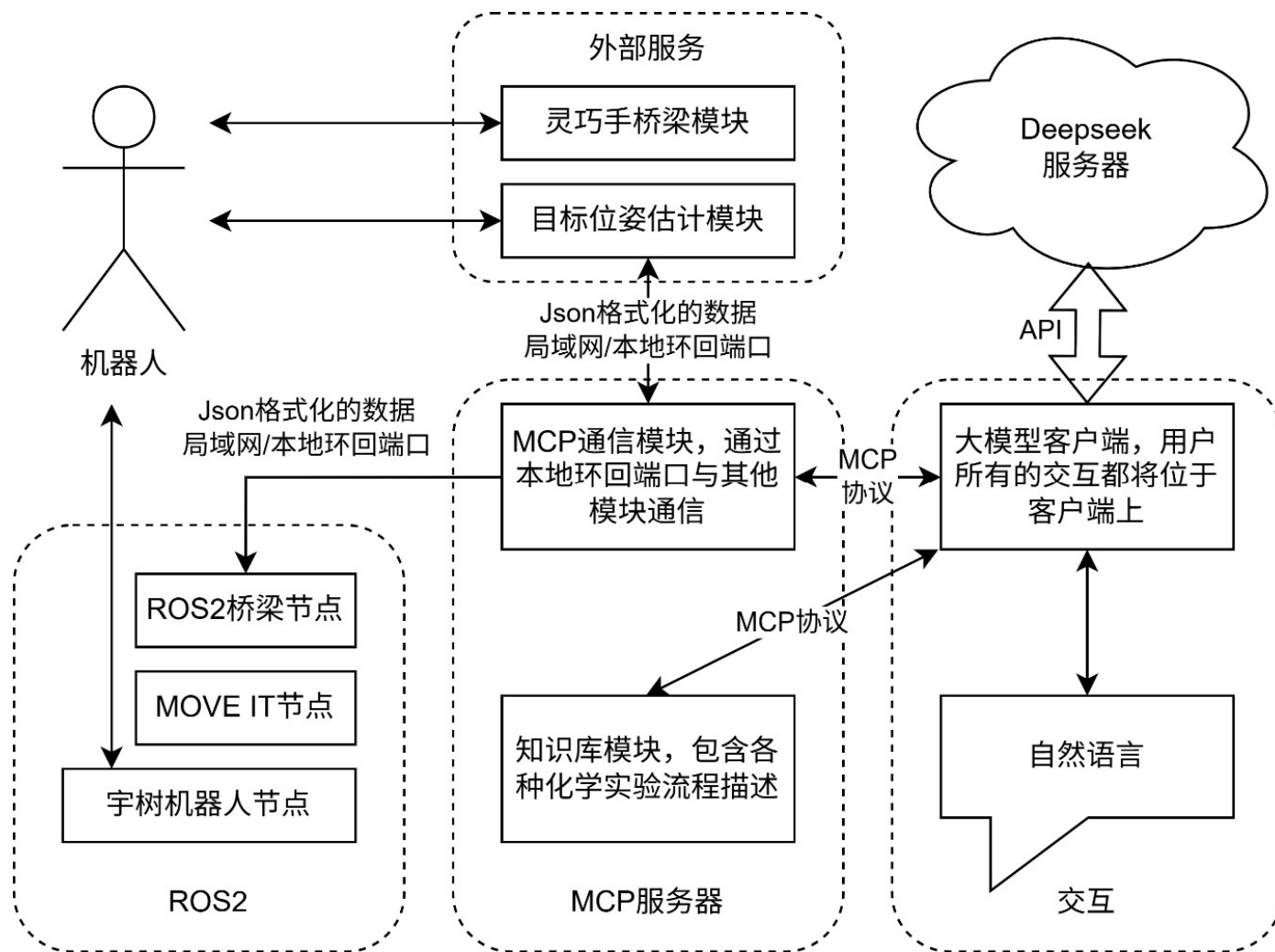
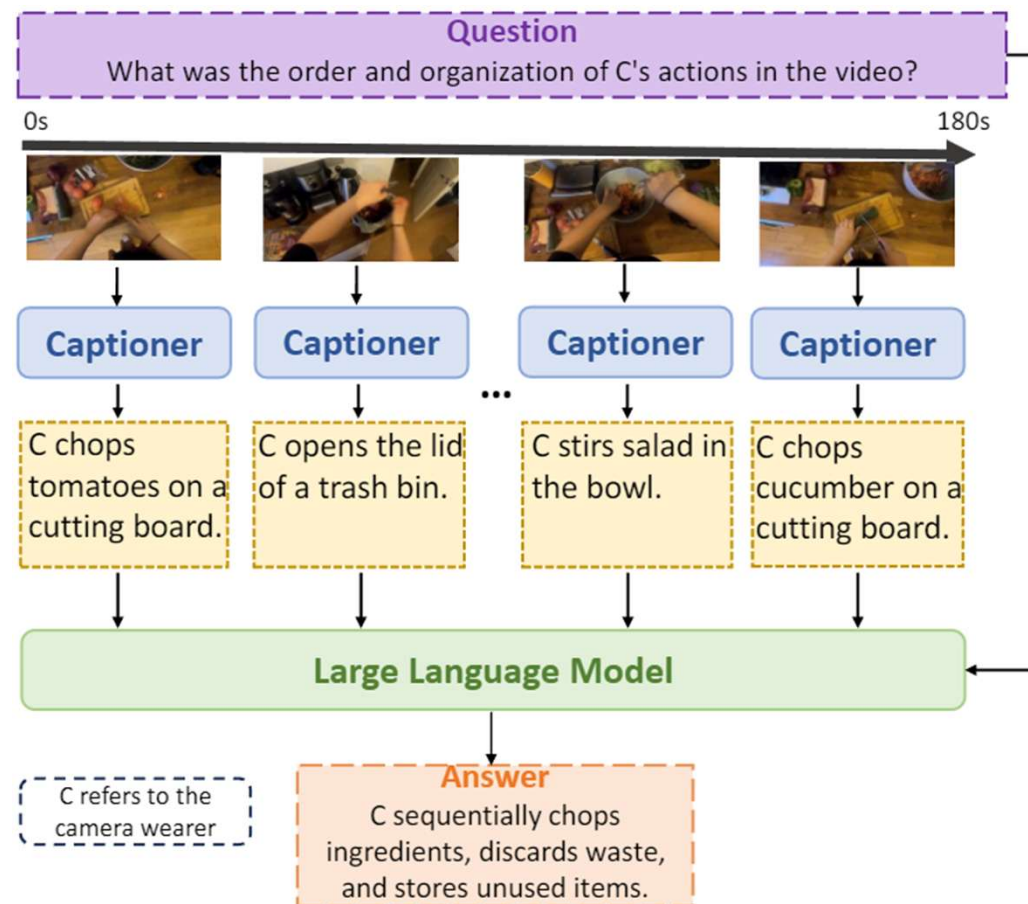


