LOOK AGAIN: ATTENTION-GUIDED TEMPORAL RESAMPLING FOR VIDEOLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Uniform frame sampling can miss the brief events needed to answer questions about long videos. We investigate whether a VideoLLM's internal attention can guide a second look at relevant moments, even when the initial view does not support a correct answer. We introduce Temporal Attention Resampling (TAR), a training-free method that contrasts question-conditioned attention with a generic reference, retrieves a contiguous temporal interval, and combines densely resampled local frames with global context. On the MLVU evaluation set, TAR improves Qwen3-VL-8B from 44.42% to 50.20%, outperforming matched 48-frame uniform sampling by 4.38 percentage points and random-window resampling by 4.65 percentage points. Accuracy improves across eight benchmarks and two Qwen-family model scales. On the annotated LVBench subset, questions whose retrieved windows overlap evidence gain 6.43 points, compared with 1.19 points for those without overlap. This association is consistent with improved temporal evidence acquisition. TAR requires no parameter updates, temporal supervision, or external retriever.

1 INTRODUCTION

A brief event can determine the answer to a question about a long video (Wu et al., 2024; Zhou et al., 2025; Wang et al., 2025b; Fu et al., 2025). A decisive action, object state, or interaction may occupy only a small fraction of the timeline. Uniform frame sampling allocates the same visual capacity to every part of that timeline, regardless of the question. Under a limited frame budget, it can therefore miss the relevant event or capture it too sparsely to resolve the answer. Increasing the budget broadens coverage, but also adds observations that may be irrelevant to the question.

This limitation concerns evidence acquisition: the model must first observe sufficient visual evidence before it can reason about it. A wrong answer does not necessarily imply a failure of semantic understanding; the initial observation may simply provide insufficient evidence. This distinction motivates a coarse-to-fine strategy: process the whole video, identify a promising interval, and revisit that interval at higher temporal resolution. Our central question is whether the answering VideoLLM itself provides a useful signal for *when to look again*.

Prior work on images shows that internal multimodal signals can guide a second look at question-relevant spatial regions (Khayatkhoei et al., 2025). Extending this principle to video requires both temporal coherence and global context. An action may unfold across neighboring frames, while its interpretation may depend on events elsewhere in the video. Moreover, raw attention can reflect generic visual saliency rather than the evidence needed for a particular question. A useful retrieval signal must distinguish question-specific emphasis from this shared response.

We address these challenges with Temporal Attention Resampling (TAR). Starting from a coarse uniform view, TAR compares attention elicited by the question with attention elicited by a generic video prompt. It aggregates their ratio over spatial locations, retrieves the highest-scoring contiguous temporal window, and samples new frames directly from that interval in the source video. The final answering view combines 32 global frames with 16 local frames, retaining broad coverage while increasing detail in the selected region. Retrieval uses only the question; answer candidates enter the final answering prompt. No model parameters are updated, and inference requires neither temporal annotations nor an external retriever.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Our experiments first establish the potential of temporal selection, then test whether attention can realize part of that potential without privileged information. On a 200-question diagnostic subset, an answer-informed temporal oracle raises accuracy from 35% to 65%. On the 502-question MLVU evaluation set, TAR improves Qwen3-VL-8B from 44.42% to 50.20%. It also outperforms uniform sampling with the same final frame count and random-window resampling with the same global–local allocation. Gains extend to all 16 combinations of eight benchmarks and two Qwen-family model scales. On LVBench, larger gains accompany retrieved windows that overlap annotated evidence, linking the accuracy results to temporal evidence acquisition while leaving the causal mechanism open.

Our contributions are threefold:

- We quantify the headroom for answer-informed temporal selection on a diagnostic subset, motivating a second look at selected video intervals.
- We introduce a training-free procedure that converts question-specific relative attention into a contiguous-window retrieval decision and a global–local answering view.
- We evaluate this procedure across eight benchmarks and two model scales, using matched final-frame controls, attention ablations, and localization diagnostics to assess its benefits and limits.

2 Related Work

2.1 Long-video understanding and adaptive observation

VideoLLMs accommodate longer visual contexts through advances in compression, training, and inference (Li et al., 2026; Bai et al., 2025; Chen et al., 2026; Qiu et al., 2026). Benchmarks including MLVU, Video-MME, LongVideoBench, and LVBench assess whether these models can use evidence distributed across long videos (Zhou et al., 2025; Fu et al., 2025; Wu et al., 2024; Wang et al., 2025b). Under a limited visual budget, broad temporal coverage must be balanced against detailed observation of relevant events.

Adaptive input construction addresses this trade-off in several ways. LongVU compresses spatiotemporal redundancy (Shen et al., 2024), and Adaptive Keyframe Sampling balances relevance with coverage (Tang et al., 2025). Other methods learn frame selectors or perform training-free selection based on relevance, diversity, and temporal structure (Hu et al., 2025; Huang et al., 2025; Sun et al., 2025). TAR focuses on temporal resampling: it selects a coherent interval and acquires denser observations from the source video, supplementing a global view.

