

ReQuest: Rethinking-based Question-Aware Frame Selection for Long-Form Video QA

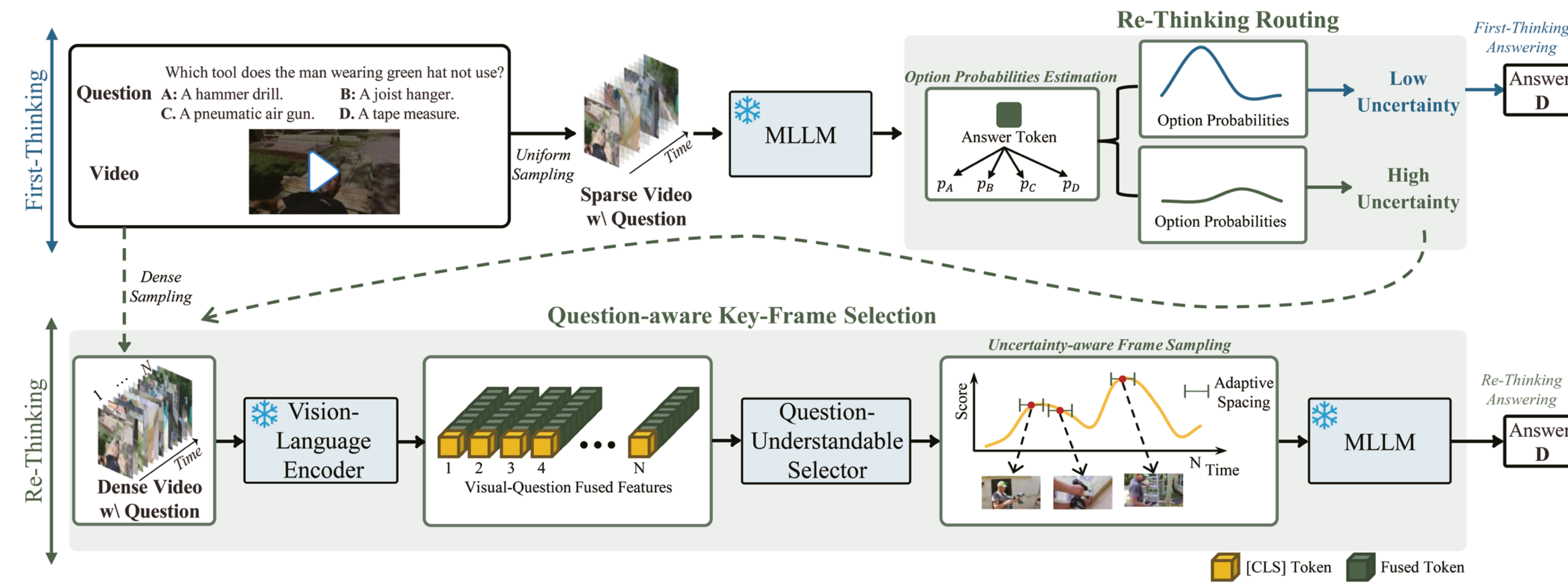
Minkuk Kim^{1*}, Suyong Yun^{1*}, Young Tae Kim¹, Jinyoung Moon², Jinwoo Choi¹, Seong Tae Kim¹
¹ Kyung Hee University ² Electronics and Telecommunications Research Institute (ETRI)



TL;DR ReQuest is a plug-and-play, question-adaptive keyframe selection pipeline for long-form video QA that requires no MLLM fine-tuning. It re-thinks with dense frames only when the model is uncertain, achieving consistent gains on Video-MME, MLVU, and LongVideoBench, especially on medium/long videos.

