M-LLM Based Video Frame Selection for Efficient Video Understanding

Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman,
Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, Trishul Chilimbi

Introduction

Multi-Modal LLMs (M-LLMs) are powerful for video understanding, but long videos pose computational challenges.

Uniform frame sampling in current methods has trade-offs:

--Too few frames → miss key content

--Too many frames → redundant & expensive

Our solution: an adaptive frame selector that takes dense frames as inputs and select much fewer important frames for the M-LLM as inputs.

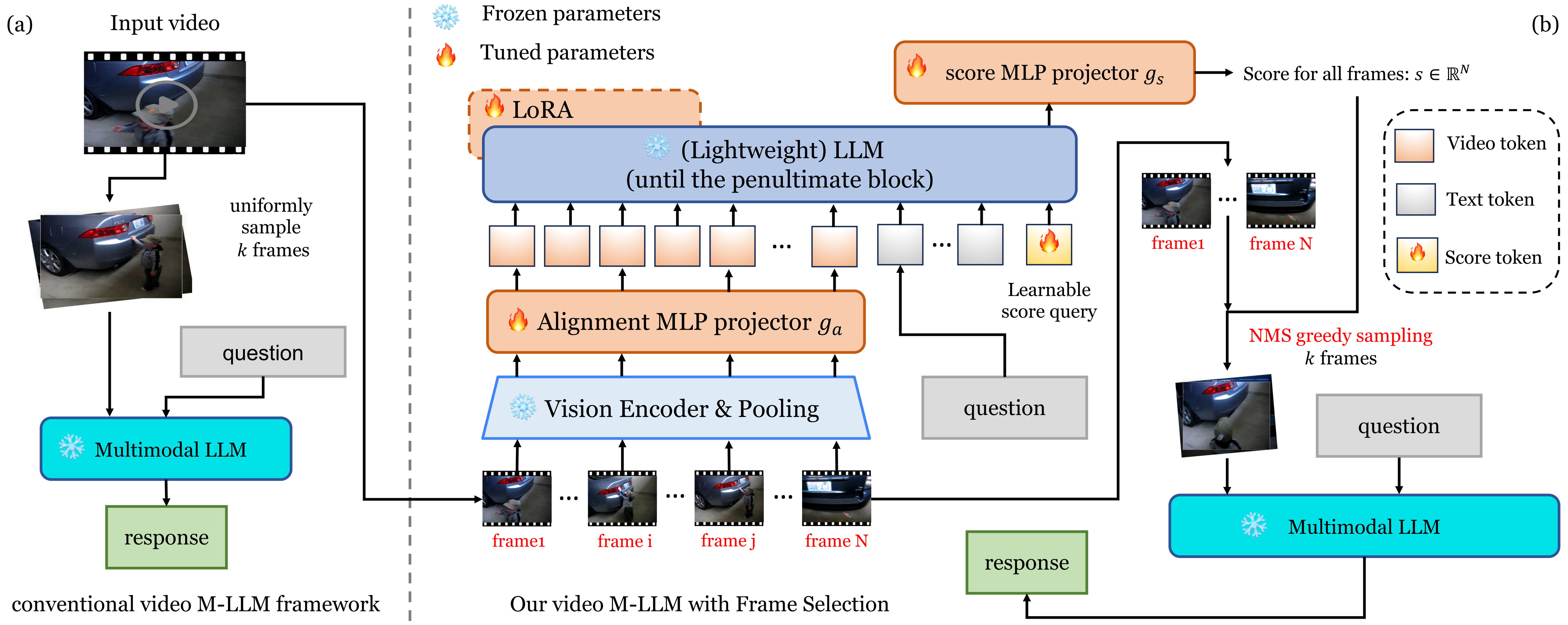
Selecting relevant frames are much easier than understanding every details. A light-weight M-LLM is ideal for the frame selector.

Uniform sampling (8 frames)

Our frame selector (8 frames)

