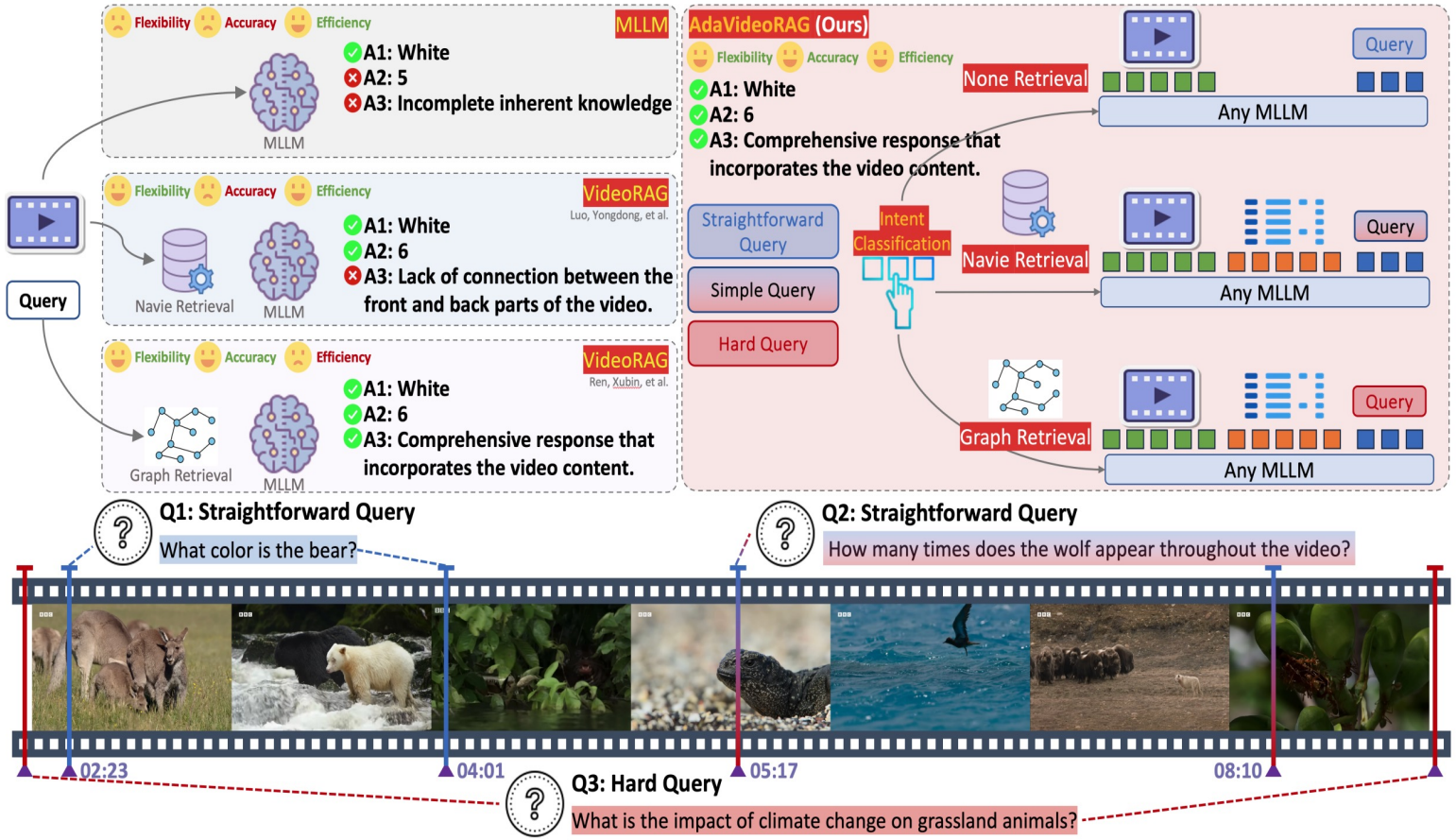
AdaVideoRAG: Omni-Contextual Adaptive Retrieval-Augmented Efficient Long Video Understanding

Zhucun Xue[1], **Jiangning Zhang**[†1], Xurong Xie[1], Yuxuan Cai[3], Yong Liu[1], Xiangtai Li[4], Dacheng Tao[4],
[1]Zhejiang University [2] YouTu Lab [3]Huazhong University of Science Technology
[4]Nanyang Technological University

