA	日常活动，200个类别	5800	58000

论文一：VaQuitA



(a) Accuracy w.r.t. depth p. (b) Score w.r.t. depth p.

Figure 6. Video QA performance on Activity Net-QA [43] using pretrained LLama 2 [29] and LLaMA [28]. Best viewed in color.

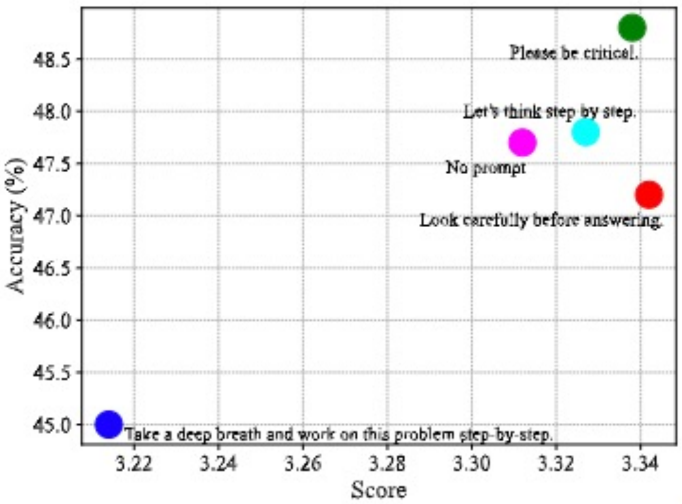


Figure 7. Accuracy and score results on Activity Net-QA [43] dataset of different prompt designs. Best viewed in color.

