NeurIPS 2026 | ILLINOIS | Microsoft Research | Google DeepMind

ReToken

Improving Long-Context VLMs with Visual Retrieval Token

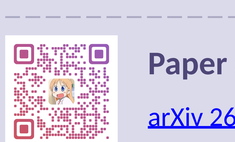
Yao Xiao[1], Reuben Tan[2], Zhen Zhu[1,3]*, Yuqun Wu[1], Jianfeng Gao[2], Derek Hoiem[1]

[1] University of Illinois Urbana-Champaign [2] Microsoft Research [3] Google DeepMind

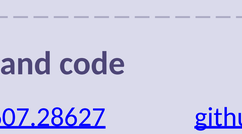
* Work completed at UIUC.

The value feature delivers a better retrieval signal than the key feature

"For the image with **a cow**, is there a truck?"



0 Irrelevant image



1 Relevant image

A target phrase extracts the retrieval cue from the full question. Here, it is "a cow".

A single learnable token retrieves well using the VLM's own representations

1 H100
for both training and long-video inference

5 hours
to train ReToken

>20% gain
in relative accuracy Visual Haystacks

Answer accuracy gain

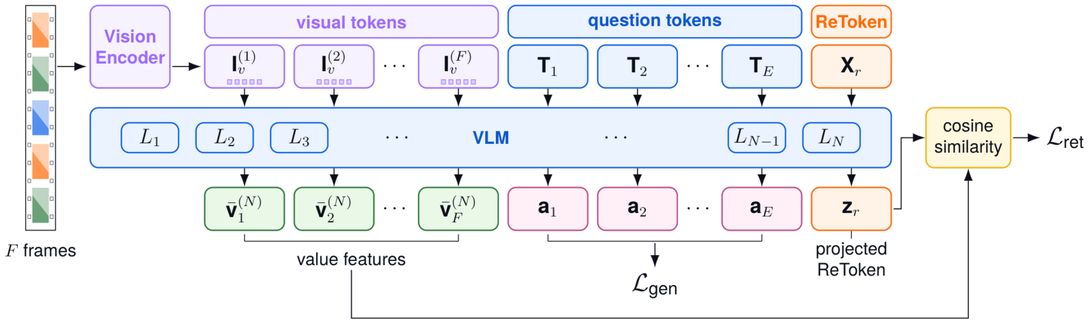
LVBench (zero-shot)	Qwen3VL-8B	+8.0
Visual Haystacks	Qwen3VL-8B	+13.4
Visual Haystacks	InternVL3.5-8B	+12.4

Paper and code

arXiv 2607.28627

github.com/avaxiao/ReToken

How ReToken learns to find evidence



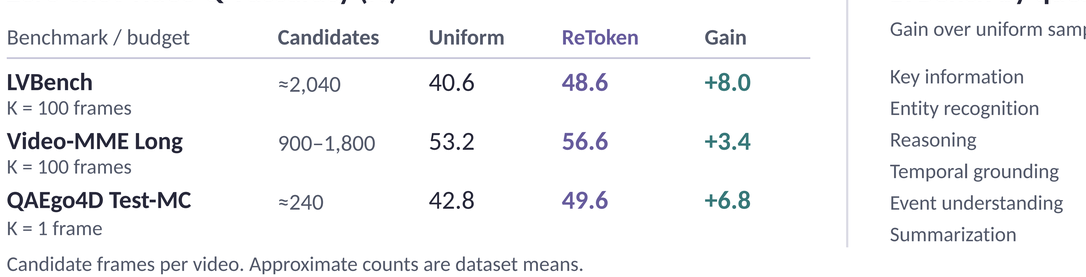
One token to train

Keep the VLM frozen. Train one retrieval token and its final-layer projection.

Image-relevance training

Train the token to rank images labeled relevant to the question above those labeled irrelevant.

Inference pipeline



Cache the video once

Store frame features in the model's visual KV cache.

Retrieve, then answer

Rank frames for the question. Generate the answer from the top-ranked frames' cache.

Video QA example

Sampled video frames

LVBench · Arsenal vs. Liverpool



Question How do Liverpool's players celebrate after the own goal?

Reference answer They hug each other.

Image-trained ReToken transfers to video

Frozen Qwen3VL-8B / cached video at 0.5 frames per second

Zero-shot video QA accuracy (%)

Benchmark / budget	Candidates	Uniform	ReToken	Gain
LVBench K = 100 frames	≈2,040	40.6	48.6	+8.0
Video-MME Long K = 100 frames	900–1,800	53.2	56.6	+3.4
QAEgo4D Test-MC K = 1 frame	≈240	42.8	49.6	+6.8

Candidate frames per video. Approximate counts are dataset means.

LVBench by question type

Gain over uniform sampling (accuracy)