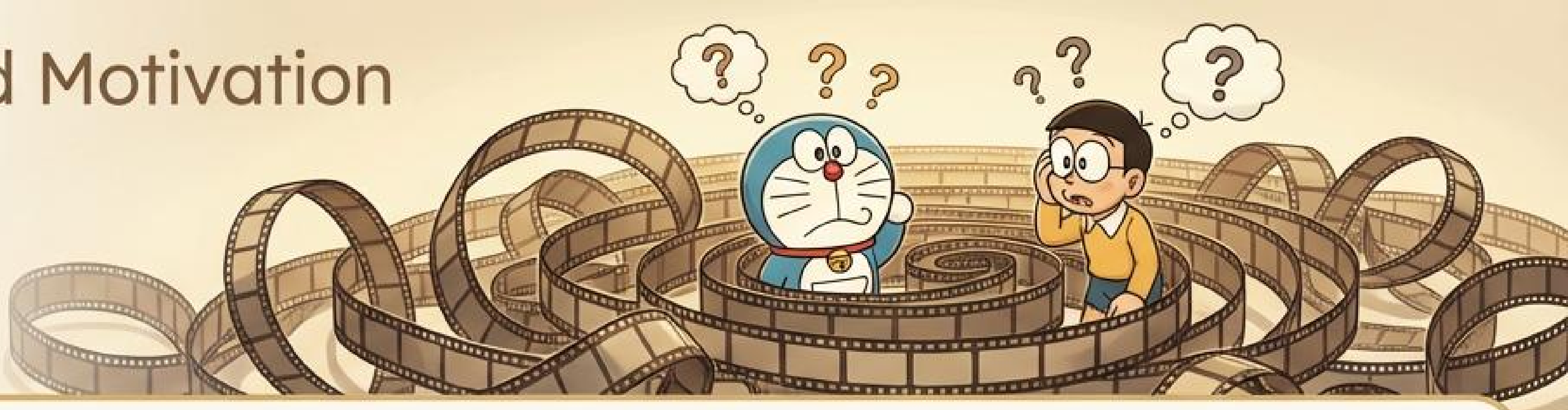
VideoRAG: Retrieval-Augmented Generation with Extreme Long-Context Videos

Xubin Ren, Lingrui Xu, Long Xia, Chao Huang from **The University of Hong Kong**;
Shuaiqiang Wang, Dawei Yin from Baidu Inc.

Research Problem and Motivation

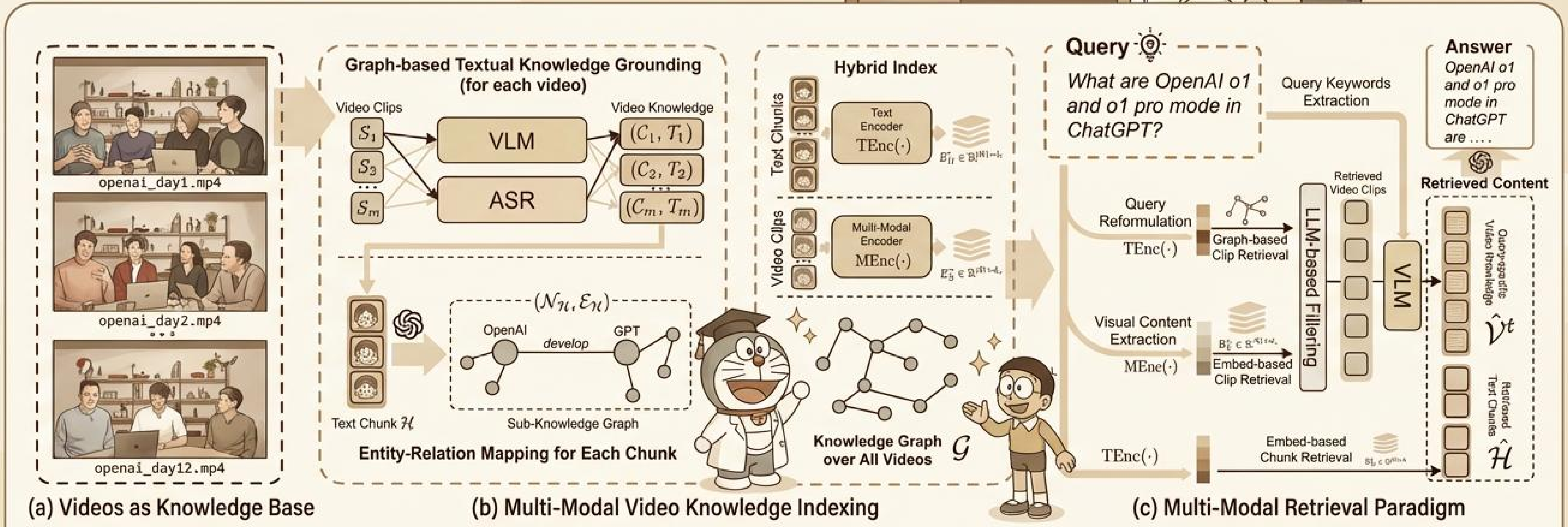


The paper addresses significant challenges in Retrieval-Augmented Generation (RAG) for understanding extremely long-context videos. VideoRAG introduces a novel framework to enhance large language models (LLMs) by integrating external knowledge through tailored retrieval mechanisms for video content.