1. 数据量小，所以随着Perceiver数量增加，性能下降；
2. 我们的实验结果表明似乎论文中提出的提示词方案没有很好的效果。

- □ VideoLLM范式：Video LLaMA
- □ 论文一：VaQuitA
- □ 论文二：Koala
- □ 总结反思

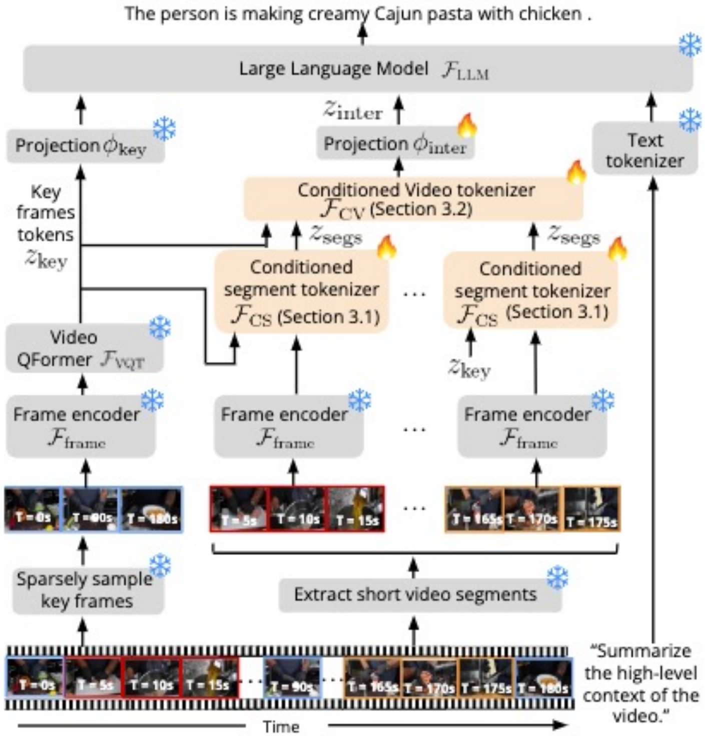
论文二：Koala

Koala: Key frame-conditioned long video-LLM

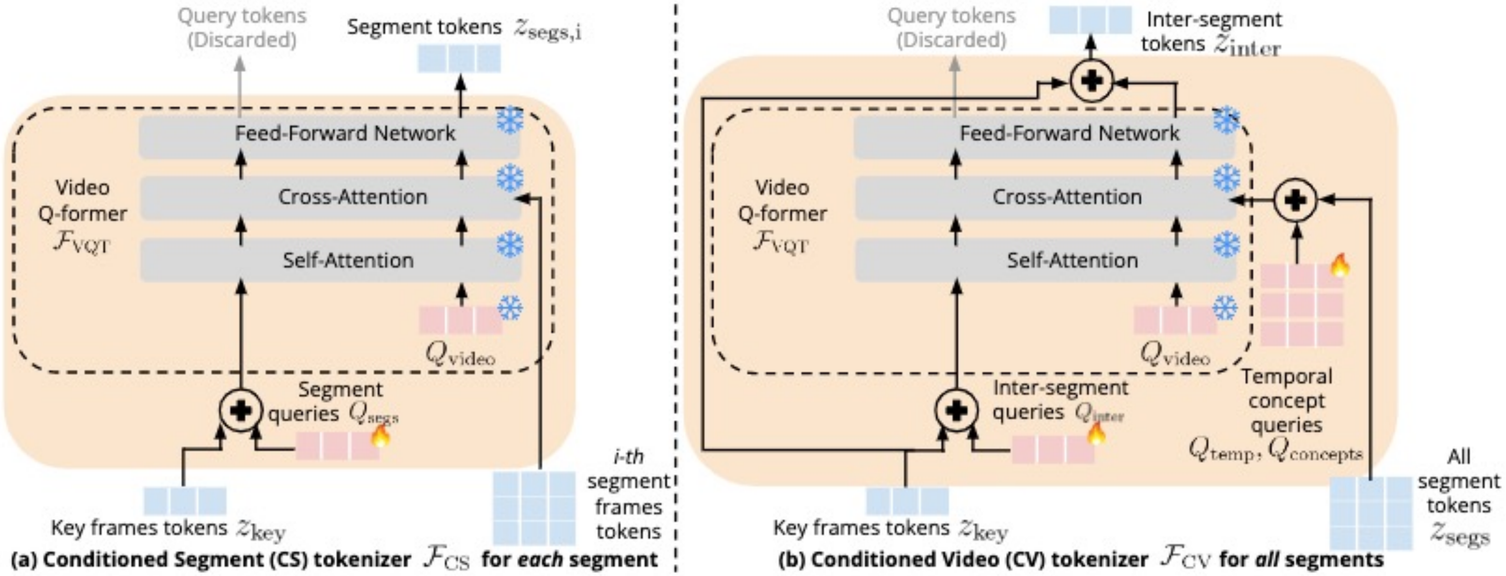
- 长视频理解的关键挑战：识别短期活动和它们之间细粒度的关系。
- 假设：现有的video tokenizer function学会为固定数量的帧聚合时空上下文，可以推广到使用相同数量的输入帧理解更长的视频。

方案：

1. 首先通过以非常粗略的采样率提取相同数量的输入帧来编码长视频的全局上下文，称为关键帧。
2. 为了减轻细粒度时空信息的损失，我们以更高的采样率提取一系列视频片段，以补充具有局部时空信息的全局上下文。


The person is making creamy Cajun pasta with chicken .
Large Language Model $\mathcal{F}_{\text{LLM}}$
z_{inter}
Projection ϕ_{key}
Projection ϕ_{inter}
Text tokenizer
Key frames tokens z_{key}
Conditioned Video tokenizer $\mathcal{F}_{\text{CV}}$ (Section 3.2)
z_{segs}
z_{segs}
Conditioned segment tokenizer $\mathcal{F}_{\text{CS}}$ (Section 3.1)
...
Conditioned segment tokenizer $\mathcal{F}_{\text{CS}}$ (Section 3.1)
Video QFormer $\mathcal{F}_{\text{VQT}}$
z_{key}
Frame encoder $\mathcal{F}_{\text{frame}}$
Frame encoder $\mathcal{F}_{\text{frame}}$
...
Frame encoder $\mathcal{F}_{\text{frame}}$
T = 0s
T = 90s
T = 180s
T = 5s
T = 10s
T = 15s
T = 165s
T = 170s
T = 175s
Sparsely sample key frames
Extract short video segments
T = 0s
T = 5s
T = 10s
T = 15s
T = 90s
T = 165s
T = 170s
T = 175s
T = 180s
"Summarize the high-level context of the video."
Time

论文二：Koala



(a) Conditioned Segment (CS) tokenizer $\mathcal{F}_{\text{CS}}$ for *each* segment

(b) Conditioned Video (CV) tokenizer $\mathcal{F}_{\text{CV}}$ for *all* segments

构建以关键帧为条件的片段信息提取器： $\mathcal{F}_{\text{CS}}(S_i \mid z_{\text{key}}) = \mathcal{F}_{\text{VQT}}(\mathcal{F}_{\text{frame}}(S_i); \text{concat}\{Q_{\text{video}}, z_{\text{key}} + Q_{\text{segs}}\}).$

