研究论文报告

From Seconds to Hours: Reviewing MultiModal Large Language Models on Comprehensive Long Video Understanding

> 链接：https://ar5iv.labs.arxiv.org/html/2409.18938

主要讲了如何将 LLM 与特定视觉模态编码器进行结合，赋予 LLM 视觉感知能力。

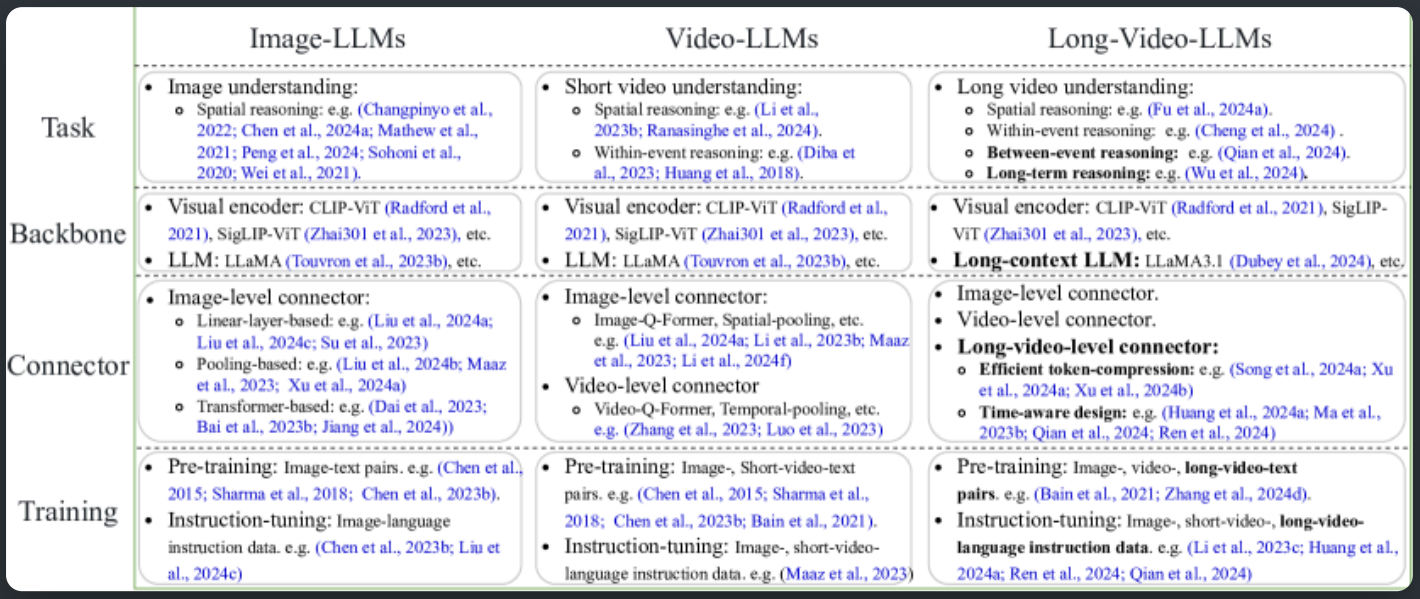


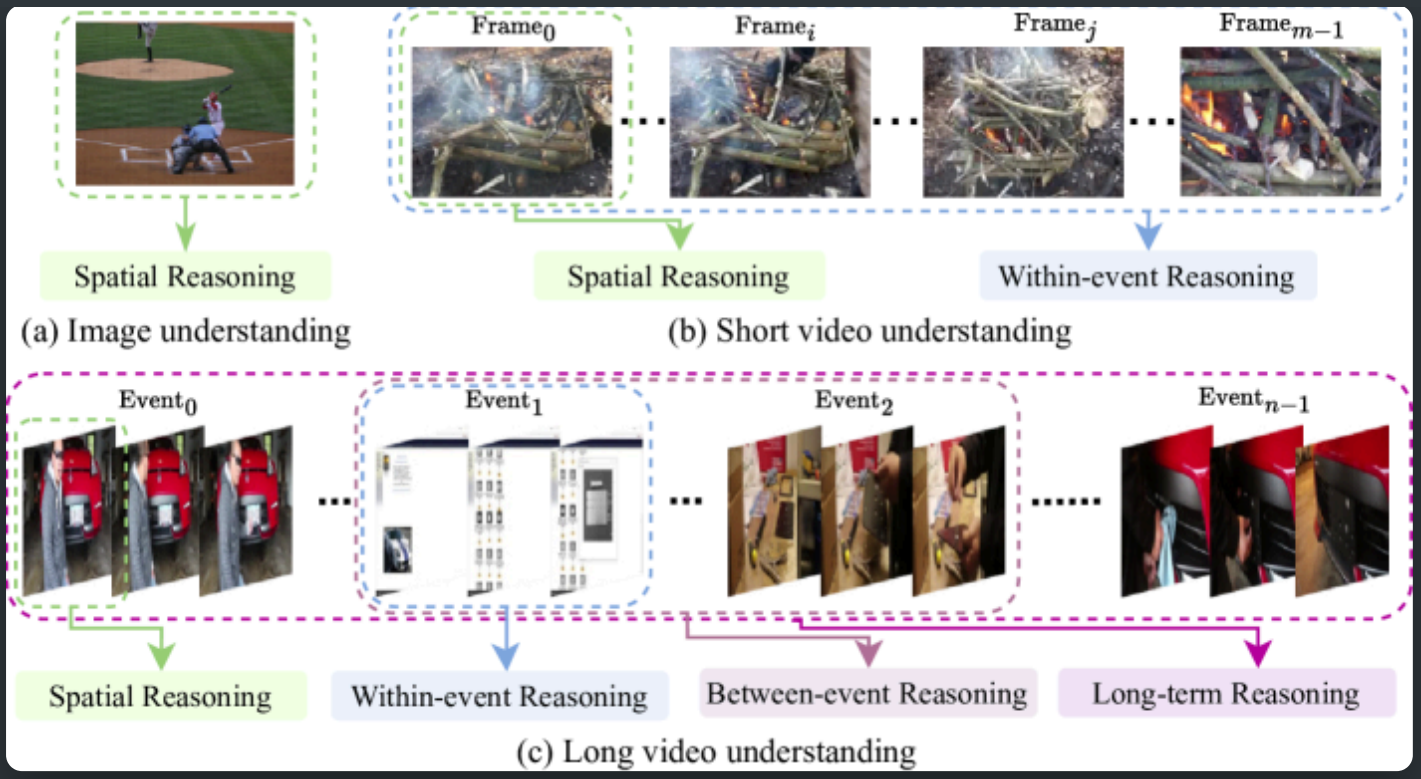
	Image-LLMs	Video-LLMs	Long-Video-LLMs

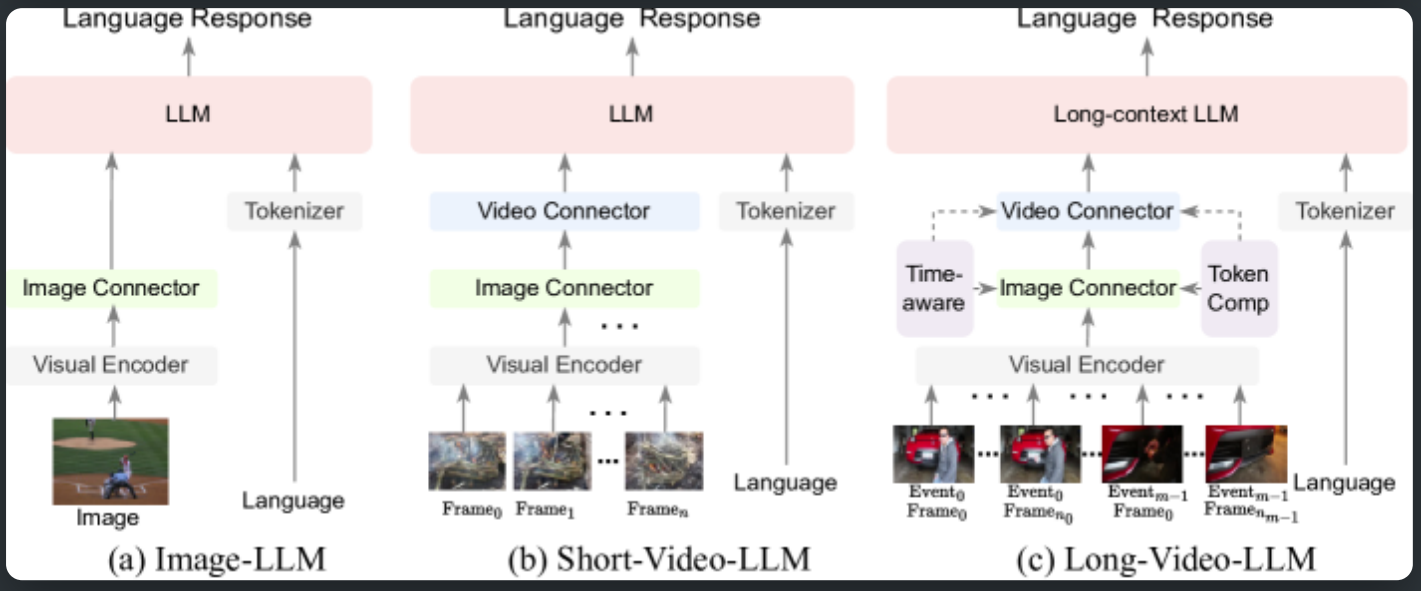
Task	• Image understanding:
	  ◦ Spatial reasoning: e.g. (Changpinyo et al., 2022; Chen et al., 2024a; Mathew et al., 2021; Peng et al., 2024; Sohoni et al., 2020; Wei et al., 2021).
		• Short video understanding:
		  ◦ Spatial reasoning: e.g. (Li et al., 2023b; Ranasinghe et al., 2024).
		  ◦ Within-event reasoning: e.g. (Diba et al., 2023; Huang et al., 2018).
			• Long video understanding:
