VideoRAG: Retrieval-Augmented Generation over Video Corpus

Soyeong Jeong*[1], Kangsan Kim*[1], Jinheon Baek*[1], Sung Ju Hwang[1,2]

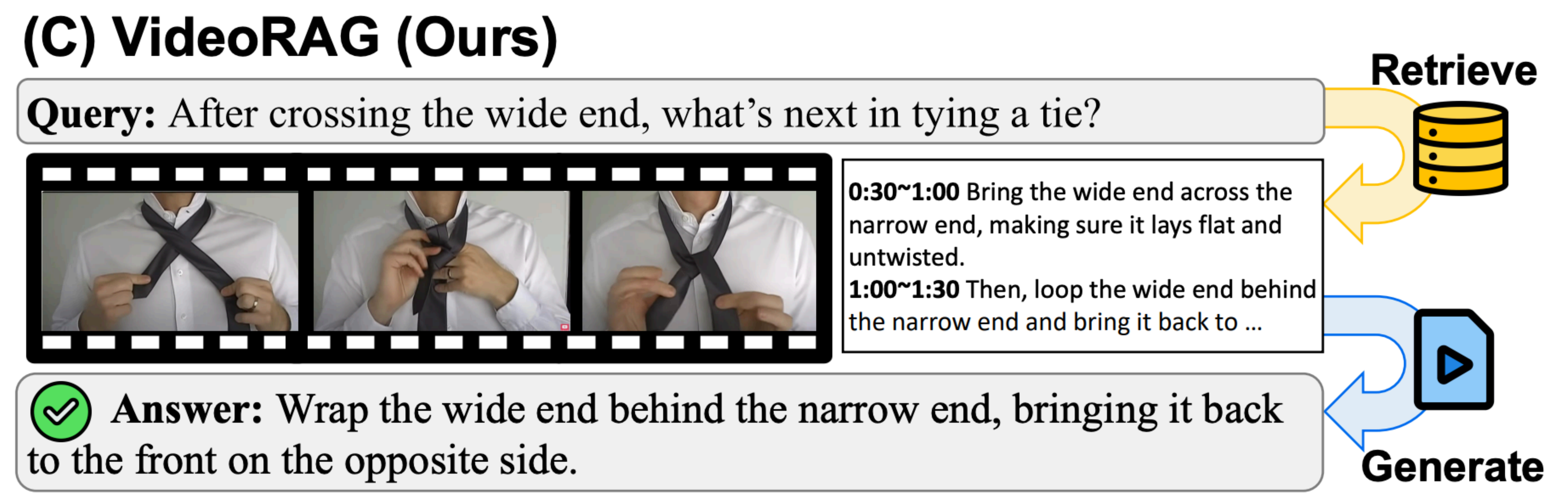
KAIST[1], DeepAuto.ai[2]

{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.

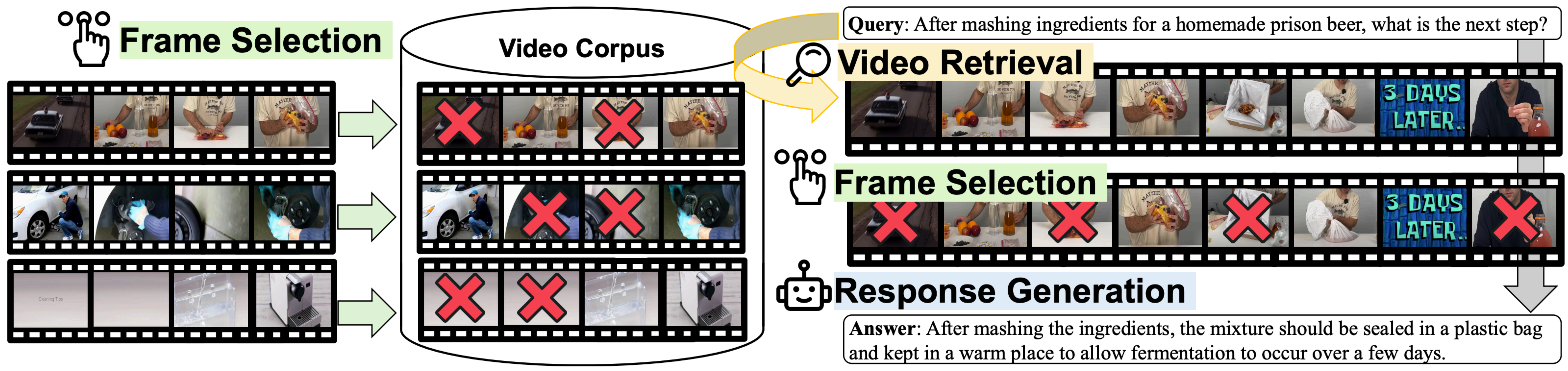
Our VideoRAG Bridges This Gap



➢ **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.