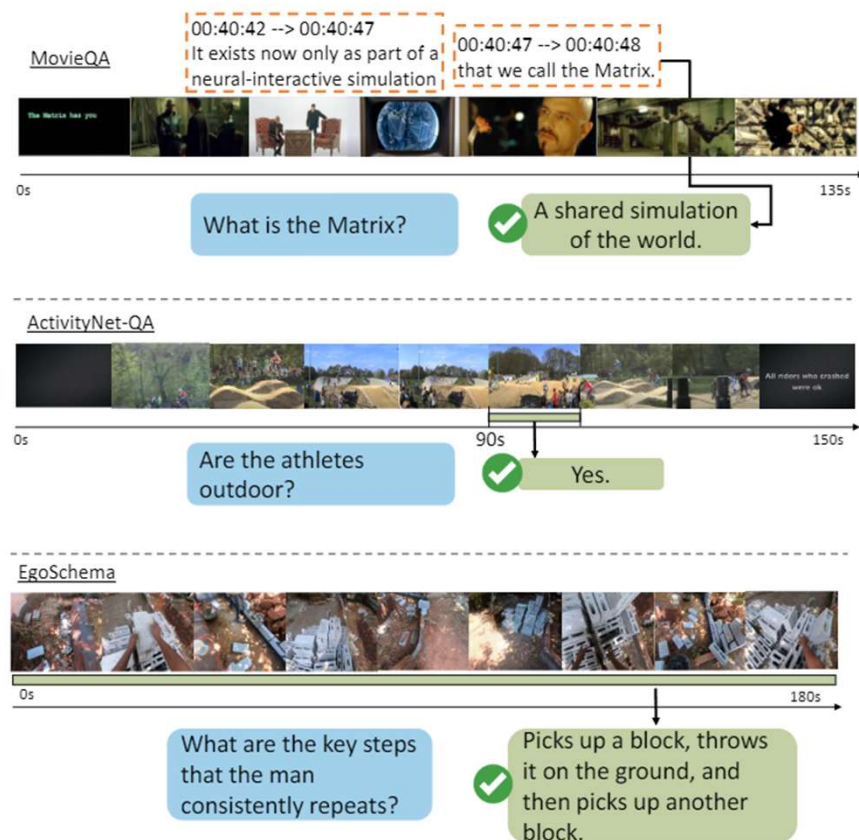


A Simple Framework for Lang-Range Video Question-Answering

- publish: EMNLP 2024
- motivate: 对长视频序列进行理解



A Simple Framework for Lang-Range Video Question-Answering



- 将长视频输入segment into多个短视频, 使用 pretrained visual captioner (BLIP2, LaViLa, LLaVA, ...)提取textual descriptions
- 按顺序拼接caption, 给LLM(如, GPT-4, LLaMA, ...)