A related line of work explicitly searches for evidence. VideoTree progressively refines a query-adaptive video representation (Wang et al., 2025c); T* reformulates temporal search as a spatial problem (Ye et al., 2025); A.I.R. integrates reasoning into iterative frame selection (Zou et al., 2026); and Deep Video Discovery searches a multi-granular video database (Zhang et al., 2026). Factor-guided coarse-to-fine reasoning likewise emphasizes evidence acquisition (Xu et al., 2026). Within this broader direction, TAR studies a single contiguous-window decision derived from the answering model's own attention. Its empirical comparisons cover uniform sampling, random windows, and a CLIP-based retriever; they do not constitute a comprehensive comparison with adaptive search systems.

2.2 Training-free use of internal multimodal signals

MLLMs Know Where to Look uses internal signals for training-free spatial revisiting (Khayatkhoei et al., 2025). Attention-guided and representation-based decoding methods also exploit internal information to improve multimodal outputs (Wang et al., 2025a; Xu et al., 2025; Cho et al., 2025), while DyFo combines dynamic visual search with external visual experts for fine-grained perception (Li et al., 2025). Together, these studies motivate using model-internal information to guide further processing.

TAR applies this idea to temporal observation. The attention signal determines which part of the original video to sample more densely; the unmodified VideoLLM then answers from the resulting

global and local frames. Its specific design combines a generic-attention reference, a contiguous retrieval window, and global context retention. The following section formalizes these choices, and the experiments test their contribution within a fixed resampling procedure.

3 Temporal Attention Resampling

TAR consists of three stages: extracting relative attention from a coarse video view, retrieving a contiguous temporal interval, and answering from global frames plus dense local observations (Fig. 1).

3.1 Problem formulation

Let V be a video of duration D, q a question, and $\mathcal{C} = \{c_1, \ldots, c_K\}$ the answer candidates. Let $U_N(V)$ sample N frames uniformly over the video and let F denote the VideoLLM. The uniform baseline predicts

$$\hat{c}_{\text{base}} = F(U_N(V), q, \mathcal{C}). \tag{1}$$

TAR first constructs a coarse view $X = U_{N_c}(V)$ to guide temporal retrieval. This view determines *where* to resample; the final answer uses a separately constructed visual input.

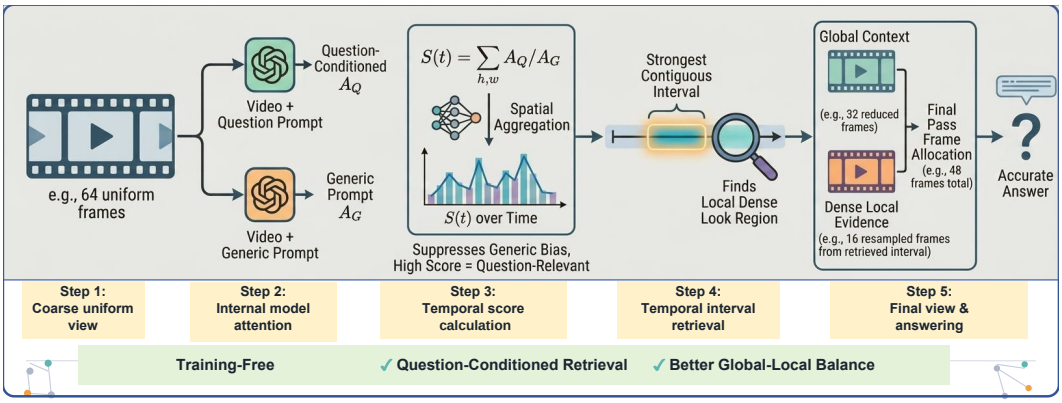


Figure 1: Relative attention retrieves a contiguous interval from a coarse video view. Dense local frames are combined with global context for final answering, without updating model parameters or adding an external retriever.

3.2 Question-specific relative attention

The retrieval prompt p_Q contains q but excludes the answer candidates. At transformer layer ℓ, we extract attention from retrieval-text tokens to visual tokens, average across heads, and map the visual-token positions to temporal–spatial coordinates. Denote the resulting question-conditioned map by $A_Q(t, h, w)$, where t indexes temporal bins and (h, w) indexes spatial locations. Processing the same coarse view with a generic prompt p_G yields a reference map $A_G(t, h, w)$. The generic prompt requests visible video content without referring to an answer hypothesis.