Question : What is the thing attached to the boy's cap? Answer : Price Tag

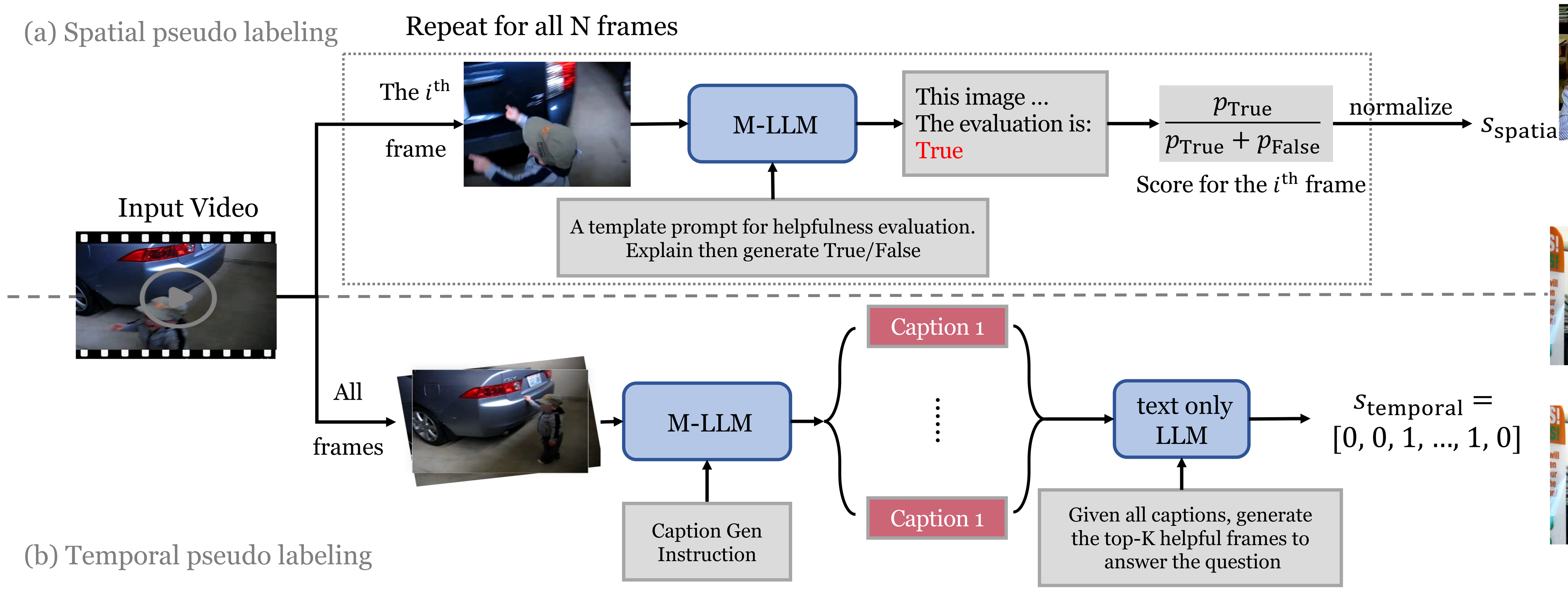
Method



Overview of our framework: a light-weight LLM is fine-tuned as the frame selector. It takes three inputs sequentially: features of uniformly sampled *N* frames (*N* is fixed, we set *N* =128), text question tokens and a learnable score query.

Not a generative model, the frame selector sends the output token of the score query to a score MLP to predict the importance score of each frame, which is used to sample frames.

How to train the selector? Need pseudo labels for the importance scores of each frame!



Results

Model	Model Size	ActivityNet-QA
Video-ChatGPT [31]	7B	35.2 / 2.8
Chat-UniVi [13]	7B	46.1 / 3.3
LLaMA-VID [21]	7B	47.4 / 3.3
LLaMA-VID [21]	13B	47.5 / 3.3
Video-LLaVA [23]	7B	45.3 / 3.3
MiniGPT4-Video [4]	7B	46.3 / 3.4
SlowFast-LLaVA [56]	7B	55.5 / 3.4
SlowFast-LLaVA [56]	34B	59.2 / 3.5
Tarsier [45]	7B	59.5 / 3.6
Tarsier [45]	34B	61.6 / 3.7
PLLaVA [55]	7B	56.3 / 3.5
PLLaVA [55]	34B	60.9 / 3.7
LLaVA-NeXT-Video [62]	7B	53.5 / 3.2
LLaVA-NeXT-Video [62]	34B	58.8 / 3.4
PLLaVA + Selector	7B + 1.5B	57.6 (1.3↑) / 3.5
PLLaVA + Selector	34B + 1.5B	62.3 (1.4↑) / 3.6
LLaVA-NeXT-Video + Selector	7B + 1.5B	55.1 (1.6↑) / 3.4
LLaVA-NeXT-Video + Selector	34B + 1.5B	60.2 (1.4↑) / 3.5

Table 1. Comparison of open-ended question answering evaluation on ActivityNet QA. Results with the "**+ Selector**" are ours.

