# VideoRAG: Retrieval-Augmented Generation over Video Corpus

Soyeong Jeong<sup>\*1</sup>, Kangsan Kim<sup>\*1</sup>, Jinheon Baek<sup>\*1</sup>, Sung Ju Hwang<sup>1,2</sup>

KA ST

KAIST<sup>1</sup>, DeepAuto.ai<sup>2</sup> {starsuzi, kksan07, jinheon.baek, sungju.hwang}@kaist.ac.kr



### Motivation: Existing RAG Systems Miss a Critical Modality — Videos

While Retrieval-Augmented Generation (RAG) has made significant progress by integrating textual and image content, it still largely overlooks videos, a modality rich in temporal and contextual information.

#### (A) Textual RAG

**Query:** After crossing the wide end, what's next in tying a tie?

Necktie



## **Our VideoRAG Bridges This Gap**

(C) VideoRAG (Ours)

A **necktie**, or simply a **tie**, is a piece of cloth worn for decorative purposes around the neck, resting under the shirt collar and knotted at the throat, and often draped down

(X) Answer: The necktie spread from Europe traces back to Croatian mercenaries serving in France during the Thirty Years' War.

#### (B) Conventional Image-Text RAG

Frame Selection

**Query:** After crossing the wide end, what's next in tying a tie?



Necktie

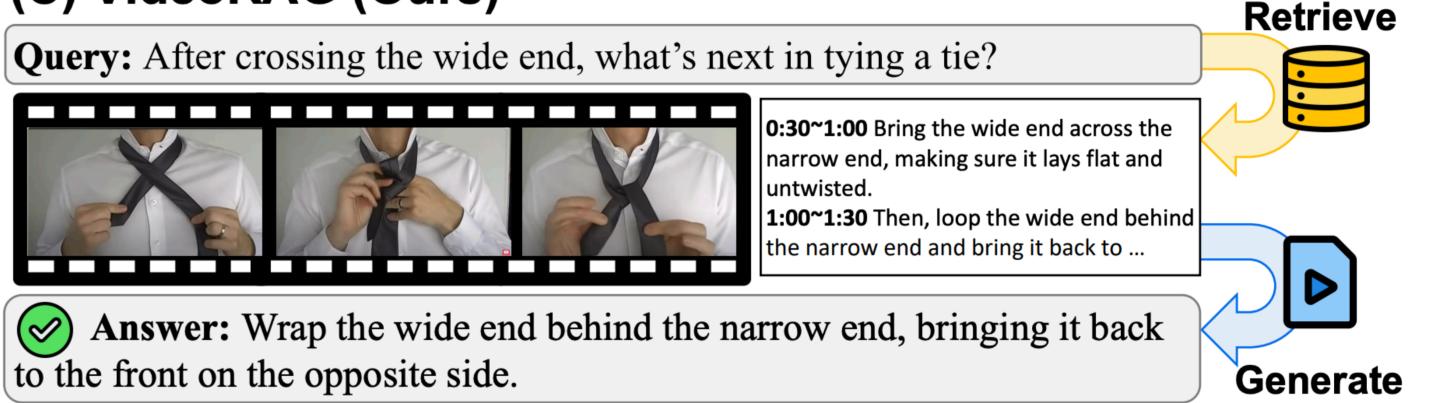
A necktie, or simply a tie, is a piece of cloth worn for decorative purposes around the neck, resting under the shirt collar and knotted at the throat, and often draped down

(X) Answer: Neckties are traditionally worn with the top shirt button fastened, and the tie knot resting between the collar points.



Generate

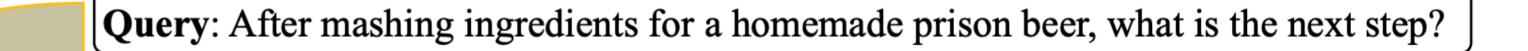
Retrieve

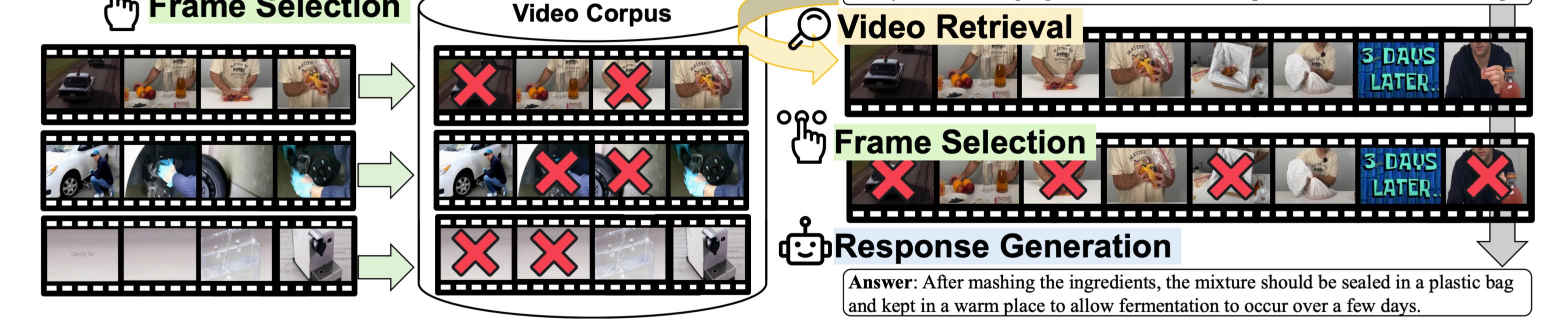


>VideoRAG retrieves and integrates both visual and textual cues from videos to generate more accurate and context-aware responses.