We use the ratio score reported in the attention ablation:

$$R(t, h, w) = \frac{A_Q(t, h, w)}{A_G(t, h, w) + \epsilon}, \qquad \epsilon = 10^{-4}. \tag{2}$$

The ratio emphasizes locations that receive more attention under the question relative to the generic prompt. Locations receiving similar attention under both prompts receive lower relative scores. This is a retrieval heuristic, not a calibrated probability of evidence relevance. Small reference values can yield large ratios, so values above the 99.5th percentile are clipped before normalization. Let $\widetilde{R}$ denote the resulting processed map. We aggregate it spatially:

$$S(t) = \sum_{h,w} \widetilde{R}(t, h, w). \tag{3}$$

This operation converts the spatial attention map into a one-dimensional temporal score. The next stage uses that score to select a coherent interval rather than treating each highly attended frame independently.

3.3 CONTIGUOUS TEMPORAL RETRIEVAL

Let T be the number of temporal attention bins. We choose a window spanning approximately one eighth of the coarse sequence:

$$L = \max\{1, \lfloor T/8 + 0.5 \rfloor\}, \qquad i^* \in \operatorname*{arg\,max}_{1 \le i \le T-L+1} \sum_{t=i}^{i+L-1} S(t). \tag{4}$$

Here, i^* is a maximizing start index and the retrieved bins are $W^* = \{i^*, \ldots, i^* + L - 1\}$. Their associated timestamps determine an interval $[a^*, b^*] \subseteq [0, D]$ in the source video. The interval is resampled directly from the video rather than by duplicating the coarse input frames.

Selecting a window rather than isolated high-scoring bins preserves a temporal neighborhood in which actions and state transitions can unfold. The window score also pools support across adjacent bins. This choice favors coherent local evidence, although a single fixed-width interval cannot cover multiple distant events. The selected window is therefore a proposal for further observation, not a guaranteed evidence interval.

3.4 GLOBAL–LOCAL ANSWERING

Temporal retrieval identifies where to increase sampling density, but the selected interval may omit context needed to interpret the event. We retain whole-video coverage by constructing

$$X_{\text{final}} = [U_{N_g}(V); U_{N_l}(V[a^*, b^*])], \qquad \hat{c}_{\text{TAR}} = F(X_{\text{final}}, q, \mathcal{C}), \tag{5}$$

where the semicolon denotes combining the global and local streams. Candidate answers are included only at this stage. The default configuration uses $N_c = 64$, $N_g = 32$, and $N_l = 16$, for 48 frames in the final answering view. One third of that view is allocated to the retrieved interval, while the global stream retains coverage of the full video.

The complete procedure requires question-conditioned and generic-reference attention extraction, followed by final answering. Both retrieval passes process the coarse video and require access to internal attention tensors. Consequently, the 48-frame final view describes the answering budget rather than total inference cost. Section 5 discusses the associated computational trade-offs.

4 EXPERIMENTS

Our evaluation addresses three questions: whether temporal selection offers useful headroom, whether TAR improves answering beyond changes in frame allocation, and whether its gains align with better temporal evidence acquisition. We first report overall results, then examine the retrieval signal, sampling budget, and localization behavior.

4.1 SETUP

Models and inputs. We evaluate Qwen3-VL-8B-Instruct (Bai et al., 2025) and Qwen3.6-35B-A3B. For the former, we extract attention from layer 22. For the latter, a mixture-of-experts model with hybrid attention, we use a full-attention layer (layer 28) and disable the thinking mode. All evaluations use video frames only, without audio or subtitles. The default retrieval view contains 64 frames and the final answering view contains 32 global and 16 local frames.

Evaluation sets. Our primary MLVU evaluation (Zhou et al., 2025) uses 502 multiple-choice questions across nine tasks; reported MLVU results refer to this evaluation set. We also evaluate 1,549 LVBench examples (Wang et al., 2025b), 1,337 labeled LongVideoBench validation examples (Wu et al., 2024), and 2,700 Video-MME questions (Fu et al., 2025). Additional evaluations cover NExT-QA (Xiao et al., 2021), EgoSchema-500 (Mangalam et al., 2023), MotionBench (Hong et al., 2025), and TOMATO (Shangguan et al., 2025). Together, these sets test long-video understanding, temporal reasoning, and motion understanding.

Comparisons and statistics. We report multiple-choice accuracy and changes in percentage points (pp). Uniform-64 is the default baseline. On MLVU, Uniform-48 matches the final frame count and Random-window matches the 32-global/16-local allocation. These controls match final answering inputs, not total computation. We use the two-sided exact McNemar test for paired accuracy comparisons (Appendix A) and summarize random-window variation over three deterministic seeds. Percentage-point differences are computed before rounding and may therefore differ slightly from subtraction of the displayed accuracies.

4.2 Oracle temporal headroom

Before evaluating attention-based retrieval, we ask how much temporal selection can help when the correct answer is available. On a 200-question diagnostic subset, we divide each video into eight intervals and select the interval assigning the highest probability to the correct answer. This answer-informed oracle raises accuracy from 35% under uniform sampling to 65% (Fig. 2). Adding oracle spatial focusing yields 66%. The large temporal gain motivates retrieving a better view, while the smaller spatial increment applies only to this diagnostic setup. Because the oracle uses the correct label, its result measures privileged selection headroom rather than deployable retrieval performance.

