Seeing the Scene Matters: Revealing Forgetting in Video Understanding Models with a Scene-Aware Long-Video Benchmark

Supplementary Material

A. SceneQA Further Analysis

Figure 1 analyzes the impact of input frame length on question answering performance. SceneQA shows a slight improvement as the number of frames increases, indicating that visual-only models benefit from longer temporal context. In contrast, SceneQA-Audio achieves its best performance at 32 frames, with accuracy gradually declining as the frame length increases, suggesting that longer input sequences may introduce noise or redundant information in the audio modality. This trend indicates that while extended visual context can be beneficial, incorporating audio signals reduces the need for longer inputs and may even be negatively affected by excessively long temporal context.

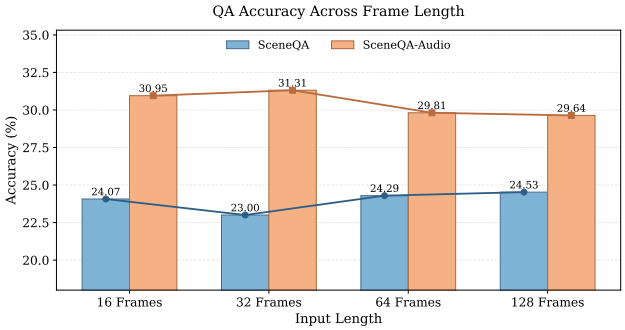

Figure 1. **QA Accuracy Across Frame Lengths**. Performance of SceneQA and SceneQA-Audio across different input frame lengths (16, 32, 64, and 128 frames). While SceneQA shows slight improvement with longer inputs, SceneQA-Audio performance peaks at moderate frame lengths and slightly declines for longer sequences.

Table 1 summarizes the temporal span distribution of SceneQA and SceneQA-Audio in our dataset. The timespan is calculated based on the duration of context required by models to answer each question, averaged across models. Both subsets cover a wide range of temporal distances, from short-range events under 250 seconds to long-range dependencies exceeding 2000 seconds. Most samples are concentrated in shorter temporal ranges (below 500 seconds), following a natural distribution, while we intentionally include longer-context questions to encourage long-range reasoning. Importantly, SceneQA and SceneQA-Audio exhibit comparable average spans within each bucket, ensuring consistent temporal coverage across subsets.

Table 1. Temporal span distribution of SceneQA and SceneQA-Audio. We report the number of samples and average temporal span (seconds) for each temporal range.

Temporal Range (s)	**SceneQA**		**SceneQA-Audio**	
	#Samples	**Avg. Span**	**#Samples**	**Avg.Span**
$[0, 250)$	2110	163.13	2953	158.95
$[250, 500)$	700	342.91	769	337.94
$[500, 750)$	256	608.29	187	588.53
$[750, 1000)$	57	853.72	34	861.50
$[1000, 1250)$	68	1139.16	22	1109.36
$[1250, 1500)$	16	1356.12	29	1344.28
$[1500, 1750)$	9	1582.33	4	1641.75
$[1750, 2000)$	15	1895.07	9	1883.56
$[2000, +\infty)$	56	2453.14	10	2270.70

B. Runtime Latency Analysis

We report the runtime of Scene-RAG breaking it down into offline preprocessing (video embedding, scene tiling, and audio captioning) and online inference (query rewriting and retrieval of relevant scene/audio context) in Table 2. A direct same-compute comparison is fundamentally flawed for long videos, since feeding equivalent frames directly causes out-of-memory while processing frames sequentially forces the baseline to treat the long video as disjointed short clips, destroying the global context required for reasoning.

Table 2. Runtime latency breakdown (video Length: 2,767s).

Stage	Visual Enc. (InternVideo2)	Audio (QwenAudio2)	LLM (Qwen3)	Total (Seconds)
Offline Preprocess	273.52	3.20	-	**276.72**
Online Inference	1.04	0.19	61.06	**62.29**

C. Annotation Details with example

Our benchmark is entirely manually annotated, specifically, Fig. 2 depicts the detailed annotation process for SceneQA. In terms of the Distance Definition, it is defined as the cue's observable interval. Regarding reproducibility, we will make our code and data publicly available before the deadline of camera ready.