### **Approach: VideoRAG with Adaptive Frame Selection**

We propose VideoRAG, a novel framework that retrieves query-relevant videos from a large corpus and adaptively selects the most informative frames for both retrieval and generation.

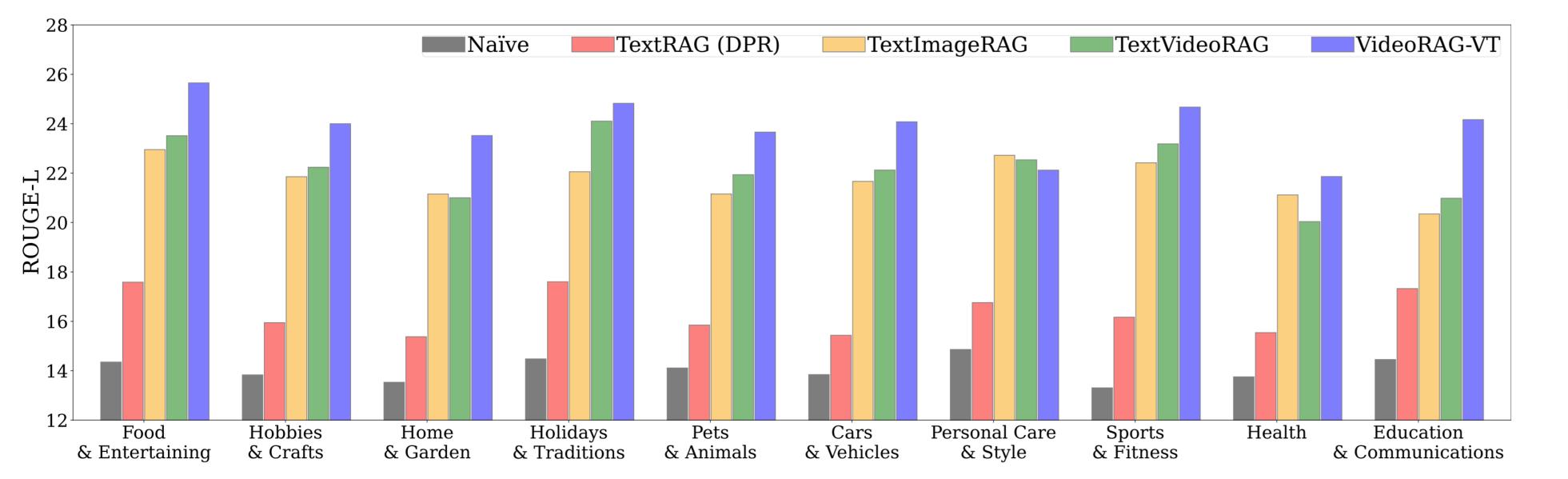




### **Results: VideoRAG Outperforms Text- and Image-Based RAG**

VideoRAG outperforms prior RAG baselines, highlighting the value of videos as external knowledge. Adaptive frame selection further improves retrieval and generation by focusing on informative segments.

		WikiHowQA with HowTo100M				Synthetic QA with HowTo100M				Retrieval		<b>R@1</b>	R@5	R@10
	Methods	<b>ROUGE-L</b>	<b>BLEU-4</b>	BERTScore	<b>G-Eval</b>	<b>ROUGE-L</b>	<b>BLEU-4</b>	BERTScore	<b>G-Eval</b>	al	Uniform	0.054	0.193	0.288
LLaVA-Video (7B)	NAÏVE Text <b>d</b> AC (DM25)	14.08	1.352 2.327	83.43 84.66	1.579 1.633	10.68	1.574	84.51 86.03	1.634	Ens. Visu	Adaptive (Ours)	0.079	0.249	0.367
	TEXTRAG (BM25) TEXTRAG (DPR) TEXTIMAGERAG TEXTVIDEORAG	17.22 16.65 22.43 22.81	2.327 2.173 4.222 4.388	84.00 84.61 86.88 86.97	1.033 1.591 2.022 1.979	14.70 14.58 25.19 23.41	2.382 2.397 6.149 5.435	80.03 85.85 88.56 88.40	1.681 1.686 2.175 2.278		Uniform Adaptive (Ours)	0.097 <b>0.118</b>	0.305 <b>0.324</b>	0.448 <b>0.453</b>
	VIDEORAG-V VIDEORAG-VT	<b>22.01</b> <b>24.95</b> 24.93	<u>5.080</u> <b>5.276</b>	<u>80.97</u> <u>87.85</u> <b>87.92</b>	$\frac{2.140}{2.142}$	<u>29.38</u> <b>29.74</b>	<u>7.530</u> <b>8.043</b>	<b>89.77</b> 89.72	2.278 =		Generation	<b>ROUGE-L</b>	BLEU-4	BERTScore
	ORACLE-V ORACLE-VT	26.19 25.37	5.480 5.237	88.41 87.95	2.225 2.166	32.16 32.31	8.769 8.885	90.34 90.46	2.884 2.938		Uniform Adaptive (Ours)	21.04 <b>23.24</b>	3.249 <b>3.963</b>	86.07 <b>87.13</b>



### Implications

> Develop and release a benchmark dataset for video-based RAG.

> Design more advanced frame selection strategies for better efficacy.