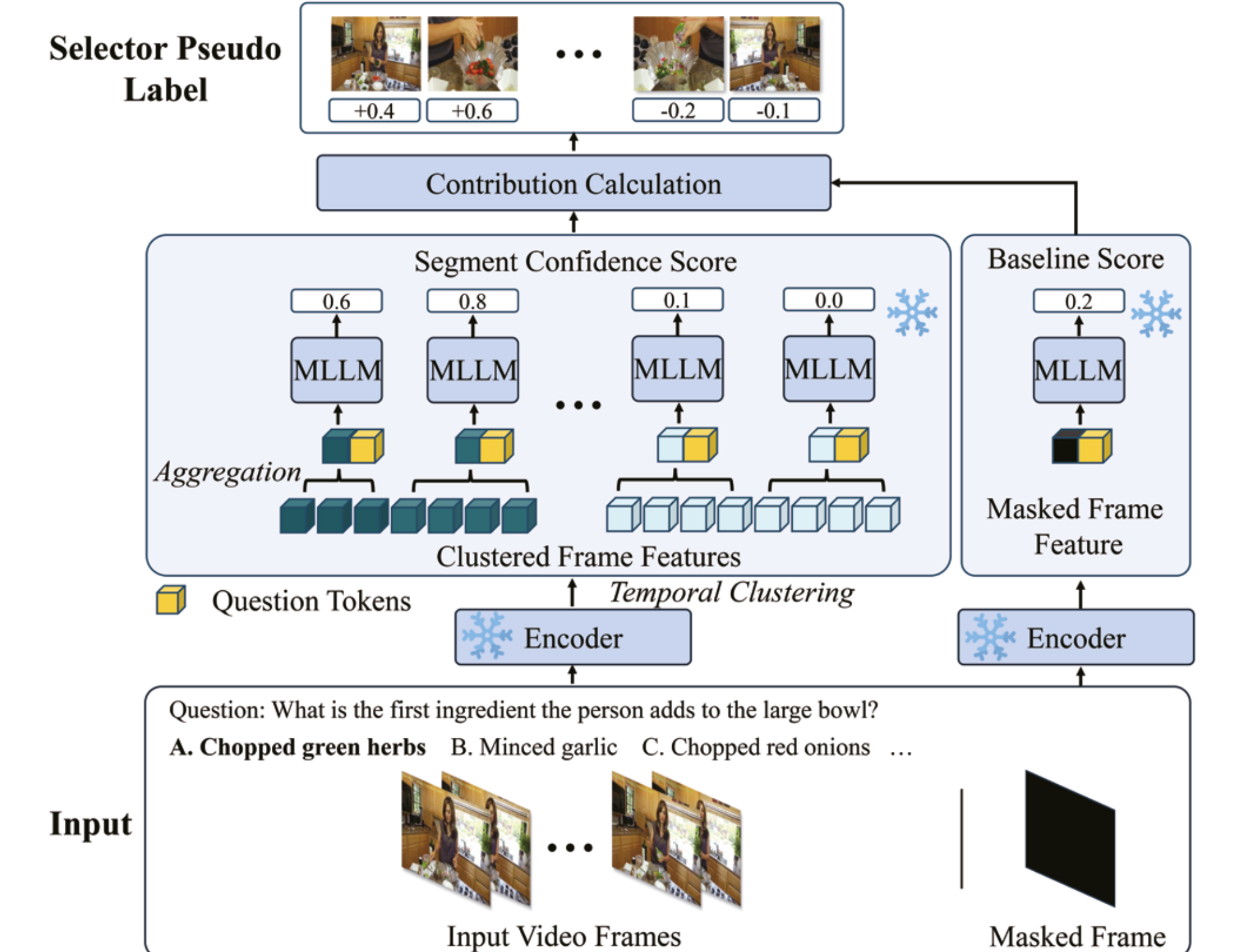
Uncertainty-Guided Re-thinking Reasoning

- First-Thinking:** MLLM answers from uniform sparse frames; answer-token entropy measures uncertainty
 - Uncertain questions to Re-thinking; confident ones answer directly
- Re-thinking:** dense frames scored by a lightweight question-aware selector (13M)
 - We picks K informative, non-redundant frames lwith uncertainty guided diversity sampling



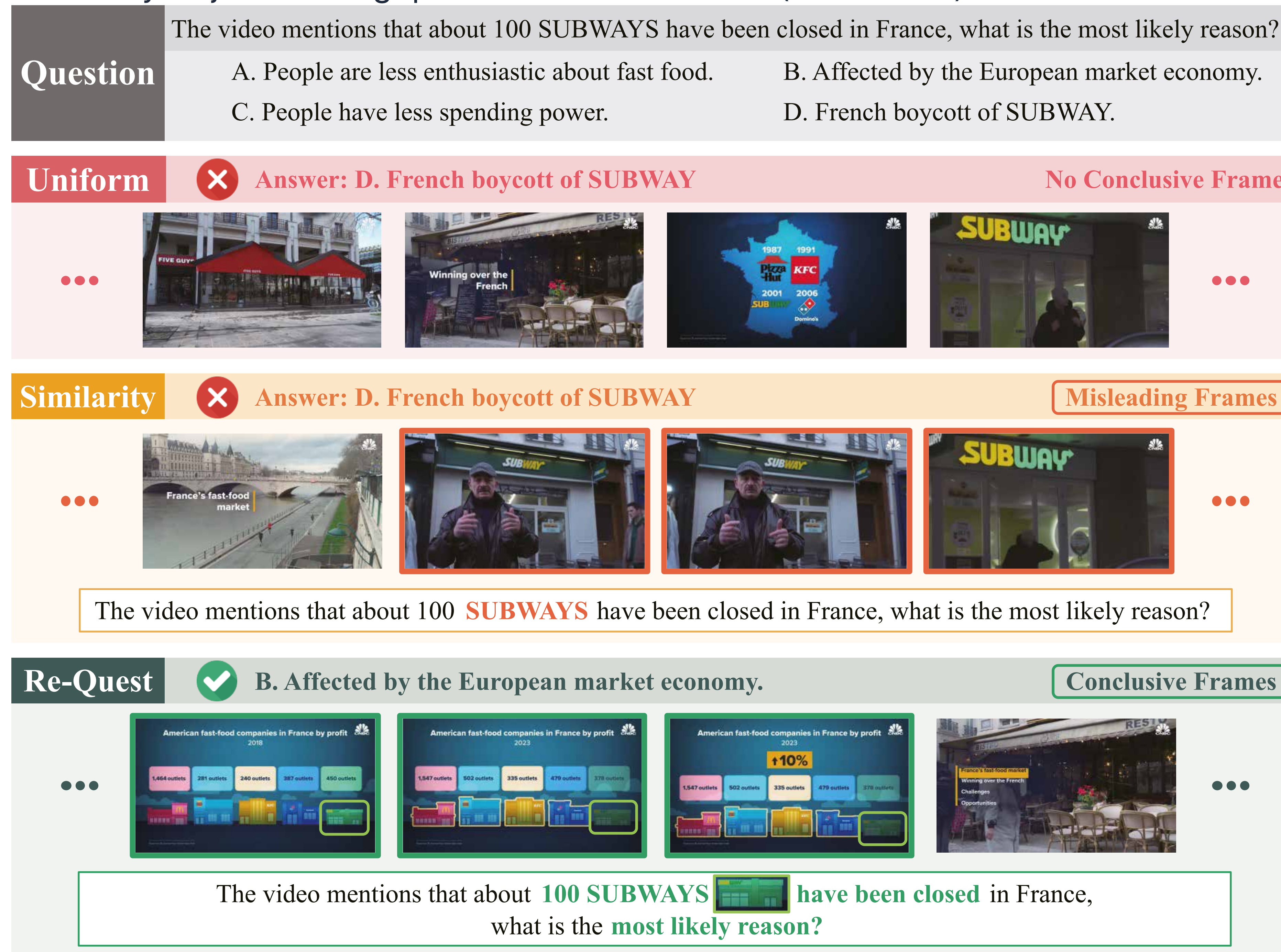
Question Understandable Selector Training