D. Related Work: RAG for Long Video

Recent efforts to integrate Retrieval-Augmented Generation (RAG) with Multimodal Large Language Models (MLLMs) for long video understanding can be broadly categorized into online and offline approaches. In this work, we focus

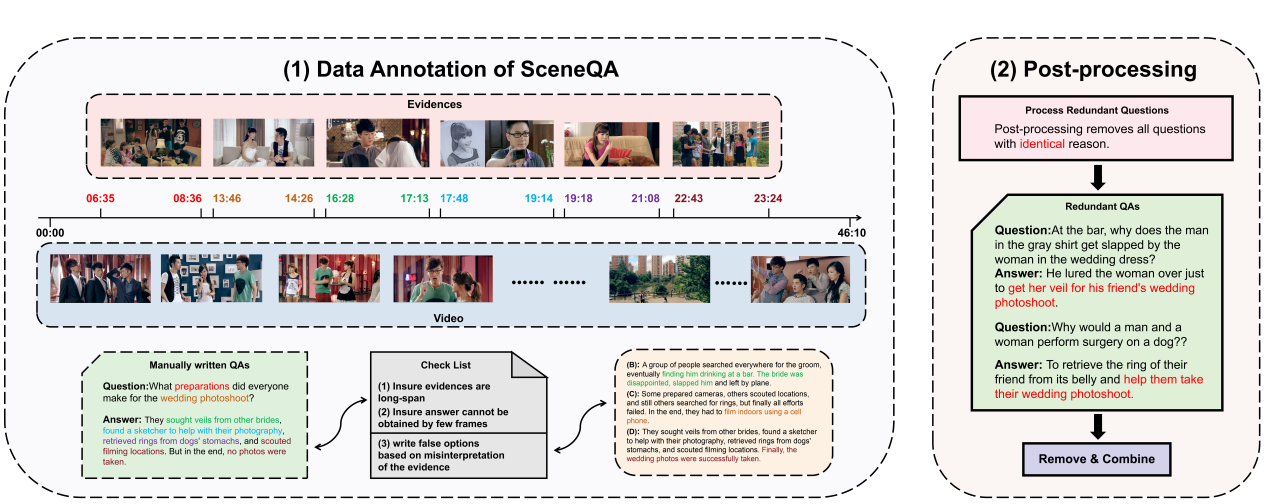

Figure 2. Overview of the Annotation pipeline. Zoom in for better visibility.

on offline RAG, where visual features are pre-extracted and reused for efficient query-based retrieval.

VideoRAG [3] adopts a straightforward pipeline: frames are sampled, and stored in a vector database for similarity search. While effective, uniform frame sampling can cause significant information loss. Video-RAG [4] enhances alignment by jointly modeling visual and textual semantics for more coherent retrieval. However, its dense representation leads to high memory usage and slower retrieval. Q-Frame [9] improves efficiency through query-aware frame selection, reducing redundant storage while maintaining relevance. Yet, it still requires handling large-scale video data. To address this, MemVid [7] proposes a memory-enhanced framework that organizes features into hierarchical memory slots, enabling more compact and context-aware retrieval for long videos.

Although MemVid improves efficiency in memory construction and retrieval, it assumes videos are composed of continuous clips. In contrast, our work targets scene-based reasoning, where semantically related content can be discontinuous and scattered across different segments of the video, posing new challenges beyond existing long video RAG methods.

E. Implementation Details of Scene-RAG

We implement Scene-RAG using PyTorch. For visual representation, we utilize the pre-trained InternVideo2-6B [5] backbone, frozen during inference. Audio streams are processed using Qwen-Audio2 [1] to extract captions for speech and background sound. For the Large Language Model (LLM) backbone, we employ Qwen3-14B [6]. The algorithm steps are summarized in Alg. 1.