Captioner	Caption Type	Ego4D Pre-training	Acc. (%)
EgoVLP (Qinghong Lin et al., 2022)	clip-level	✓	46.6
LLaVA (Liu et al., 2023b)	frame-level	✗	45.2
BLIP-2 (Li et al., 2023d)	frame-level	✗	50.6
LaViLa (Zhao et al., 2023)	clip-level	✓	55.2
Oracle	clip-level	-	66.0

LLM	Model Size	Acc. (%)
Mistral (Jiang et al., 2023)	7B	50.8
Llama3-8B (Touvron et al., 2023)	8B	52.2
Llama3-70B (Touvron et al., 2023)	70B	56.8
GPT-3.5 (OpenAI, 2023a)	175B	55.2
GPT-4 (OpenAI, 2023b)	1.8T	61.2

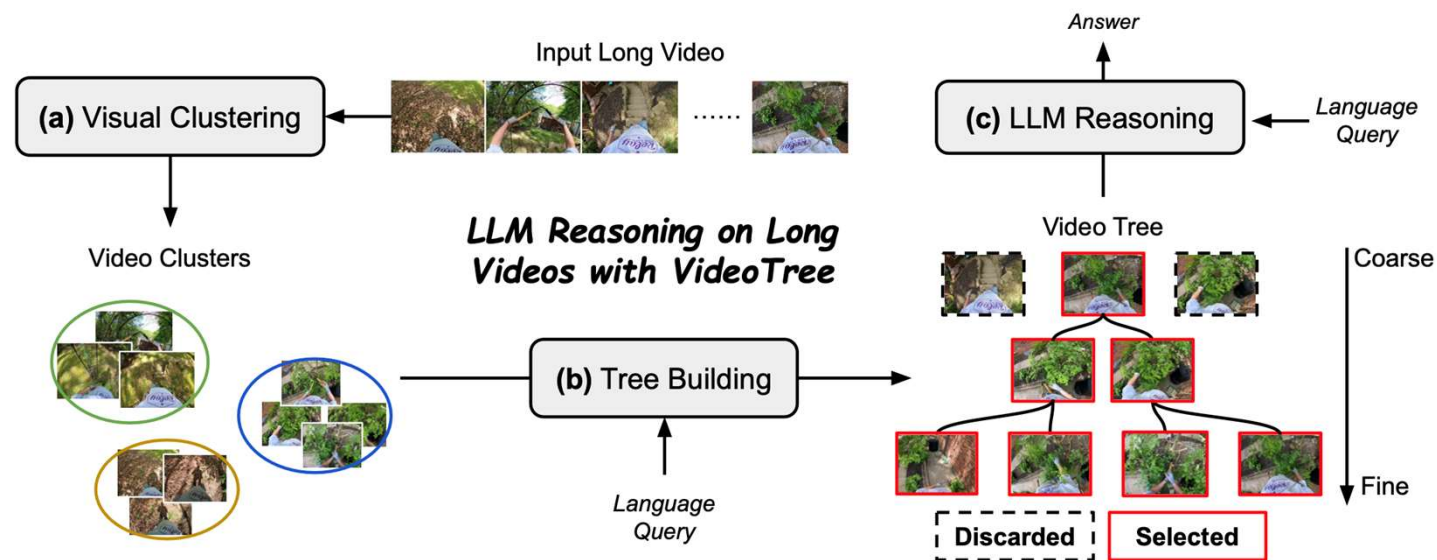
Clip length (s)	1	2	4	8
Acc. (%)	55.2	54.8	53.4	53.4

Clip sampling rate	1	1/2	1/4	1/8
Acc. (%)	55.2	55.2	54.6	53.2

A Simple Framework for Long-Range Video Question-Answering

- 思考:
- 采样的时候是均匀采样, 忽略了视频结构
 - 可能有大量无关信息
 - 可能重要的视频部分被跳过了, 或者描述不详细
 - 一个连续的动作或者场景被切分

VideoTree: Adaptive Tree-based Video Representation for LLM Reasoning on Long Videos

