FAR

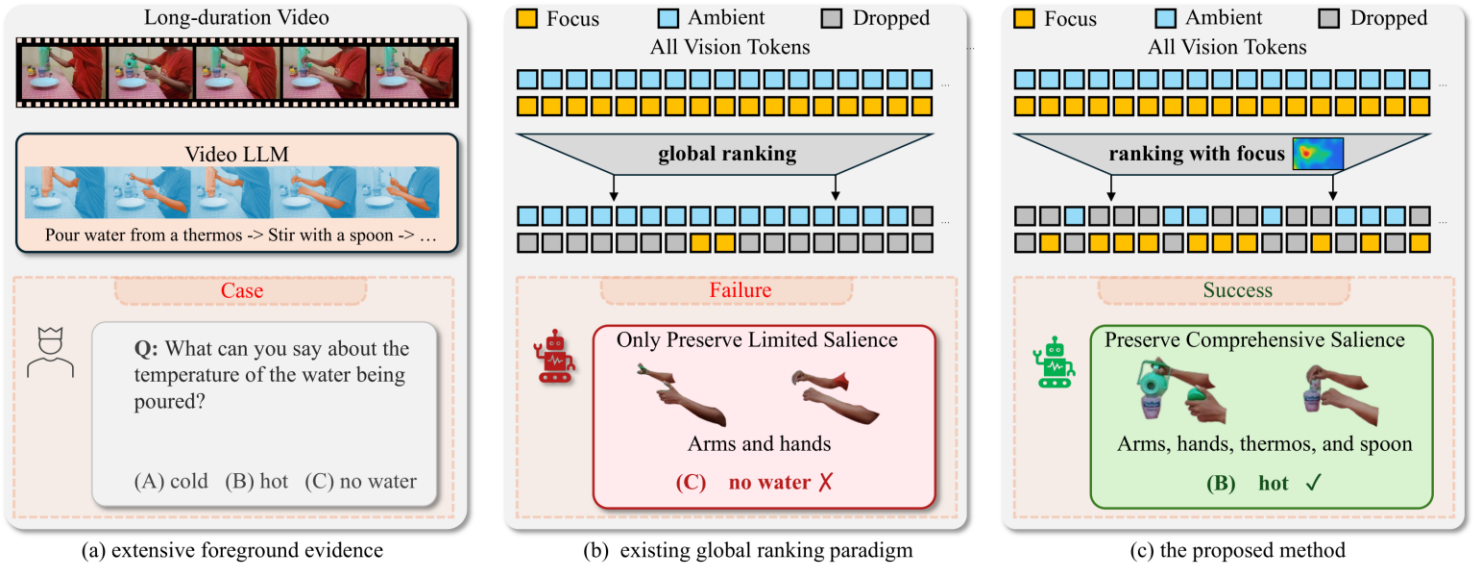
Training-free Focus-Ambient Retention for Memory-Efficient Video Large Language Models

Haozheng Zeng, Qi Bi, Gui-Song Xia

Wuhan University / University of Amsterdam

Training-free visual token retention before LLM prefill

Where long-video inference consumes memory



(a) extensive foreground evidence (b) existing global ranking paradigm (c) the proposed method