Hyperparameters. The TV-L1 smoothing utilizes a regularization weight $\mu = 0.5$. The sensitivity parameter for scene detection is set to $\alpha = 1.5$ based on validation set performance. We filter out short segments with duration $L_{\min} < 3.0s$ to minimize noise. For retrieval, we maintain a memory bank size dynamic to the video length, retrieving the top-K ($K = 10$) most relevant scenes for final generation.

Algorithm 1 Scene-RAG Retrieval Pipeline

Require: Video $\mathcal{V}$, user query $\mathcal{Q}$, parameters μ, α and $L_{\min}$
Ensure: Answer $\mathcal{A}$

// Stage 1: Scene Tiling
Compute the raw similarity sequence $s_1, \ldots, s_n$ from $\mathcal{V}$
$x^* \leftarrow \arg\min_x \left[\frac{1}{2} \sum_t (x_t - s_t)^2 + \lambda \sum_t |x_t - x_{t-1}| \right]$.
Define the threshold $k \leftarrow \mu_{x^*} + \alpha \sigma_{x^*}$
Extract scene segments $\mathcal{S} = \{S_1, \ldots, S_m\}$ where $x_t^* \geq k$ and $|S_i| \geq L_{\min}$
// Stage 2: Memory Construction
for each segment $S_i \in \mathcal{S}$ **do**
    $v_i \leftarrow \text{InternVideo2}(S_i.\text{visual})$
    $a_i \leftarrow \text{QwenAudio}(S_i.\text{audio})$
    Memory bank $\mathcal{M} \leftarrow \mathcal{M} \cup \text{Concat}(v_i, a_i)$
end for
// Stage 3: Query Retrieval & Reasoning
Sub-queries $\{q_1, \ldots, q_p\} \leftarrow \text{Decompose}(\mathcal{Q}, \text{Qwen3})$
Indices $\mathcal{I} \leftarrow \emptyset$
for each q_j **do**
    $\mathcal{I} \leftarrow \text{TopK}(\mathcal{I} \cup (\text{Sim}(q_j, \mathcal{M}))$ // Global TopK
end for
Context $\mathcal{C} \leftarrow \text{Gather}(\mathcal{S}, \mathcal{I})$
$\mathcal{A} \leftarrow \text{LLM}(\mathcal{Q}, \mathcal{C})$

Table 3. Ablation over SceneTiling $(\alpha, L_{\min})$ and Scene-RAG retrieval size K on VideoMME [2]. Model: Longva [8].

Setting	TV-L1 α	TV-L1 $L_{\min}$(s)	Scene-RAG K	Avg. Result	Gain
Full Model (Ours)	1.5	3.0	10	62.4	-
$\alpha \downarrow$ (0.5)	0.5	3.0	10	61.2	-1.2
$\alpha \uparrow$ (2.0)	2.0	3.0	10	61.7	-0.7
$L_{\min} \downarrow$ (2s)	1.5	2.0	10	61.9	-0.5
$K \downarrow$ (5)	1.5	3.0	5	62.1	-0.3
$K \uparrow$ (15)	1.5	3.0	15	62.0	-0.4

Ablation Study. We conduct a controlled ablation to isolate the contributions of (i) adaptive scene-sensitivity thresholding, $k = \mu_x + \alpha\sigma_x$, (ii) TV-L1 smoothing for scene-tiling continuity (with $\alpha = 1.5$), and (iii) the number of retrieved scenes K used during memory-bank retrieval. First, removing the adaptive thresholding and using a fixed cutoff leads to a notable drop in scene-boundary accuracy, confirming that dynamic scaling with α better adapts to local motion statistics. Second, disabling TV-L1 smoothing (with default $L_{\min} = 3$) causes fragmented scene boundaries and increases false splits, demonstrating the importance of regularized temporal gradients for stable segmentation. Finally, we vary the top-K retrieved scenes (with default $K = 10$) and observe that too small a value underutilizes contextual history, while overly large K introduces irrelevant or noisy scenes. Table 3 summarizes these findings, where we evaluated a grid search of hyperparameters. We did not exhaustively explore all settings due to the computational cost of the experiments.