			  ◦ Spatial reasoning: e.g. (Fu et al., 2024a).
			  ◦ Within-event reasoning: e.g. (Cheng et al., 2024) .
			  ◦ **Between-event reasoning:** e.g. (Qian et al., 2024).
			  ◦ **Long-term reasoning:** e.g. (Wu et al., 2024).

Backbone	• Visual encoder: CLIP-ViT (Radford et al., 2021), SigLIP-ViT (Zhai301 et al., 2023), etc. • LLM: LLaMA (Touvron et al., 2023b), etc.	• Visual encoder: CLIP-ViT (Radford et al., 2021), SigLIP-ViT (Zhai301 et al., 2023), etc. • LLM: LLaMA (Touvron et al., 2023b), etc.	• Visual encoder: CLIP-ViT (Radford et al., 2021), SigLIP-ViT (Zhai301 et al., 2023), etc. • **Long-context LLM:** LLaMA3.1 (Dubey et al., 2024), etc.

Connector	• Image-level connector:
	  ◦ Linear-layer-based: e.g. (Liu et al., 2024a; Liu et al., 2024c; Su et al., 2023)
	  ◦ Pooling-based: e.g. (Liu et al., 2024b; Maaz et al., 2023; Xu et al., 2024a)
	  ◦ Transformer-based: e.g. (Dai et al., 2023; Bai et al., 2023b; Jiang et al., 2024))
		• Image-level connector:
		  ◦ Image-Q-Former, Spatial-pooling, etc. e.g. (Liu et al., 2024a; Li et al., 2023b; Maaz et al., 2023; Li et al., 2024f)
		• Video-level connector
		  ◦ Video-Q-Former, Temporal-pooling, etc. e.g. (Zhang et al., 2023; Luo et al., 2023)
			• Image-level connector.
			• Video-level connector.
			• **Long-video-level connector:**
			  ◦ **Efficient token-compression:** e.g. (Song et al., 2024a; Xu et al., 2024a; Xu et al., 2024b)
			  ◦ **Time-aware design:** e.g. (Huang et al., 2024a; Ma et al., 2023b; Qian et al., 2024; Ren et al., 2024)