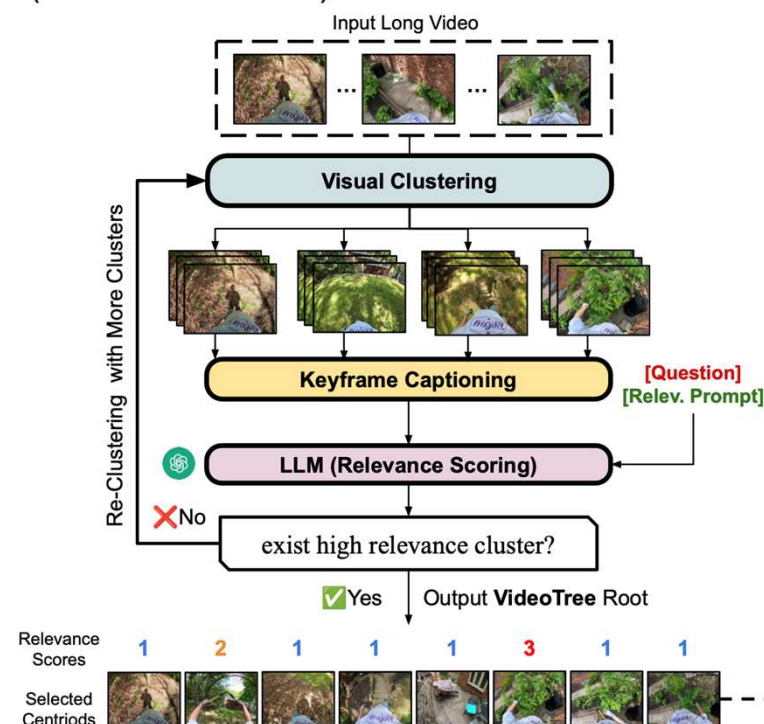


- publish: CVPR 2025
- motivate: 解决问题:
 - 1. 信息冗余, 大量查询无关信息
 - 2. 查询没有粒度的区分. 部分信息密集区域可能需要更细粒度的时序理解

VideoTree: Adaptive Tree-based Video Representation for LLM Reasoning on Long Videos

- Cluster:
 - 使用 pretrained visual encoder 提取 feature
 - 使用 K-means 进行聚类
- Captioning:
 - 使用 pretrained captioner 对每个聚类的中间的一个或几个 key frame 进行 caption, 作为这个聚类的文本描述
- Relevance Score:
 - 使用 LLM, 生成 Query 和每个聚类的文本描述的相关性级别 (Highly Relevant, Somewhat Relevant, Not Relevant)

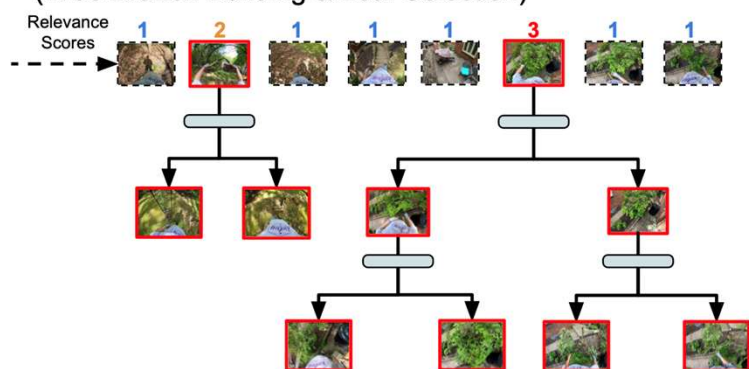
Step-1: Adaptive Breadth Expansion (Tree Root Initialization)



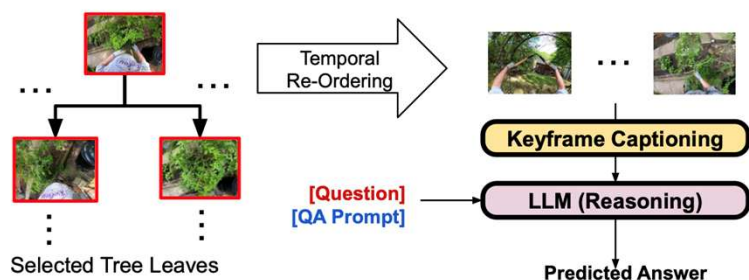
VideoTree: Adaptive Tree-based Video Representation for LLM Reasoning on Long Videos

Step 2: Relevance-guided Depth Expansion

(Tree Branch Building & Leaf Selection)



Step 3: LLM Reasoning with VideoTree



• Depth Expansion

- 对于Somewhat Relevant级别的cluster, 生成一层宽度为 W 的树
- 对于Highly Relevant级别的cluster, 生成两层宽度为 W 的树
- 对于Not Relevant级别的cluster, 不生成树

• LLM Reasoning

- 按顺序将frame和对应的frame tree进行排序
- 从所有cluster中提取相应的文本描述
- 按照时间顺序拼接caption
- 将caption和query给LLM进行回答

VideoTree: Adaptive Tree-based Video Representation for LLM Reasoning on Long Videos

- 思考:
- 对于全部的frame进行caption会很慢, 且会消耗较多的计算资源
 - 如果使用GPT-4等商业大模型, token消耗会特别多
- K-Means的超参数不好设置