F. Labeling Challenges.

Annotating long and complex videos presents significant challenges, especially for the SceneQA and I-VQA tasks. Both questions and answers can be ambiguous. For instance, a reference to "the person in red" may correspond to multiple individuals appearing at different times, while visually similar scenes, such as different classrooms, can cause confusion when identifying the correct location. These ambiguities require precise temporal localization and careful contextual verification. Annotators often need to rewatch the entire video, confirm spatial and temporal references, and cross-check with others to ensure that each question–answer pair aligns with a single, unambiguous narrative. On average, annotating one QA pair takes about 36 minutes, excluding the additional time required for review and verification. If a scene cannot be reliably characterized due to visual blurring, semantic ambiguity, or unclear narrative boundaries, we will directly abandon annotation for that scene. This ensures that all scene-level questions in SceneBench possess clear semantic support and verifiability.

G. Ethical Concern and Data Publication

We do not own the video data. Instead, we collect access to publicly available videos in accordance with their original licensing conditions. We make reasonable efforts to ensure that the collected videos are legally redistributable and do not contain privacy-sensitive, illegal, or otherwise inappropriate content. The dataset will be released on HuggingFace.

H. Supplementary Task Examples

We also provide examples of I-VQA, Comment Prediction, and Title Prediction in Figure 3 and Figure 4.

References

[1] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024. 2

[2] Chaoyou Fu, Yuhan Dai, Yondong 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. *arXiv preprint arXiv:2405.21075*, 2024. 3

[3] Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. Videorag: Retrieval-augmented generation over video corpus, 2025. 2

[4] Yongdong Luo, Xiawu Zheng, Xiao Yang, Guilin Li, Haojia Lin, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. Video-rag: Visually-aligned retrieval-augmented long video comprehension, 2024. 2

[5] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024. 2

[6] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 2

[7] Huaying Yuan, Zheng Liu, Minghao Qin, Hongjin Qian, Yan Shu, Zhicheng Dou, Ji-Rong Wen, and Nicu Sebe. Memory-enhanced retrieval augmentation for long video understanding, 2025. 2

[8] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 3

[9] Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms, 2025. 2

IVQA

Question: The answer to all of the following questions is "Coal." Please select the question that best matches the video description.

(A) What traditional fuel powers the steam train shown in the mountain sightseeing tour?
(B) What black combustible mineral is loaded onto the trucks in the scene?
(C) In the local street barbecue scene, what type of fuel is used to grill the skewers?
(D) What type of energy do the three people in the video use to heat the water in the kettle?

Question: The answer to all of the following questions is "Because no customers were willing to visit his restaurant." Please select the question that best matches the video description.

(A) Why does the small green character slam the metal?
(B) Why is the small green character screaming and stomping his feet?
(C) Why does the red crab stick out his tongue after tasting the liquid?
(D) Why is the red crab attacking the small green character ?

Figure 3. Examples QAs of SceneBench. Overview of the problem set.

Comment prediction

Question: Which of the following is most likely to be a comment from viewers on this video?

(A) It inspires me and I'm sure others that we don't need all the gear and massive boats to do it. Keep up the awesome content!
(B) Bro I was worried something happened to you it been so long since you posted anything expect you was climbing AMA DABLAM I checked all social media platforms glad to see you back !
(C) You might not have realized it yet, but you are becoming a huge runner across the world. Congratulations on getting 3rd place!
(D) You guys are amazing saving wildlife in every episode and enlightening us on the true conditions of our oceans realtime!!! Stay safe and keep making these amazing videos!!!

Title Prediction

Question: Which of the following options is most likely to be the title of the video?

(A) Cycling through the Japanese countryside Explore ASUKA
(B) Cerro Torre Climb & Fly full movie
(C) December in Paris Vlog 3-Day Itinerary Moulin Rouge Show, Palais Garnier Opera
(D) Driving Las Vegas in 8K HDR Dolby Vision - Zion Utah to Las Vegas Nevada

Figure 4. Examples QAs of SceneBench. Overview of the problem set.