Training	• Pre-training: Image-text pairs. e.g. (Chen et al., 2015; Sharma et al., 2018; Chen et al., 2023b). • Instruction-tuning: Image-language instruction data. e.g. (Chen et al., 2023b; Liu et al., 2024c)	• Pre-training: Image-, Short-video-text pairs. e.g. (Chen et al., 2015; Sharma et al., 2018; Chen et al., 2023b; Bain et al., 2021). • Instruction-tuning: Image-, short-video-language instruction data. e.g. (Maaz et al., 2023)	• Pre-training: Image-, video-, **long-video-text pairs**. e.g. (Bain et al., 2021; Zhang et al., 2024d). • Instruction-tuning: Image-, short-video-, **long-video-language instruction data**. e.g. (Li et al., 2023c; Huang et al., 2024a; Ren et al., 2024; Qian et al., 2024)

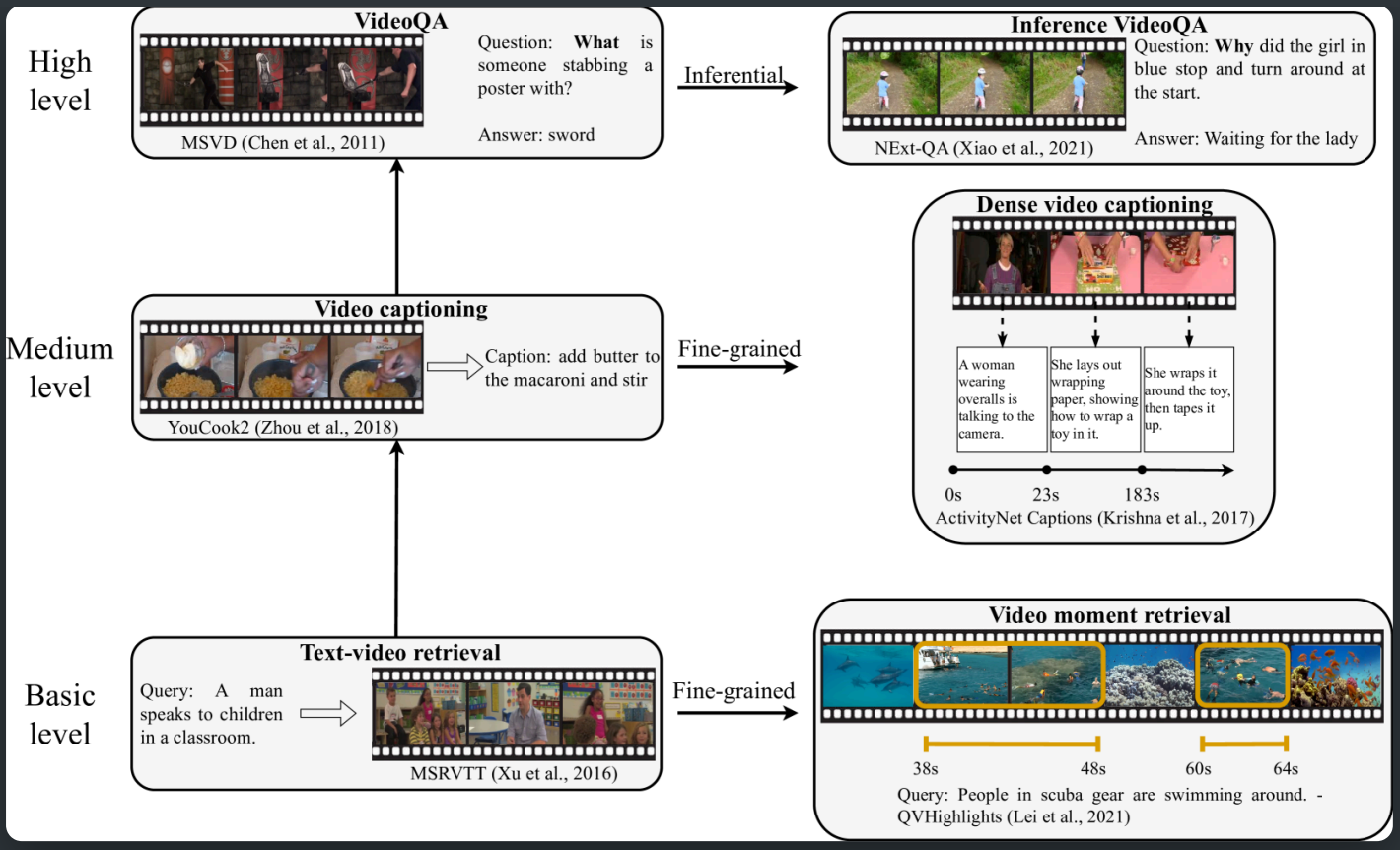




Video-Language Understanding: A Survey from Model Architecture, Model Training, and Data Perspectives

link: https://arxiv.org/abs/2406.05615

相关性不强，只讲了各种视频处理的区别。



⭐Video Token Merging for Long-form Video Understanding

link: https://papers.nips.cc/paper/2024/hash/194fa4536bf36f35a4505d20cd5dd6fc-Abstract-Conference.html