Github： https://github.com/xzc-zju/AdaVideoRAG

Introduction



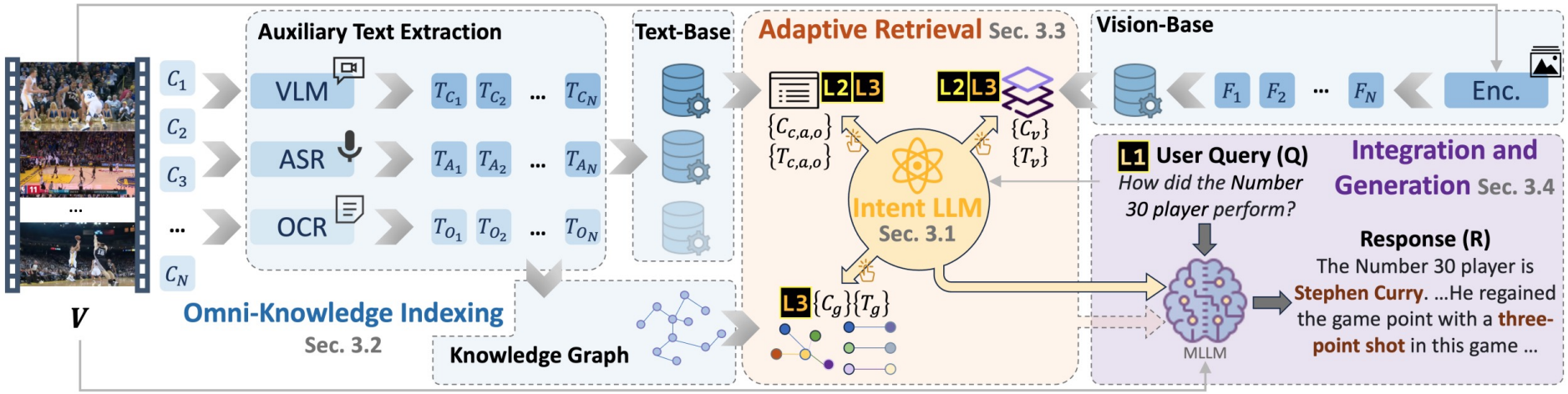
Motivation

- **Inherent Limitations of MLLMs**: Struggle with long video processing due to fixed context windows, failing to model long-term dependencies effectively
- **Shortcomings of Traditional VideoRAG**: adopt a "one-size-fits-all" fixed retrieval paradigm:

Contribution

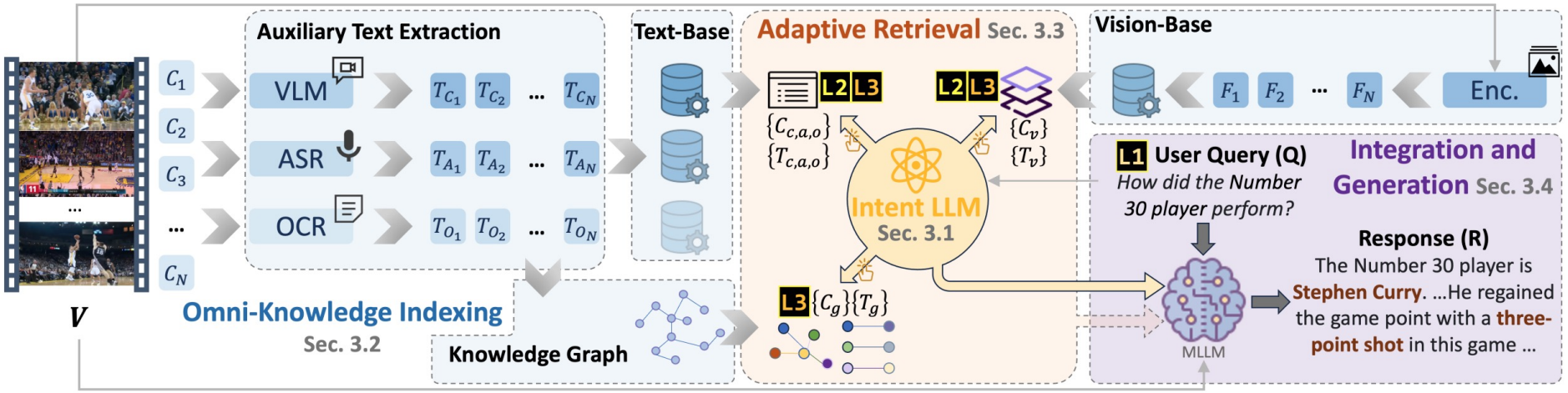
- **Proposing the AdaVideoRAG Framework**: a lightweight intent classifier is used to dynamically assign retrieval schemes.
- **Designing the Omni-Knowledge Indexing Module**: Extract information from multimodal signals and build three types of databases to support hierarchical knowledge access:
- **Constructing the HiVU Benchmark Dataset**