Model	Model Size	EgoSchema
SlowFast-LLaVA [56]	7B	47.2
SlowFast-LLaVA [56]	34B	55.8
Tarsier [45]	7B	56
Tarsier [45]	34B	68.6
LLaVA-NeXT-Video [62]	7B	45.8
LLaVA-NeXT-Video [62]	34B	48.6
Idefics2 [15]	8B	56.6
Qwen2-VL [46]	7B	64.6
LLaVA-NeXT-Video + Selector	7B + 1.5B	47.2 (1.3↑)
LLaVA-NeXT-Video + Selector	34B + 1.5B	50.6 (2.0↑)
Idefics2 + Selector	8B + 1.5B	57.9 (1.3↑)
Qwen2-VL + Selector	7B + 1.5B	65.9 (1.1↑)

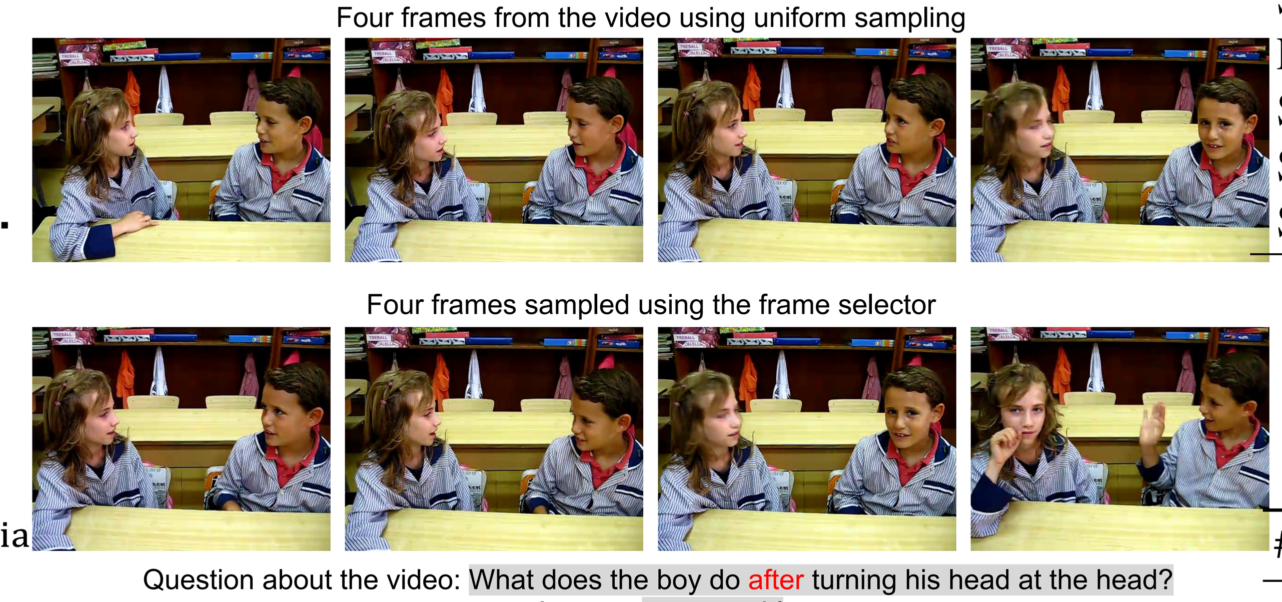
Table 2. Comparison of multi-choice question answering evaluation on EgoSchema. Results with the "**+ Selector**" are ours.

Effectiveness of our proposed methods on three video question answering benchmarks

More visualization of the frames selected

Four frames from the video using uniform sampling

Four frames sampled using the frame selector



Question about the video: What does the boy do after turning his head at the head?
Answer: wave at girl

Four frames from the video using uniform sampling

Four frames sampled using the frame selector

Question about the video: Why did the man with the cap move his hands at the start?
Answer: drink water

Selection Method	ANet-QA	NExT-QA
Uniform sampling	53.5	62.4
Scores from CLIP similarity	53.7	62.2
Pseudo labels from SeViLA	54.0	63.2
Spatial pseudo labels	54.2	63.6
Spatial & temporal pseudo	55.5	63.9
Scores from trained selector	55.1	63.4

Ablation study on the frame selection method: uniform *v.s.* from pseudo labels *v.s.* from predicted scores (using pseudo)

# frames	Acc@C	Acc@T	Acc@D	Acc	Speed (s)
4	67.2	61.2	73.9	66.4	0.56
8	68.7	62.5	76.9	68.1	0.92
16	69.1	63.6	76.8	68.7	1.71
32	69.5	64.3	78.4	69.3	3.40
128 → 4	68.5	64.5	75.7	68.5	0.76
128 → 8	69.3	64.9	77.5	69.3	1.12
128 → 16	69.4	64.8	78.5	69.5	1.91
128 → 32	69.2	65.6	78.7	69.6	3.50

Performance and inference speed of LLaVA-NeXT-Video on NExT-QA with varying input frames. First 4 rows: uniform sampling, last 4 rows: sampling using the selector.