做token merge。

方案：

1. Naive VTM：在 Transformer 中插入 VTM block。会忽略掉重要 token。
2. Region-concentrated VTM：50% 的目标 token 从画面中心抽样。会忽略视频边缘。
3. Motion-based VTM：利用运动信息作权重，认为“运动信更明显”的 token 更重要。
4. Learnable VTM：训练一个模块来估计每个 token 的重要程度，基于 saliency 分数来决定保留/合并哪些 token。
    - 主路径处理 Transformer self-attention，辅路径估计 token 的 saliency 分数。
    - 作者：在内存消耗上降低了 84%，吞吐量提高 6.89x。

整体流程：

- 在每一 VTM block（插入于 Transformer 层）中：
    1. 对当前 token 集 X 计算 saliency（learnable 路径）
    2. Partition $X \to T$（target）+S（source）依据 saliency／策略
    3. 对每 $x_j \in S$ 找匹配 $x_i \in T$（如相似度最大）
    4. 合并 x_j 到 x_i，生成新的 token set Y（|Y| < |X|）
    5. 下一 Transformer 层以 Y 为输入。

Classroom Behavior Recognition Using Computer Vision: A Comprehensive Review（Sensors, 2025）

> link; https://www.mdpi.com/1424-8220/25/2/373

本文针对基于计算机视觉的课堂行为识别问题开展系统性综述。行为计算基于视觉线索，能够大规模、实时地捕捉课堂中教师与学生状态。但当前在“使用计算机视觉识别课堂行为”的研究中，尚缺乏对整体研究现状、目标分类、识别技术、未来趋势的统一共识。

⭐From Seconds to Hours: Reviewing MultiModal Large Language Models on Comprehensive Long Video Understanding

> link：https://github.com/Vincent-ZHQ/Comprehensive-Long-Video-Understanding-Survey
> paper link：https://arxiv.org/pdf/2409.18938

对小时级视频进行处理，讲解了多个大模型的处理能力，有不同模型的基准测试表单。

重点模型：NVILA，LONGVILA，TimeMarker

Empowering LLMs with Pseudo-Untrimmed Videos for Audio-Visual Temporal Understanding

