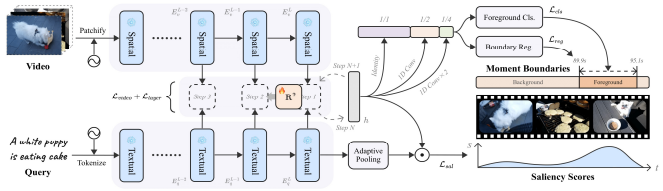
Video Temporal Grounding with Context-aware Text Query

Jimin Han, Dong-Jin Kim
School of Data Science, Hanyang University

Introduction

Ours : VTG with Context-aware Text Query

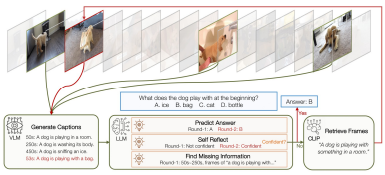
Generating Context-aware Text Queries using MLLMs/LLMs
Using the queries at inference without additional training
Enhancing performance by emphasizing key frames



R^2- Tuning

Video Temporal Grounding (VTG)

- Finding segments of video corresponding to a given text query
- Given text query is short and concise



VideoAgent

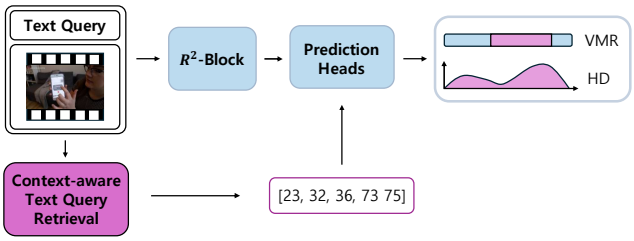
LLaVA-NeXT

LLM / MLLM based Video Understanding

- Leveraging LLM capabilities for video understanding
- MLLM / LLM used to retrieve **missing or crucial information**
- MLLM integrating multiple modalities for reasoning

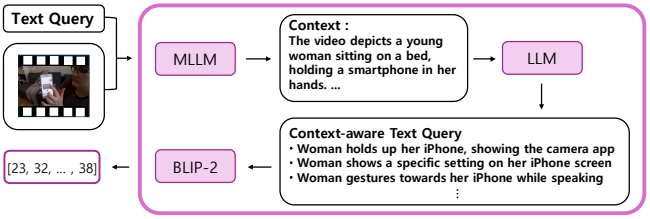
Framework

Model Overview



- Based on **context** of video, **generating text queries in detail**
- **Retrieve frames related to queries**, and use them at **test-time**

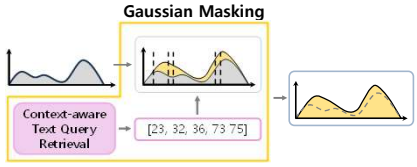
Context-aware Text Query Retrieval



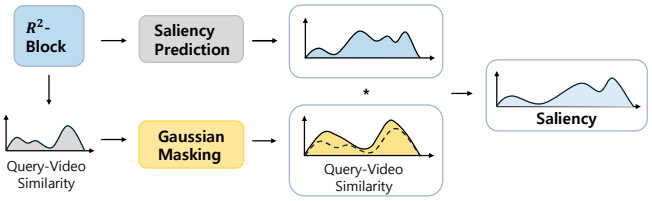
- **MLLM** extracts the overall context of a video
- **LLM** generates 5 context-aware text queries based on context
- **BLIP-2 retrieves 5 frames** based on the generated queries

Inference

- Apply Gaussian masking tailored to the two tasks

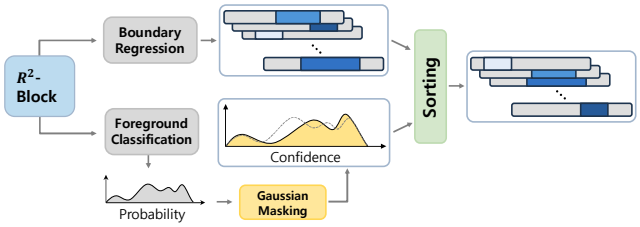


1) Highlight Detection



- **Frame-level saliency prediction**
- Emphasis on information around retrieved frames

2) Video Moment Retrieval



- Predict various boundaries, and **sort them by confidence**
- Emphasis on confidence around retrieved frames

Experiments

- **Overall performance gains across both tasks**

Highlight Detection Results

Model	HIT@1			mAP		
	Fair	Good	Very Good	Fair	Good	Very Good
R^2-Tuning	79.55	77.55	64.13	75.33	64.32	39.45
w/ Context-aware Text Query	**81.74**	**79.55**	**66.52**	**77.76**	**66.18**	**40.48**

Benchmark: QVHIGHLIGHTS

Video Moment Retrieval Results

Model	Recall@1			mAP		
	@0.3	@0.5	@0.7	@0.5	@0.75	Avg*
R^2-Tuning	78.71	67.74	51.87	68.54	49.94	47.85
w/ Context-aware Text Query	**79.81**	**68.39**	**52.52**	**68.82**	**50.08**	**48.07**

Avg* : Average from [0.5, 0.95]

Benchmark: QVHIGHLIGHTS

Conclusion

- **Propose a Context-aware Text-Query approach at VTG**
- **Generated queries capture more diverse, detailed meaning**
- **Observe consistent improvements within two VTG tasks**