Method



- **Query Intent Classification**: Classify by query complexity levels and trigger different retrieval strategies;
- **Omni-Knowledge Indexing**: Extract information from multimodal signals and build three types of databases;
- **Adaptive Retrieval Paradigm**: Match retrieval methods according to intent levels, ranging from no retrieval to graph retrieval;
- **Integration and Generation:** Integrate retrieval results and input them into MLLM to generate the final answer.

Method



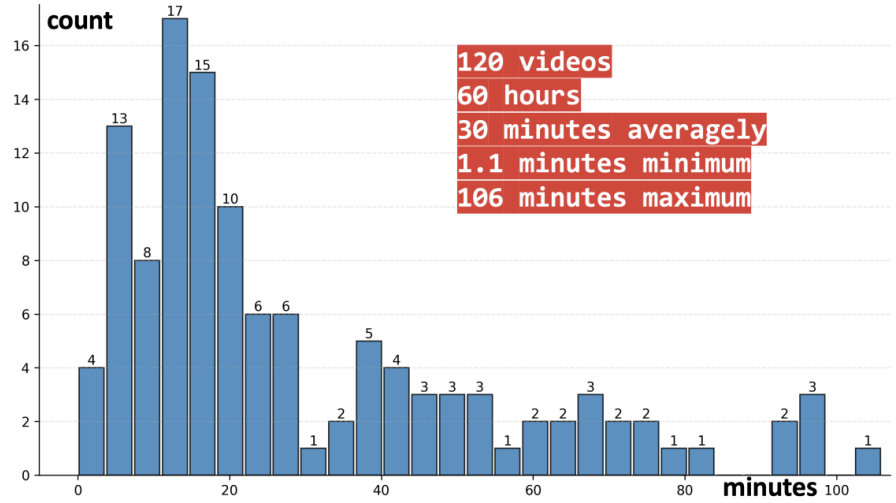
- **Level-1**: - Straightforward Reasoning, No complex logic; answers directly from the video
  - Only input video segments and queries, relying on the inherent capabilities of MLLMs.
- **Level-2**: - Simple Reasoning, Single-step spatio-temporal/causal reasoning
  - Input visual evidence ($\{Cv\}$), text evidence ($\{Cc,a,o\},\{Tc,a,o\}$), and queries.
- **Level-3**: - Hard Reasoning, Multi-hop cross-modal reasoning; requires deep semantic correlation
  - On the basis of Level-2, add graph evidence ($\{Cg\},\{Tg\}$).

$$R = \begin{cases} \text{MLLM}(\{C_n\}, Q) & \text{if } L \text{ is Level-1,} \\ \text{MLLM}(\{C_v\}, \{C_{c,a,o}\}, \{T_{c,a,o}\}, Q) & \text{if } L \text{ is Level-2,} \\ \text{MLLM}(\{C_v\}, \{C_{c,a,o}\}, \{T_{c,a,o}\}, \{C_g\}, \{T_g\}, Q) & \text{if } L \text{ is Level-3.} \end{cases}$$

HiVU: Hierarchical Video Understanding Benchmark

- 120 video samples, covering diverse scenarios such as movie clips, documentaries, lectures, and short videos, with a total duration of over 150 hours

Difficulty Level	Key Requirement	Example
L1 (Basic Perception)	Frame/clip-level basic perception, no reasoning required	"What animal appears at the 8th minute of the video?"
L2 (Temporal Reasoning)	Single-step spatio-temporal associative reasoning, requiring local temporal logic	"After the protagonist enters the room, what item does he/she touch first?"
L3 (Cross-Modal Causal Inference)	Multi-hop cross-modal causal reasoning, requiring global semantic association	"How does the color tone of the frame change when the background music rhythm speeds up? What impact does this change have on the audience's emotions?"