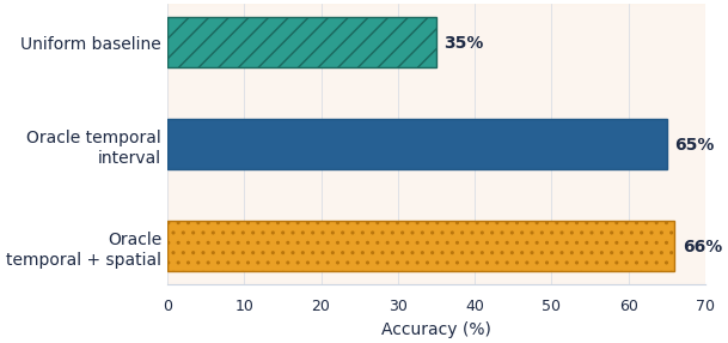


Figure 2: Oracle intervention on the 200-question diagnostic subset. Oracle temporal selection recovers 30 percentage points over uniform sampling, while additional oracle spatial focusing provides only a 1-point improvement.

Table 1: Main results. TAR retrieves from a 64-frame coarse view and answers from 32 global plus 16 local frames. Uniform answers directly from 64 frames. Δ denotes the accuracy change in percentage points. Both models use video-only inputs.

	Qwen3-VL-8B			Qwen3.6-35B-A3B		
Benchmark	Uniform	TAR	Δ	Uniform	TAR	Δ
MLVU	44.42	**50.20**	+5.78	50.20	**53.78**	+3.58
Video-MME	58.39	**60.73**	+2.33	64.11	**67.81**	+3.70
LVBench	34.60	**36.99**	+2.39	39.38	**43.06**	+3.68
LongVideoBench	55.27	**59.09**	+3.81	57.07	**61.63**	+4.56
NExT-QA	77.77	**80.37**	+2.60	81.07	**83.87**	+2.80
EgoSchema-500	67.60	**70.40**	+2.80	66.80	**69.60**	+2.80
MotionBench	51.91	**54.57**	+2.66	60.89	**63.79**	+2.90
TOMATO	29.77	**32.10**	+2.33	33.90	**37.53**	+3.63

4.3 Results across benchmarks and model scales

Table 1 compares TAR with uniform sampling across eight benchmarks. Accuracy improves in all 16 model–benchmark combinations. Qwen3-VL-8B gains 2.33–5.78 points, including an increase

from 44.42% to 50.20% on MLVU. Qwen3.6-35B-A3B gains 2.80–4.56 points, with MLVU accuracy increasing from 50.20% to 53.78%. On Video-MME, the latter model corrects 252 baseline errors while introducing 152 regressions, giving a net gain of 100 questions and a paired exact $p = 7.44 \times 10^{-7}$ (Appendix B).

The consistency of these gains suggests that temporal resampling is useful across the evaluated tasks and Qwen-family model scales. However, the comparison changes both the retrieval procedure and the final frame allocation. We next use matched controls on MLVU to determine whether the selected interval contributes beyond this allocation change.

Table 2: Controlled ablations on MLVU with Qwen3-VL-8B. Panels A and B use the default 64-frame coarse view when retrieval is required; Panel C uses the 32-frame setting. Random-window reports the mean and population standard deviation over three seeds. Compare variants within each panel.

Answering view or retrieval variant	Setting	Accuracy (%)
A. Does the selected interval matter?		
Uniform-64	64 final	44.42
Uniform-48	48 final	45.82
Random window (32G+16L)	48 final	45.55 ± 0.19
TAR (32G+16L)	48 final	**50.20**
B. Which attention signal is useful?		
Question A_Q	48 final	46.41
Generic A_G	48 final	46.61
Difference $[A_Q - A_G]_+$	48 final	47.41
Ratio $A_Q/(A_G + \epsilon)$	48 final	**50.20**
C. Does answer conditioning help? (32-frame setting)		
Uniform-32	32 initial	43.23
Question only (TAR)	32 initial	**47.21**
Question + top-two candidates	32 initial	46.61
Candidate-token score fusion	32 initial	44.42

4.4 Controlled retrieval ablations

Frame allocation. Table 2A separates retrieval from final-frame allocation on MLVU. Uniform-48 achieves 45.82%, slightly above Uniform-64 at 44.42%, but remains 4.38 points below TAR. Random-window uses the same 32-global/16-local structure and interval duration as TAR, yet reaches only $45.55 \pm 0.19\%$ across three seeds, compared with 50.20% for attention retrieval. Thus, a smaller uniform view and an arbitrary local view both fall short of attention-guided resampling. The comparison supports a contribution from interval selection under a matched final answering budget. The standard deviation describes variation among random-window seeds, not uncertainty in the paired advantage of TAR.