- Query MLLM per video segment
 - correct-answer probability;
 - masked visual input gives text-prior baseline
- Contribution $\Delta c = p_{\text{seg}} - p_{\text{base}}$ isolates net visual influence
- A lightweight selector is trained with Δc as a pseudo label



Retrieving Evidence, Not Keywords

- Similarity selectors over-focus on lexical matches ("SUBWAY") → misleading frames
- ReQuest retrieves informative frames (store statistics) matching the question's intent
- On key-object-missing questions: 53.0% vs 48.0 (feature sim)



Main Results: Plug-and-Play Gains Across Backbones

- LLaVA-Video: VideoMME +3.0, MLVU +7.0, LVB +1.9; largest gain on medium/long videos
- Zero-shot selector transfer improves LLaVA-OneVision and Qwen3-VL without retraining
- Even 512-frame Qwen3-VL benefits: fewer, better frames beat denser uniform sampling

Model	LLM Size	#Frames	LVB	MLVU m-avg	Video-MME (w.o. sub.)			
					Overall	Short	Medium	Long
Frame-Voyager	7B	8	—	65.6	57.5	67.3	56.3	48.9
LongVU	7B	1fps	—	65.4	60.9	64.7	58.2	59.5
NVILA	8B	1024	57.7	70.1	64.0	75.0	62.2	54.8
LLoVi	—	—	—	55.1	54.7	62.1	53.2	48.8
VideoTree	—	—	—	60.4	60.6	67.8	59.9	54.2
LLaVA-Video [†]	7B	32	58.0	64.7	62.6	76.2	59.3	52.2
+ REQUEST	7B	32	60.1	71.7	65.6	77.0	64.1	55.8
LLaVa-OneVision [†]	7B	32	56.6	63.1	58.7	70.3	56.6	49.2
+ REQUEST*	7B	32	60.2	68.8	60.9	71.7	58.8	52.3
Qwen3-VL [†]	8B	512	62.7	74.0	70.0	78.6	70.1	61.2
+ REQUEST*	8B	≤ 512	66.3	76.2	71.1	80.0	70.8	62.4

Cost-Accuracy: Routing Makes Dense Observation Practical

- Dense observation only for routed samples (640/900 LLaVA-Video, 390/900 Qwen3-VL)
- vs. similarity: lower latency (14.7→12.7s), long acc (54.3→55.8) with 0.3s selection cost

Frame Selection	Answer Model	#Frames	Latency Breakdown (s)			Total (s)	Long Acc.
			Feature Ext.	Selection	MLLM Inf.		
No Selection (Baselines)							
Uniform	LLaVA-Video	32	0.5	–	1.6	2.1	52.2
Uniform	Qwen3-VL-8B	512	3.4	–	11.6	15.0	61.2
Key Frame Selection (Dense Observation)							
Similarity	LLaVA-Video	1fps→32	13.1	1.3×10^{-3}	1.6	14.7	54.3
REQUEST	LLaVA-Video	1fps→32	9.4	0.3	2.7	12.7	55.8
REQUEST	Qwen3-VL-8B	1fps→≤ 512	8.8	0.3	16.2	25.1	62.4