Result

☐ Improving open-sourced MLLMs with AdaVideoRAG on MLVU_test benchmark.

Model	Params	Frames	TR	AR	NQA	ER	PQA	SQA	AO	AC	TQA	AVG	Gain
GPT-4o	-	0.5fps	83.7	**68.8**	42.9	47.8	57.1	**63.6**	**46.2**	35	**48.7**	**54.9**	-
Video-LLaVA	7B	8	64.4	35.9	25.4	34	26	25	13.1	16.9	23.8	29.4	-
Video-LLaVA + AdaVideoRAG	7B	8	73.9	33.1	46.2	38	41.9	31.3	21.2	16.9	38.5	37.9	28.9%
Qwen2.5-VL	7B	2fps-768	46.7	15.4	16.9	35.8	38	38.9	24.6	13.6	31	29.0	
Qwen2.5-VL + AdaVideoRAG	7B	2fps-768	78.9	30.8	44.1	37.7	48	36.1	33.3	15.3	40.5	40.5	39.8%
Qwen2.5-VL	72B	2fps-768	73.3	33.3	59.3	47.2	40	41.7	37.7	16.9	26.2	41.7	
Qwen2.5-VL + AdaVideoRAG	72B	2fps-768	82.2	41	54.2	41.5	44	47.2	35.1	15.1	45.2	45.1	8%
VideoLLaMA3	7B	1fps-180	76.9	43.6	68.3	54.7	58	34.3	25	33.3	34.9	47.7	
VideoLLaMA3 + AdaVideoRAG	7B	1fps-180	**83.8**	47.1	**69.2**	**62.3**	**64**	38.9	34.8	**35.6**	42.9	53.2	11.6%

☐ Comparison with state-of-the-art VideoRAG on Video-MME dataset

Model	Params	Frames	Short	Medium	Long	Overall	Gain
GPT-4o	-	384	80	**70.3**	**65.3**	**71.9**	
Qwen2.5-VL	7B	2fps-768	55.6	47.1	38.8	47.2	
Qwen2.5-VL + VideoRAG [32]	7B	32	70.3	51.5	43.3	55.0	+7.9
Qwen2.5-VL + AdaVideoRAG	7B	2fps-768	72.8	59.1	47.7	59.9	+12.7
VideoLLaMA3	7B	1fps-180	76.7	62.8	53.2	64.2	
VideoLLaMA3 + VideoRAG [32]	7B	32	81.5	63.3	57.1	67.3	3.1
VideoLLaMA3 + AdaVideoRAG	7B	1fps-180	**80.3**	65.4	59.8	68.5	4.3

Result

Performance on HiVU

Metric	Level-2		Level-3		Overall		Level-2		Level-3		Overall	
	VideoLLaMA3	VideoLLaMA3 w/ AdaVideoRAG	VideoLLaMA3	VideoLLaMA3 w/ AdaVideoRAG	VideoLLaMA3	VideoLLaMA3 w/ AdaVideoRAG	VideoLLaMA3 w/ VideoRAG [32]	VideoLLaMA3 w/ AdaVideoRAG	VideoLLaMA3 w/ VideoRAG [32]	VideoLLaMA3 w/ AdaVideoRAG	VideoLLaMA3 w/ VideoRAG [32]	VideoLLaMA3 w/ AdaVideoRAG
Comprehensiveness	42.72%	57.28%	26.00%	74%	35.98%	64.02%	47.21%	52.79%	41.23%	58.77%	45.33%	54.67%
Empowerment	36.81%	63.19%	25.11%	74.89%	30.88%	69.12%	45.18%	54.82%	42.95%	57.05%	43.81%	56.19%
Trustworthiness	36.81%	63.19%	26.45%	73.55%	30.58%	69.42%	48.1%	51.9%	43.87%	56.13%	46.4%	53.6%
Depth	34.09%	65.91%	22.87%	77.13%	26.23%	73.77%	43.98%	56.02%	40.53%	59.47%	40.88%	59.12%
Density	38.63%	61.37%	25.11%	74.89%	31.03%	68.97%	46.36%	53.64%	42.56%	57.44%	44.1%	55.9%
Overall Winner	37.27%	62.73%	22.87%	77.13%	30.58%	69.42%	46.17%	53.83%	42.23%	57.77%	44.1%	55.9%