Attention signal. Having established that the interval matters, we compare the signals used to select it. Table 2B holds the coarse view, window length, and final allocation fixed. Question attention A_Q reaches 46.41%, while generic attention A_G reaches 46.61%; question conditioning alone therefore does not yield a stronger result in this comparison. The positive difference $[A_Q - A_G]_+$ reaches 47.41%, and the ratio in Eq. 2 reaches 50.20%. The ratio exceeds the difference by 2.79 points and Uniform-64 by 5.78 points. The strongest result comes from comparing question attention with its generic reference, rather than using either map alone. This finding supports the ratio as a retrieval score in the tested configuration; Section 4.7 separately examines agreement with annotated evidence.

Answer conditioning. We also test whether adding an initial answer hypothesis improves retrieval. Under the 32-frame setting, question-only retrieval achieves 47.21%, compared with 46.61% for a prompt incorporating the top-two predicted candidates and 44.42% for candidate-token score fusion (Table 2C). Uniform-32 reaches 43.23%. Neither variant outperforms question-only retrieval,

supporting the simpler prompt used by TAR. The comparison does not isolate the source of the decrease: initial prediction errors, prompt changes, and score aggregation may each contribute.

4.5 Sensitivity to frame budget and attention layer

We next examine whether the benefit depends on a particular sampling budget. Figure 3a varies the initial view from 32 to 128 frames. The corresponding final views contain 24, 48, and 96 frames. Gains over the uniform baseline are positive in all three settings: 3.98, 5.78, and 2.99 points. The 64-frame comparison has an exact McNemar $p = 0.0042$; the 32- and 128-frame comparisons have $p = 0.0569$ and $p = 0.1591$, respectively. The gains are positive across budgets, with the clearest paired evidence at 64 frames. Because each setting changes both the coarse view and the final answering budget, this experiment characterizes the combined configuration rather than either budget in isolation. The largest observed gain at 64 frames does not establish a universally optimal choice.

Effect of attention layer. Figure 3b varies the attention layer while holding the remaining settings fixed. Layers 5 and 11 reach 45.62% and 47.01%, whereas layers 17, 22, 28, and 34 reach 50.20–50.80%. Layers 28 and 34 tie at the highest reported accuracy. Useful retrieval signals therefore appear across several middle-to-late layers, and the default layer 22 is not uniquely optimal. These results describe sensitivity on the evaluated set; a fixed layer-selection protocol would be needed to assess generalization without benchmark-specific tuning.

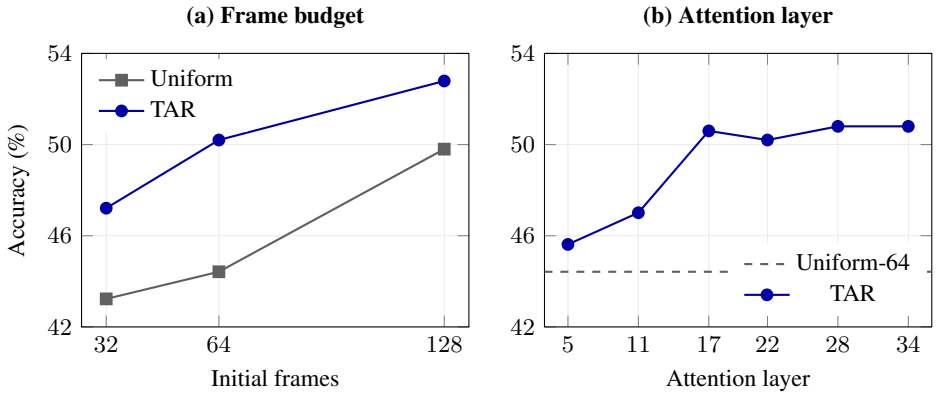


Figure 3: Sensitivity on MLVU with Qwen3-VL-8B. (a) Initial-frame budgets; TAR uses 24/48/96 final frames, respectively. (b) Attention layers at the default budget; the dashed line is Uniform-64. Lines connect evaluated settings only. Exact values are provided in Appendix C.

Table 3: External-retriever comparison on Video-MME. TAR and CLIP use the same 32G+16L answering view. The upper panel compares overall accuracy; the lower panel reports video-length groups. Changes are relative to CLIP.

Overall comparison		Accuracy (%)	Δ (pp)
Uniform-64		58.39	−0.56
CLIP (32G+16L)		58.95	–
TAR (32G+16L)		**60.73**	+1.78
Video length	**CLIP**	**TAR**	**Δ (pp)**
Short	70.67	**72.56**	+1.89
Medium	55.39	**57.73**	+2.34
Long	50.78	**51.89**	+1.11

4.6 Comparison with CLIP retrieval

The preceding controls compare internal attention variants. To assess an external alternative, we compare TAR with CLIP retrieval on Video-MME using the same 32-global/16-local final answering view. We use a CLIP ViT-L/14 (Radford et al., 2021) backbone, scoring coarse frames by cosine similarity with the question text embedding and selecting the same window length as TAR by maximizing summed similarity. TAR reaches 60.73%, compared with 58.95% for CLIP and 58.39% for Uniform-64 (Table 3). The paired comparison with CLIP has $p = 0.00935$. This supports the internal attention signal relative to the evaluated external similarity baseline, without requiring a separate retrieval model.