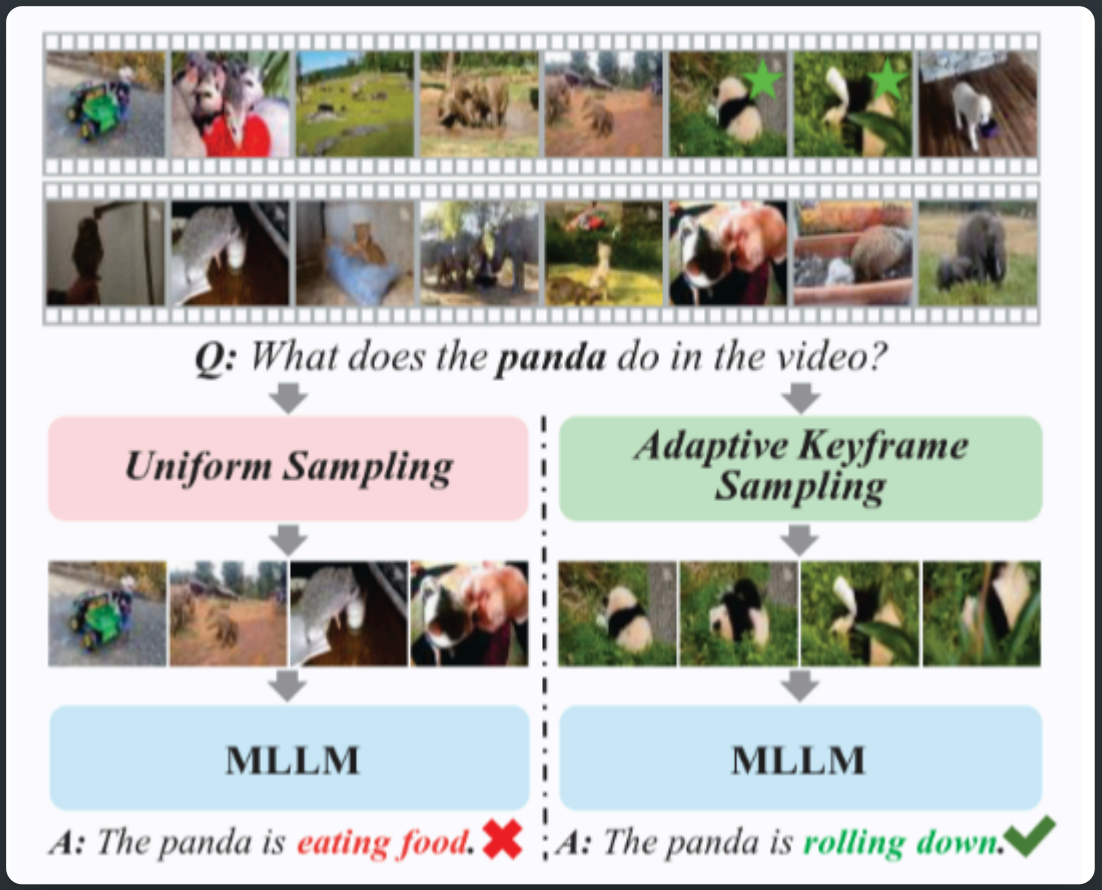
> link：https://arxiv.org/abs/2403.16276

论文目标：让多模态大模型具备时间感知力，把音视频事件，文本描述，时间区间对齐，完成时间定位+对话式问答。

⭐Adaptive Keyframe Sampling for Long Video Understanding

> link: https://openaccess.thecvf.com/content/CVPR2025/papers/Tang_Adaptive_Keyframe_Sampling_for_Long_Video_Understanding_CVPR_2025_paper.pdf
> code：https://github.com/ncTimTang/AKS

自适应关键帧采样（AKS）



关键帧选择原则：相关性、覆盖。

算法流程：

首先，使用一个轻量视觉-语言模型对每帧 F_t 与提示 Q 计算匹配分数 $r(Q, F_t)$。这提供帧与提示之间的“相关性”评分。

然后，通过递归“划分 + 选取”的方式保证覆盖：将时间轴分成若干区间（bins），在每个区间内部根据得分选择若干帧，从而兼顾覆盖各区。

此算法被称为 ADA（Adaptive Sampling），比简单的 TOP（只按得分排序选）或 BIN（仅按均匀时间分箱）更平衡。论文中将这三者与均匀抽样（UNI）做比较。

⭐MovieChat: From Dense Token to Sparse Memory for Long Video Understanding

> link: https://openaccess.thecvf.com/content/CVPR2024/papers/Song_MovieChat_From_Dense_Token_to_Sparse_Memory_for_Long_Video_CVPR_2024_paper.pdf

整体流程

1. 输入：一段“长视频”
2. 视频编码：使用滑动窗口从视频中提取视觉特征或帧 token。
3. 记忆机制：将这些 token 输入一个短期记忆缓冲区；当缓冲满后，将其“整理／合并”转入一个长期记忆中。
4. 最终，短期 + 长期记忆的 token 经过投影/融合后，进入大语言模型做理解／问答／对话。
5. 支持两种模式：
    - Global mode：针对整段视频的理解／问答
    - Breakpoint mode：针对某一特定时间点或片段（事件发生点）做理解／问答。

缺点

- 虽然提高了长视频理解能力，但作者自己指出“感知能力有限”和“时间处理不够精细”。
- 模型可能对“极细粒度动作”区分仍然有难度，因为其聚焦于长时理解而非细胞级动作识别。
- 如果教学视频中包含大量类似但意图不同的动作，可能还需要专门的动作识别子模块 +上下文信息，而不只是“记忆压缩”机制。