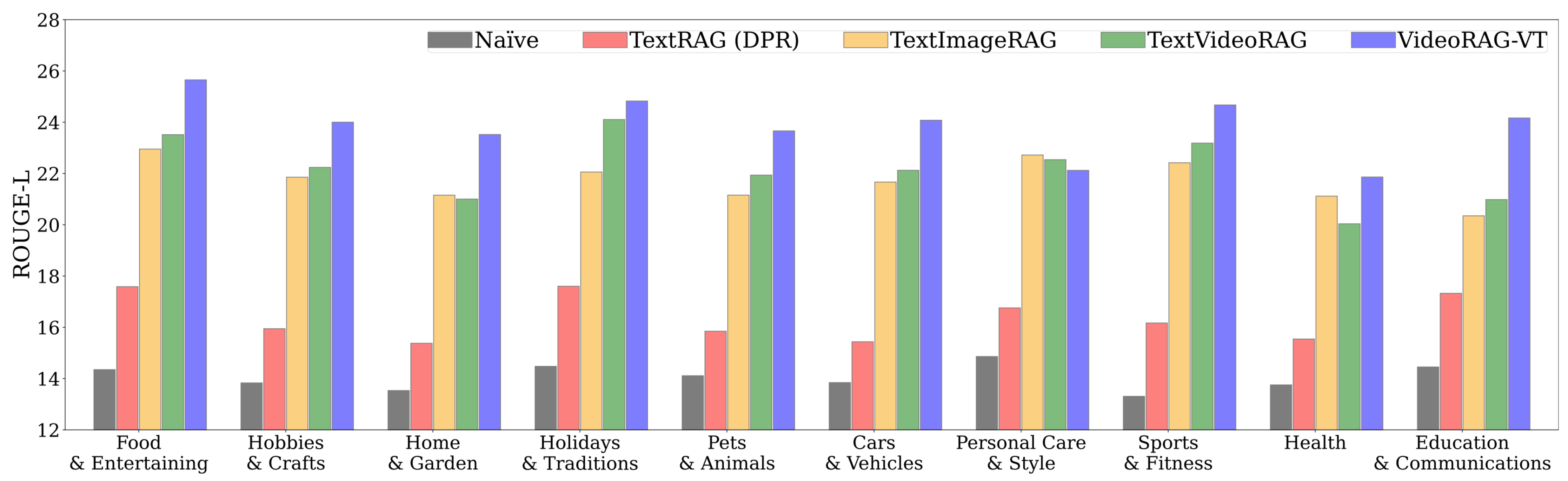


Results: VideoRAG Outperforms Text- and Image-Based RAG

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.

	Methods	WikiHowQA with HowTo100M				Synthetic QA with HowTo100M			
		ROUGE-L	BLEU-4	BERTScore	G-Eval	ROUGE-L	BLEU-4	BERTScore	G-Eval
LLaVA-Video (7B)	Naïve	14.08	1.352	83.43	1.579	10.68	1.574	84.51	1.634
	TextRAG (BM25)	17.22	2.327	84.66	1.633	14.70	2.382	86.03	1.681
	TextRAG (DPR)	16.65	2.173	84.61	1.591	14.58	2.397	85.85	1.686
	TextImageRAG	22.43	4.222	86.88	2.022	25.19	6.149	88.56	2.175
	TextVideoRAG	22.81	4.388	86.97	1.979	23.41	5.435	88.40	2.278
	VideoRAG-V	**24.95**	5.080	87.85	2.140	29.38	7.530	**89.77**	**2.479**
	VideoRAG-VT	24.93	**5.276**	**87.92**	**2.142**	**29.74**	**8.043**	89.72	2.476
	Oracle-V	26.19	5.480	88.41	2.225	32.16	8.769	90.34	2.884
	Oracle-VT	25.37	5.237	87.95	2.166	32.31	8.885	90.46	2.938

	Retrieval	R@1	R@5	R@10
Visual	Uniform	0.054	0.193	0.288
	Adaptive (Ours)	**0.079**	**0.249**	**0.367**
Ens.	Uniform	0.097	0.305	0.448
	Adaptive (Ours)	**0.118**	**0.324**	**0.453**
	Generation	**ROUGE-L**	**BLEU-4**	**BERTScore**
	Uniform	21.04	3.249	86.07
	Adaptive (Ours)	**23.24**	**3.963**	**87.13**



Implications

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

➢ Design more advanced frame selection strategies for better efficacy.