The aggregate advantage varies across capabilities. TAR improves in 9 of 12 categories, with a 7.19-point gain on OCR, but trails CLIP on attribute, temporal, and spatial perception (Appendix C, Table 6). The temporal and spatial perception gaps are 5.45 and 5.56 points, respectively. By video length, gains are positive for short, medium, and long groups (Table 3), although long videos show the smallest gain at 1.11 points. These breakdowns qualify the aggregate result and identify settings in which the retrieval signal remains less effective.

4.7 Temporal localization and answer improvement

Finally, we examine whether the accuracy gains align with retrieval of annotated evidence. We compare retrieved windows with temporal annotations on 1,534 LVBench examples with parseable annotations. Recall@1 under any overlap is 23.34%, mean temporal IoU is 3.20%, and ground-truth interval coverage is 13.93% (Appendix C, Table 7). The selected interval often fails to closely match the annotated evidence. These metrics describe absolute localization quality; establishing performance above chance would additionally require a matched random-window localization baseline.

To relate localization to answering, Fig. 4 separates questions according to whether the retrieved window overlaps annotated evidence. Examples with overlap improve from 41.06% to 47.49% (6.43 points); those without overlap improve from 32.57% to 33.76% (1.19 points). The larger gain in the overlap group is consistent with a benefit from acquiring relevant evidence. This relationship is observational: the groups differ in baseline accuracy, and overlap is not randomly assigned. The result also concerns the annotated subset, whereas the overall 2.39-point LVBench gain in Table 1 covers all 1,549 examples. Taken together, the accuracy and localization results support the value of targeted observation while showing that reliable temporal retrieval remains an open challenge.

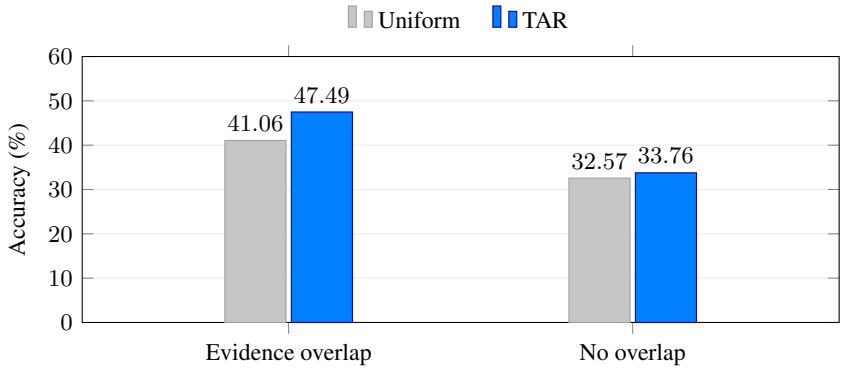


Figure 4: Answer accuracy by retrieved-window overlap on the annotated LVBench subset. The overlap group gains 6.43 points, versus 1.19 points without overlap. Groups differ in baseline accuracy and are not randomized; the comparison is observational. Full localization metrics appear in Appendix C.

5 Discussion and Limitations

What the controls establish. The results support a targeted claim: internal relative attention can guide useful temporal resampling within the evaluated global–local pipeline. On MLVU, the gain

persists against both matched uniform sampling and random-window resampling, indicating that the choice of interval matters beyond the final frame count. The attention ablation further favors the ratio over the tested raw and difference scores. These comparisons isolate components of the present procedure; broader comparisons with adaptive search methods are still needed to assess its relative strengths.

Answering budget and total cost. The frame savings apply to the final answering pass: 32 global and 16 local frames total 48 frames, which is 25% fewer than the 64-frame uniform baseline. Attention retrieval adds computation, so this reduction does not imply lower end-to-end inference cost; constructing the retrieval signal requires two attention-extraction passes over a coarse view, and extracting attention may materialize more intermediate state than ordinary generation. Final frame count, visual-token count, latency, and FLOPs therefore measure different aspects of resource use. A complete accuracy–cost curve requires end-to-end measurements that include video decoding, both attention-extraction passes, final answer generation, and peak memory under the same hardware and input-resolution settings. Comparing configurations at matched total latency or computation would clarify whether the accuracy gains justify retrieval overhead and distinguish the benefit of selecting evidence from that of spending additional inference resources.