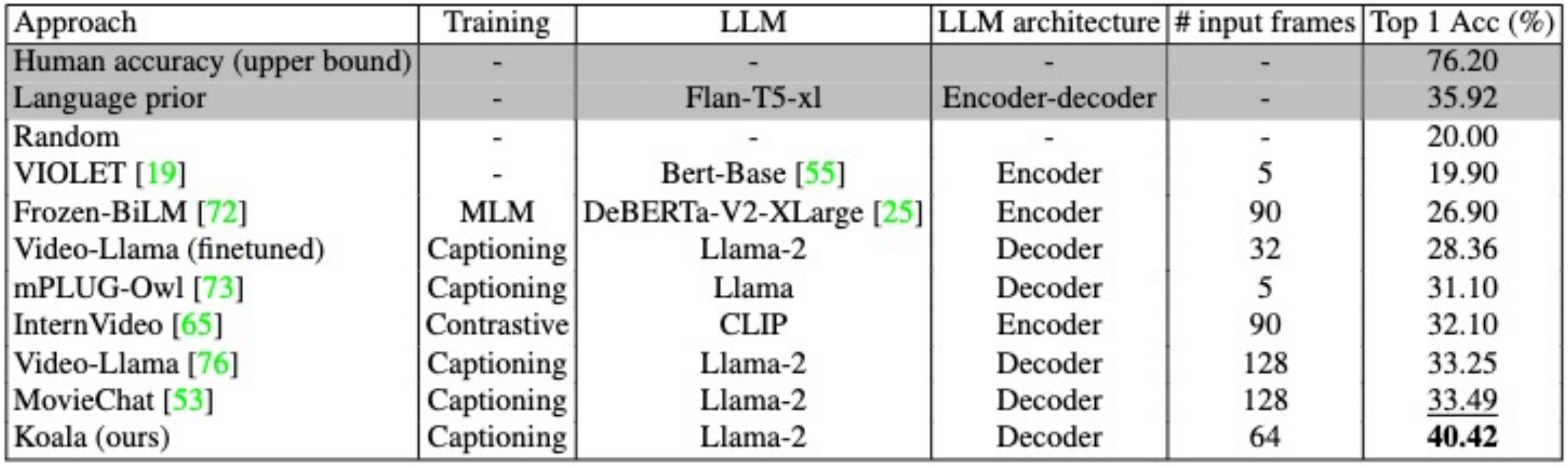
构建以关键帧为条件的片段关系聚合器：

$Q_{\text{final},i,t} = z_{\text{segs},i,t} + Q_{\text{temp},i} + Q_{\text{concepts},t}.$ Q_temp加在片段上，Q_concept加在令牌上

$$\mathcal{F}_{\text{CV}}(z_{\text{segs}} \mid z_{\text{key}}) = z_{\text{key}} + w\mathcal{F}_{\text{VQT}}(Q_{\text{final}}; \text{concat}\{Q_{\text{video}}, z_{\text{key}} + Q_{\text{inter}}\})$$

论文二：Koala

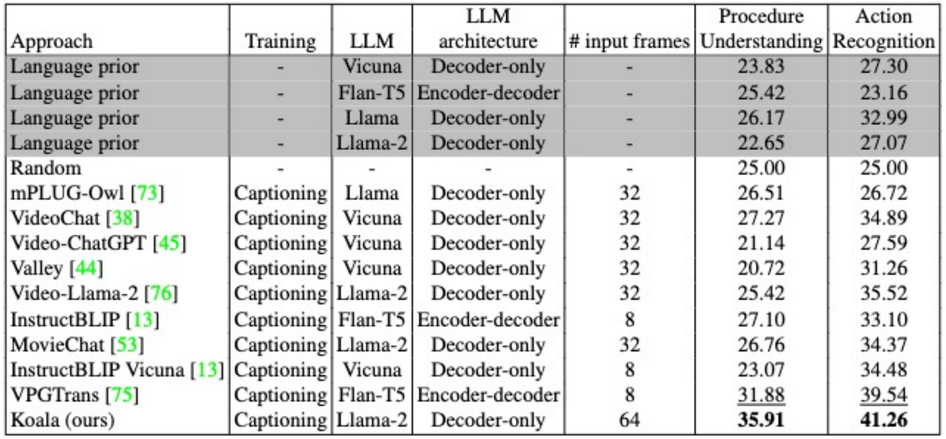
- 训练：从HowTo100M指令视频数据集中过滤出250K的子集，时长从4分钟到30分钟不等。
- 测试：在两个零样本长视频数据集上评估：EgoSchema和SeedBench。