Ablation on graph-based knowledge retrieval, vision-based embedding retrieval and auxiliary text retrieval components

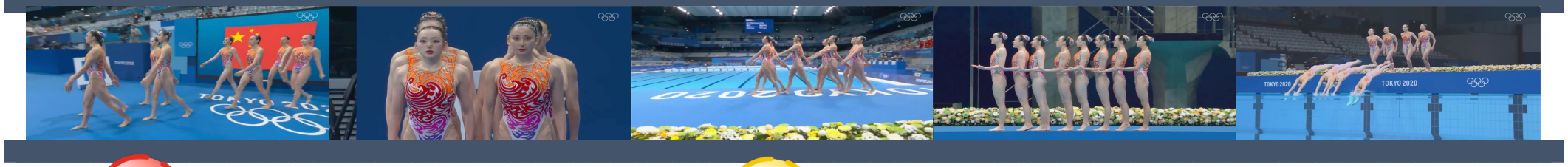
Metric	w/o Graph	All	w/o Vision	All	w/o Text	All
Comprehensiveness	38.92%	61.08%	50.13%	49.87%	33.17%	66.83%
Empowerment	47.79%	52.21%	48.42%	51.58%	40.53%	59.47%
Trustworthiness	47.79%	52.21%	46.31%	53.69%	39.79%	60.21%
Depth	46.31%	53.69%	49.47%	50.53%	30.33%	69.67%
Density	51.73%	48.27%	46.84%	53.16%	35.36%	64.64%
Overall Winner	45.82%	54.18%	48.23%	51.77%	31.25%	68.75%

Result

multiple-choice questions

Query: Which technique is primarily used when editing the landscape photo in the video?

(A) 7 (B) 3 (C) 8 (D) 2 (E) 5 (F) 4

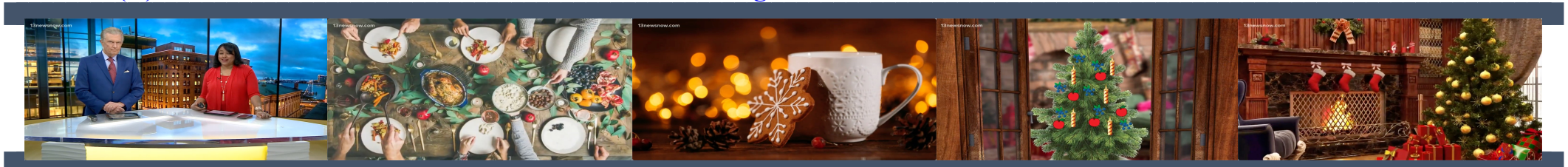


MLLM(VideoLLaMA): F

MLLM(VideoLLaMA) with AdaVideoRAG: C

Query: What is the genre of this video?

(A) It is a news report that introduces the history behind Christmas decorations.
(B) It is a documentary on the evolution of Christmas holiday recipes
(C) It is a travel vlog exploring Christmas markets around the world
(D) It is a tutorial on DIY Christmas ornament crafting



MLLM(VideoLLaMA) with VideoRAG: C

MLLM(VideoLLaMA) with AdaVideoRAG: A

Result

Sensemaking questions

Query: How does Jean's makeup routine reflect her personal style and preferences?

MLLM(VideoLLaMA) : Jean's makeup routine reflects her personal style and preferences through her choice of products, application techniques, and overall aesthetic. She opts for a natural yet polished look, using a mix of drugstore and high-end products. Her preference for a minimalist approach is evident in her simple eye makeup and the use of a single lip color.

MLLM(VideoLLaMA) with AdaVideoRAG : Jean's makeup routine reflects her personal style and preferences through her emphasis on a 'Clean Girl' look, which focuses on creating a dewy, hydrated appearance without heavy makeup. She prefers using products like the Mac lipstick and the Giorgio Armani luminous silk in the shade 575 for a natural yet polished finish. Her approach to skincare and makeup suggests she values a fresh and radiant look, often achieved through careful application techniques and product selection that enhances her natural features without appearing cakey or overdone.

Thanks

Github： https://github.com/xzc-zju/AdaVideoRAG