Localization and distributed evidence. The LVBench analysis associates evidence overlap with larger answer gains, but absolute localization remains weak. Attention is a retrieval heuristic rather than a faithful causal attribution, and overlap is not randomly assigned. Long or multi-interval annotations and questions requiring distant events also challenge a single fixed-width window. Multi-window retrieval and question-adaptive window widths are natural extensions, but their value must be tested against the additional observation cost.

Interpreting retrieval as a proposal. The distinction between locating evidence and improving an answer helps reconcile the accuracy gains with weak absolute localization. A retrieved interval need not recover the complete annotated event to supply a useful additional observation, and the global stream can still support questions whose local window misses the evidence. Conversely, overlap alone does not ensure that the sampled frames capture the decisive action or that the model interprets it correctly. These possibilities are consistent with the pipeline, but the current diagnostics do not distinguish among them. A more direct test would compare attention-selected and random windows within the same questions, measure their annotation overlap, and track which added observations change the answer. Such an analysis would separate retrieval quality from the model's ability to use retrieved content. It would also help determine when to allocate more frames to local detail and when to preserve broader temporal coverage, particularly for questions whose required evidence spans several distant events.

Model access and evaluation scope. The method requires internal attention tensors and is therefore unsuitable for black-box APIs. Our implementation targets Qwen-family attention layouts; applying the same principle to another backbone requires a reliable mapping from visual tokens to time. Both evaluated models belong to that family, all tasks are multiple-choice, and inputs exclude audio and subtitles. Transfer to unrelated model families, open-ended generation, and multimodal evidence settings remains untested. The answer-informed oracle is also restricted to a diagnostic subset and should be interpreted as privileged selection headroom, rather than a deployable system or a bound on every evaluated benchmark.

6 Conclusion

We introduced TAR, a training-free approach that uses a VideoLLM's relative attention to decide when to look again. A coarse view guides contiguous-window retrieval, and dense local observations complement global coverage in the final answering pass. Accuracy improves across eight benchmarks and two Qwen-family model scales; matched controls support the importance of interval selection, and LVBench diagnostics associate evidence overlap with larger gains. These results show that internal attention can help acquire useful temporal evidence even though localization remains imperfect and retrieval adds computation, pointing to better temporal coverage and full accuracy–cost accounting as next steps.

AI USE STATEMENT

AI tools were used to assist with language editing and proofreading, including improving grammar, clarity, and overall readability. The ideas, arguments, and final content remain my own, and I reviewed and approved all AI-assisted revisions.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, et al. Scaling rl to long videos. *Advances in Neural Information Processing Systems*, 38:172842–172870, 2026.

Yeongjae Cho, Keonwoo Kim, Taebaek Hwang, and Sungzoon Cho. Do you keep an eye on what i ask? mitigating multimodal hallucination via attention-guided ensemble decoding. In *International Conference on Learning Representations*, volume 2025, pp. 39873–39893, 2025.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8450–8460. IEEE, 2025.

Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, et al. M-llm based video frame selection for efficient video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13702–13712. IEEE, 2025.

De-An Huang, Subhashree Radhakrishnan, Zhiding Yu, and Jan Kautz. Frag: Frame selection augmented generation for long video and long document understanding. *arXiv preprint arXiv:2504.17447*, 2025.

Mahyar Khayatkhoei, Prateek Chhikara, Filip Ilievski, et al. Mllms know where to look: Training-free perception of small visual details with multimodal llms. In *International Conference on Learning Representations*, volume 2025, pp. 68194–68213, 2025.

Geng Li, Jinglin Xu, Yunzhen Zhao, and Yuxin Peng. Dyfo: A training-free dynamic focus visual search for enhancing lmms in fine-grained visual understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9098–9108. IEEE, 2025.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. In *International Conference on Learning Representations*, volume 2026, pp. 109089–109117, 2026.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

Jihao Qiu, Lingxi Xie, Xinyue Huo, Qi Tian, and Qixiang Ye. Longvideo-r1: Smart navigation for low-cost long video understanding. *arXiv preprint arXiv:2602.20913*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models. In *International Conference on Learning Representations*, volume 2025, pp. 7593–7734, 2025.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024.

Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. Mdp 3: a training-free approach for list-wise frame selection in video-llms. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24090–24101. IEEE, 2025.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29118–29128. IEEE, 2025.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. Mllm can see? dynamic correction decoding for hallucination mitigation. In *International Conference on Learning Representations*, volume 2025, pp. 13712–13736, 2025a.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22958–22967. IEEE, 2025b.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3272–3282. IEEE, 2025c.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9772–9781. IEEE, 2021.

Baixuan Xu, Yinyui Xu, Tianshi Zheng, Zhaowei Wang, Weiqi Wang, Haochen Shi, Jiayu Liu, Qing Zong, Xiyu Ren, Xinyu Geng, et al. Finding the right evidence: Factor-guided coarse-to-fine reasoning for long videos. *arXiv preprint arXiv:2608.26355*, 2026.