Approach	Training	LLM	LLM architecture	# input frames	Top 1 Acc (%)
Human accuracy (upper bound)	-	-	-	-	76.20
Language prior	-	Flan-T5-xl	Encoder-decoder	-	35.92
Random	-	-	-	-	20.00
VIOLET [19]	-	Bert-Base [55]	Encoder	5	19.90
Frozen-BiLM [72]	MLM	DeBERTa-V2-XLarge [25]	Encoder	90	26.90
Video-Llama (finetuned)	Captioning	Llama-2	Decoder	32	28.36
mPLUG-Owl [73]	Captioning	Llama	Decoder	5	31.10
InternVideo [65]	Contrastive	CLIP	Encoder	90	32.10
Video-Llama [76]	Captioning	Llama-2	Decoder	128	33.25
MovieChat [53]	Captioning	Llama-2	Decoder	128	33.49
Koala (ours)	Captioning	Llama-2	Decoder	64	**40.42**

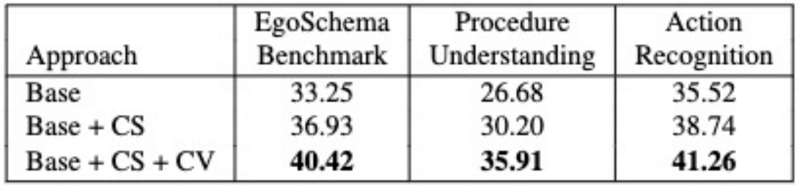
Table 1. **Zero-shot long video question answering on EgoSchema benchmark.** For all models, we report the best results obtained across varying number of input frames. Our Koala approach outperforms the base Video-Llama model despite using much fewer frames. We also include the results for a strong language prior baseline as well as human performance (highlighted in gray).

论文二：Koala



Approach	Training	LLM	LLM architecture	# input frames	Procedure Understanding	Action Recognition
Language prior	-	Vicuna	Decoder-only	-	23.83	27.30
Language prior	-	Flan-T5	Encoder-decoder	-	25.42	23.16
Language prior	-	Llama	Decoder-only	-	26.17	32.99
Language prior	-	Llama-2	Decoder-only	-	22.65	27.07
Random	-	-	-	-	25.00	25.00
mPLUG-Owl [73]	Captioning	Llama	Decoder-only	32	26.51	26.72
VideoChat [38]	Captioning	Vicuna	Decoder-only	32	27.27	34.89
Video-ChatGPT [45]	Captioning	Vicuna	Decoder-only	32	21.14	27.59
Valley [44]	Captioning	Vicuna	Decoder-only	32	20.72	31.26
Video-Llama-2 [76]	Captioning	Llama-2	Decoder-only	32	25.42	35.52
InstructBLIP [13]	Captioning	Flan-T5	Encoder-decoder	8	27.10	33.10
MovieChat [53]	Captioning	Llama-2	Decoder-only	32	26.76	34.37
InstructBLIP Vicuna [13]	Captioning	Vicuna	Decoder-only	8	23.07	34.48
VPGTrans [75]	Captioning	Flan-T5	Encoder-decoder	8	31.88	39.54
Koala (ours)	Captioning	Llama-2	Decoder-only	64	**35.91**	**41.26**

Table 3. **Zero-shot video question answering on Seed-Bench.** Compared to state-of-the-art mLLMs, our Koala approach improves the capability of the vLLM to not only understand long temporal context in procedure understanding but also to recognize short actions. We also compare to language prior baselines with different LLMs (highlighted in gray).



Approach	EgoSchema Benchmark	Procedure Understanding	Action Recognition
Base	33.25	26.68	35.52
Base + CS	36.93	30.20	38.74
Base + CS + CV	**40.42**	**35.91**	**41.26**

Table 4. **Model ablations on the zero-shot evaluation benchmarks.** We ablate the effectiveness of different queries introduced in our Koala approach on all three evaluation tasks.

- □ VideoLLM范式：Video LLaMA
- □ 论文一：VaQuitA
- □ 论文二：Koala
- □ 总结反思

总结反思

- □ 长视频理解处于上升期；
- □ 大部分改动围绕Video Qformer展开；
- □ 目前的测评方式不稳定；