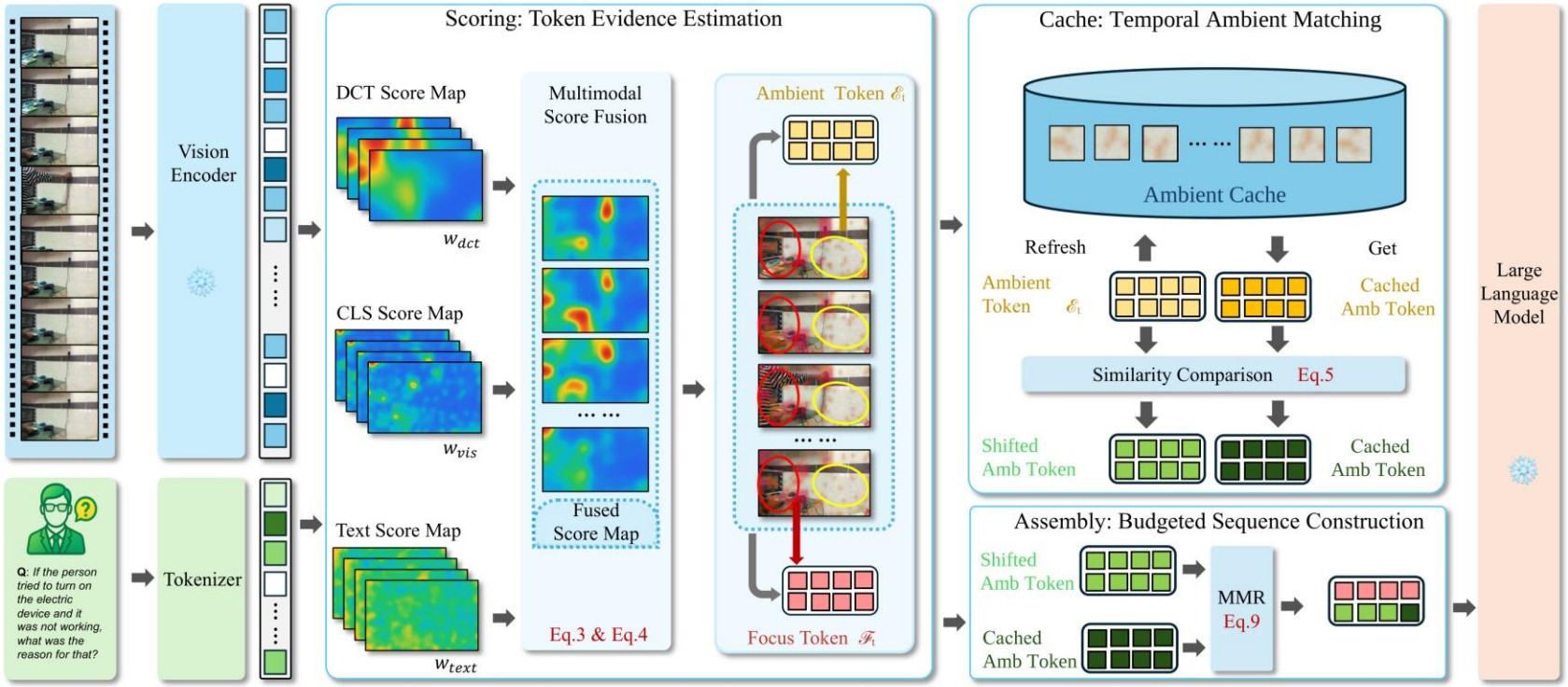
Two sources of memory pressure

Auxiliary attention tensors

Long LLM sequences and their KV entries

Memory-efficient execution and evidence selection work together

The FAR pipeline



Frame-wise scoring

Lower temporary residency

Before-prefill retention

Shorter downstream sequence

Focus and Ambient

Preserve useful evidence

Memory mechanism 1: frame-wise attention

Auxiliary attention workspace

$$\mathcal{M}_{\text{aux}}: \quad O(FHN^2) \longrightarrow O(HN^2)$$

Reduce each frame’s matrix immediately to a vector of token scores

Materialization	Peak VRAM (GiB)
Frame-wise streaming	**17.178**
Frame-wise batched	21.847
Naive video-wide	OOM

Same 64-frame input
Qwen2.5-VL-7B
H100, 94 GB

Auxiliary-matrix complexity only. This control is separate from the 50.6% cross-method result.

Memory mechanism 2: compression before prefill

$$S_{\text{dense}} = T + FN, \qquad S_{\text{FAR}} = T + \sum_{t=1}^{F} |\mathcal{I}_t|$$

T: text tokens. F: frames. N: visual tokens per frame. I_t: retained indices.

Shorter inputs reduce prefill activations and KV residency; model weights stay unchanged.

Sequence handling	Dense video input	FAR input
LLM input	All encoded visual tokens	Retained embeddings
Position tensors	Original coordinates	Gather matching positions
Downstream memory	Dense visual sequence	Reduced visual sequence

Preserving evidence under a smaller budget



Original paper QA example. Aggregate results appear on slides 8 and 9.

Ambient feature cache and LLM KV cache serve different roles

Focus

Preserve answer-critical detail

Ambient

Forward changed context

Assembly

Control redundancy within the budget

Evaluation across five video benchmarks

Benchmark	Coverage
VideoMME	Broad video understanding
EgoSchema	Egocentric reasoning
MVBench	Temporal understanding
LongVideoBench	Long-form reasoning
MLVU	Long-video comprehension

Backbones

Qwen2.5-VL
3B, 7B, 32B

GLM-4.6V-Flash-10B

Matched budgets

1%, 5%, 10%,
15%, 20%, 25%

Deterministic decoding; shared prompts and preprocessing within each backbone

Comparable scores with lower peak memory

Retention	FlashVID	FAR	Δ score
1%	47.667	46.590	-1.077
5%	53.965	53.572	-0.393
10%	55.551	55.194	-0.357
15%	56.346	56.532	+0.186
20%	56.905	56.454	-0.451
25%	**56.919**	**57.347**	**+0.428**

50.6% less

peak VRAM vs FlashVID

36.46 GiB

18.02 GiB

FlashVID, then FAR
Qwen2.5-VL-7B

Five-benchmark mean score. Peak VRAM averages dataset maxima, not mean residency.

Selection quality and component contributions

Pre-LLM comparison

R	FAR	DivPrune	PruneVid	DyCoKe
1%	46.590	47.226	46.366	40.071
5%	**53.572**	**51.492**	**51.629**	**42.682**
15%	**56.532**	**54.798**	**53.757**	**52.898**
25%	**57.347**	**56.036**	**54.744**	**53.593**

VideoMME ablation

Variant	25%	5%
Full FAR	**60.667**	**54.926**
CLS-only scoring	60.148	54.926
Without Cache	60.407	54.778
Without Assembly	60.185	54.556

Highest pre-LLM mean at 5–25% retention

Five benchmarks; controlled adapters; DyCoKe uses TTM only

Modules help overall; effects vary by budget

The remaining peak comes from vision encoding

Retention	Vision	LLM prefill	Decode
1%	16.710	14.411	14.392
10%	16.709	14.529	14.431
25%	16.709	14.727	14.496

GiB. Qwen2.5-VL-7B, MLVU, 50 samples per ratio. Values are mean per-sample phase peaks.

Vision sets the peak

Lower retention can reduce prefill residency while the overall peak stays nearly constant.

Scope and next steps

Position compatibility and compression latency remain constraints.