Xinhao Xu, Hui Chen, Mengyao Lyu, Sicheng Zhao, Yizhe Xiong, Zijia Lin, Jungong Han, and Guiguang Ding. Mitigating hallucinations in multi-modal large language models via image token attention-guided decoding. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1571–1590, 2025.

Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, et al. Re-thinking temporal search for long-form video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8579–8591. IEEE, 2025.

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. *Advances in Neural Information Processing Systems*, 38:89863–89895, 2026.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13691–13701. IEEE, 2025.

Yuanhao Zou, Shengji Jin, Andong Deng, Youpeng Zhao, Jun Wang, and Chen Chen. Air: Enabling adaptive, iterative, and reasoning-based frame selection for video question answering. In *International Conference on Learning Representations*, volume 2026, pp. 11302–11329, 2026.

A Additional Experimental Details

Prompts. The retrieval question prompt p_Q is the original multiple-choice question. The generic reference prompt p_G is the fixed instruction "Please describe the video content."

Sampling. All streams are sampled directly from the source video using timestamps. Global frames span the full duration, while local frames span the retrieved interval. Random controls use a deterministic hash of the example identifier and seed to choose an interval of identical duration. Outputs are appended to per-method JSONL files and deduplicated by example identifier during aggregation, enabling restart-safe evaluation.

Statistics. For paired accuracy comparisons, we count discordant outcomes: fixes are examples where the baseline is wrong and TAR is correct; regressions are the converse. We use a two-sided exact McNemar test, equivalent to an exact binomial test over discordant pairs under equal probability of either direction.

B Additional Control Details

The three random-window seeds obtain 45.42%, 45.42%, and 45.82% accuracy, respectively, yielding the 45.55$\pm$0.19% result in Table 2. Uniform-48 obtains 230/502 correct (45.82%), Uniform-64 obtains 223/502 (44.42%), and TAR obtains 252/502 (50.20%). Thus the main improvement persists both against the user-facing 64-frame baseline and against a matched 48-frame final-answer budget.

For the Qwen3.6 Video-MME experiment, TAR corrects 252 baseline errors and changes 152 baseline successes to errors, a net gain of 100 questions. The corresponding two-sided exact McNemar test gives $p = 7.44 \times 10^{-7}$.

C Detailed Ablation and Localization Results

This appendix preserves the numerical results underlying the main-text sensitivity plots, the capability-level external-retriever comparison, and the localization diagnostics. All results are from the same evaluations reported in the main text. For the attention-score ablation in Table 2B, gains over Uniform-64 are 1.99, 2.19, 2.99, and 5.78 percentage points for question, generic, difference, and ratio attention, respectively.

Table 4: MLVU results for Qwen3-VL-8B across frame budgets. The baseline answers directly from uniformly sampled initial frames, whereas TAR performs temporal retrieval before the final answering pass. All values are accuracy (%); Δ denotes TAR minus the baseline.

Initial Frames	Final Frames	Baseline	TAR	Δ
32	24	43.23	**47.21**	+3.98
64	48	44.42	**50.20**	+5.78
128	96	49.80	**52.79**	+2.99

Table 5: Effect of the attention layer used for temporal localization on MLVU with Qwen3-VL-8B.

Setting	**Baseline**	**L5**	**L11**	**L17**	**L22**	**L28**	**L34**
Acc. (%)	44.42	45.62	47.01	50.60	50.20	**50.80**	**50.80**
Δ (pp)	–	+1.20	+2.59	+6.18	+5.78	**+6.37**	**+6.37**

Table 6: Capability-level comparison with CLIP retrieval on Video-MME. Δ denotes Ratio Attention minus CLIP; differences are computed from unrounded scores.

Capability	CLIP	Ratio	Δ
Object Reasoning	56.17	**59.03**	+2.86
Object Recognition	63.56	**66.38**	+2.82
Information Synopsis	74.61	**76.16**	+1.55
Action Recognition	52.08	**53.99**	+1.92
Action Reasoning	55.44	**57.19**	+1.75
Counting Problem	40.67	**41.79**	+1.12
Attribute Perception	**75.11**	74.66	-0.45
Temporal Reasoning	42.37	**43.50**	+1.13
OCR Problems	58.27	**65.47**	+7.19
Spatial Reasoning	75.00	**76.79**	+1.79
Temporal Perception	**70.91**	65.45	-5.45
Spatial Perception	**68.52**	62.96	-5.56

Table 7: Temporal localization diagnostics on 1,534 LVBench examples with parseable temporal annotations.

Metric	Result
Recall@1 (any overlap)	23.34%
Mean temporal IoU	3.20%
Ground-truth interval coverage	13.93%
Predicted center in ground truth	11.15%
Ground-truth center in prediction	16.75%
Accuracy gain with overlap	+6.43 pp
Accuracy gain without overlap	+1.19 pp