Current models inadequately handle long-context videos, fail to integrate multi-modal data effectively, often losing contextual relevance, and rely on inefficient external tools for extraction, thus compromising retrieval precision and utility.

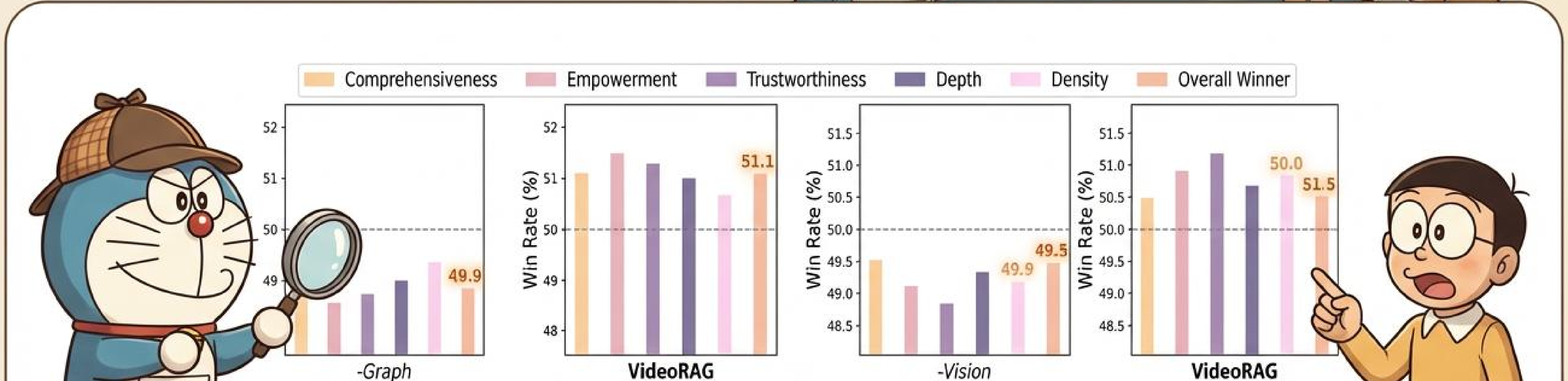
- **Inadequately handle long-context videos**
- **Fail to integrate multi-modal data effectively**
- **Losing contextual relevance**
- **Reliant on inefficient external tools**

VideoRAG Framework Overview



VideoRAG employs a dual-channel architecture combining Graph-based Textual Knowledge Grounding and Multi-Modal Context Encoding. Graph-based knowledge representation constructs a comprehensive graph capturing complex relationships across videos. Efficient retrieval is achieved via hybrid paradigms ensuring rapid and precise information extraction. Key formulas include: $\mathcal{C}_j = \mathrm{VLM}(\mathcal{T}_j, \{\mathbf{F}_1, \dots, \mathbf{F}_k\}|\mathbf{F} \in S_j$ $\boldsymbol{F} \in S_j)$ representing caption generation and $\hat{D} = \varphi(\mathcal{D}) = (\mathcal{G}, \mathbf{E}_H^t, \mathbf{E}_S^v)$ for structured output representation.

Method Components



(a) Comparison with the variant w/o the graph index-retrieval.

(b) Comparison with the variant w/o the visual modality.

VideoRAG utilizes a **graph-based grounding** to maintain semantic coherence, capturing multi-modal information information effectively.

The **multi-modal context encoding** captures visual and audio aspects preserving temporal dynamics.

This framework is evaluated using benchmark datasets such as **LongerVideos**, demonstrating superior performance against existing models including GraphRAG and LightRAG.

Experimental Results and Comparison

The LongerVideos dataset comprises **164 videos** totalling **134.6 hours** with **602** queries. VideoRAG demonstrates enhanced **comprehensiveness**, empowerment, trustworthiness, and depth compared to NaiveRAG, GraphRAG, and LightRAG. Ablation studies show the impact of graph-based and visual retrieval components. Comparisons with models such as LLaMA-VID and VideoAgent highlight VideoRAG's exceptional performance.

Method	Comprehensiveness	Empowerment	Trustworthiness
VideoRAG	52.34%	55.35%	54.49%
NaiveRAG	47.66%	44.65%	45.51%

- **Comprehensiveness** Effeccres graph comprehensiveness.
- **Empowerment** Demonstratαn: empowerment depthl.
- **Trustworthiness** Senurky empowerment and trustworthiness.
- **depth** Diving evauration of the depth.

Conclusion

- **Significant Advancements in Video Comprehension:** Effectively addresses limitations of existing models with a unique dual-channel architecture.
- **Integration of Multi-Modal Retrieval:** Successfully enhances retrieval speed and accuracy.
- **Superior Performance:** Quantitative comparisons and case studies underscore notable improvements in comprehensiveness, empowerment, and trustworthiness.