

# SceneRAG: Scene-level Retrieval-Augmented Generation for Video Understanding

Nianbo Zeng<sup>1,2</sup>, Haowen Hou<sup>1</sup>, Fei Richard Yu<sup>1,2</sup>, Si Shi<sup>1</sup>, Ying Tiffany He<sup>2</sup>

<sup>1</sup>Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen, China

<sup>2</sup>College of Computer Science and Software Engineering, Shenzhen University, China

{zengnianbo, houhaowen, yufei, shisi}@gml.ac.cn, tiffanyhe@szu.edu.cn

## Abstract

Despite recent advances in retrieval-augmented generation (RAG) for video understanding, effectively understanding long-form video content remains underexplored due to the vast scale and high complexity of video data. Current RAG approaches typically segment videos into fixed-length chunks, which often disrupts the continuity of contextual information and fails to capture authentic scene boundaries. Inspired by the human ability to naturally organize continuous experiences into coherent scenes, we present SceneRAG, a unified framework that leverages large language models to segment videos into narrative-consistent scenes by processing ASR transcripts alongside temporal metadata. SceneRAG further sharpens these initial boundaries through lightweight heuristics and iterative correction. For each scene, the framework fuses information from both visual and textual modalities to extract entity relations and dynamically builds a knowledge graph, enabling robust multi-hop retrieval and generation that account for long-range dependencies. Experiments on the LongerVideos benchmark, featuring over 134 hours of diverse content, confirm that SceneRAG substantially outperforms prior baselines, achieving a win rate of up to 72.5 percent on generation tasks.

## 1 Introduction

Video has become the dominant medium for communication, education, and entertainment in the digital era. Unlike unimodal text or audio, video integrates visual frames, spoken language, on-screen graphics, and ambient sound into a continuous multimodal stream. This richness enables deep storytelling and immersive learning, but also produces massive, unstructured data that challenges both retrieval and generative models [1, 45, 21]. In particular, static text-centric Retrieval-Augmented Generation (RAG) methods excel on documents but struggle to capture the temporal and multimodal complexity of video content [24, 16, 19].

Existing video segmentation approaches [6, 14] often rely on fixed-length windows or naive sliding clips, which frequently misalign with true scene boundaries and yield fragmented narratives, ultimately degrading retrieval precision and downstream performance. Moreover, current RAG systems [5, 24, 9] treat each segment as an isolated text block (e.g., subtitle snippet), ignoring inter-scene dependencies such as recurring characters, thematic motifs, and long-range references. Retrieval-augmented generation pipelines that process segments in isolation thus overlook critical narrative continuity, temporal coherence, and entity tracking.

A key insight to overcome these limitations comes from psychology [47]: humans naturally segment continuous experiences into discrete, meaningful “scenes” through an *event segmentation* process. This ability enables us to efficiently organize, comprehend, and recall long sequences of events. Film editors, for example, leverage this principle through continuity editing, guiding viewers seamlessly across cuts and locations [11, 29, 12]. By breaking down long videos into self-contained, semantically

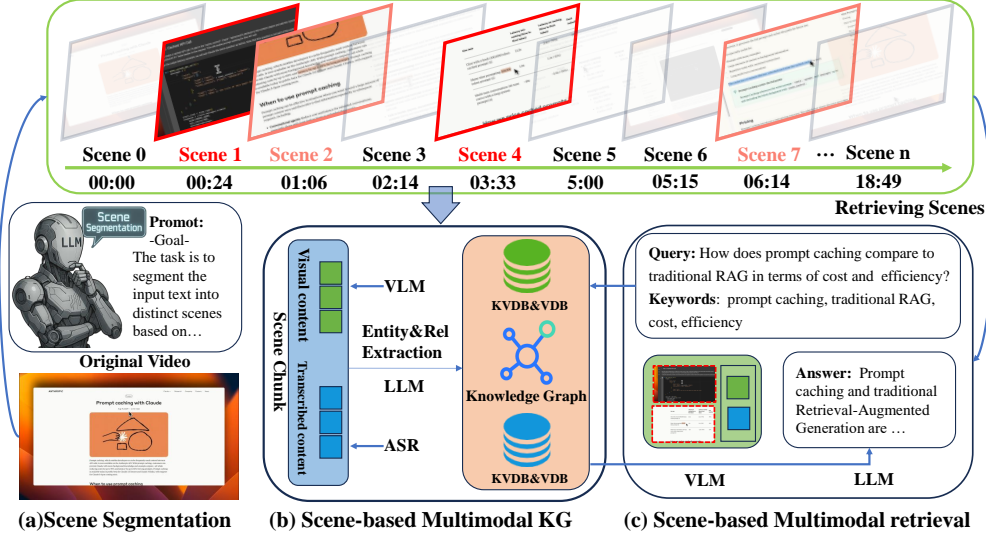


Figure 1: Overview of SceneRAG. Given a long video, SceneRAG segments it into scenes highlighted in green using an LLM and heuristics. For a query, relevant scenes are highlighted in red, and retrieved but irrelevant scenes are outlined in pink through knowledge graph-based retrieval.

coherent scenes—typically centered on a specific location, speaker, or topic—humans can more easily follow narratives and track salient entities and actions. Emulating this perceptual segmentation thus provides a principled foundation for automated video indexing, more precise retrieval, and robust cross-modal alignment in long-form video understanding.

To address these limitations, we propose **SceneRAG**, a unified framework that begins with human-inspired, LLM-driven scene segmentation. By reasoning over transcripts, subtitles, and temporal cues, the LLM produces semantically coherent scene boundaries aligned with narrative flow [8, 50]. To further enhance precision and robustness, we incorporate simple yet effective heuristics—such as Mute processing and segment alignment—along with a boundary correction mechanism to refine coarse boundaries and ensure accurate scene transition alignment. Next, SceneRAG constructs a lightweight knowledge graph [13] where nodes represent scenes, entities, and events, and edges capture semantic co-occurrence and temporal adjacency [22]. By enabling multi-hop retrieval over this graph, the model recovers long-range dependencies that flat retrieval cannot model. Given a user query, SceneRAG retrieves a relevant subgraph, aggregates multimodal signals, and issues a unified prompt to an LLM, resulting in context-rich generation.

Our contributions are as follows.

- **Human-inspired Scene Segmentation.** We introduce an LLM-based algorithm that fuses vision, audio, and subtitle cues to generate semantically coherent scene boundaries, emulating cognitive event segmentation, thereby refining existing segmentation methods.
- **Unified RAG Pipeline.** We propose a novel framework that cohesively combines dynamic segmentation, advanced graph construction, and retrieval-augmented generation, significantly enhancing the robustness and accuracy of long-form video understanding.
- **Comprehensive Evaluation.** Experiments on public benchmarks—including Lecture, Documentary and Entertainment datasets—demonstrate SceneRAG’s substantial improvements in retrieval precision and generation quality compared to state-of-the-art baselines.

## 2 Related work

**Multimodal Video Understanding as Foundation.** Recent video-language models have advanced joint representations across vision, text, and audio, enabling more comprehensive video understanding.

ViViT [3] uses transformers for spatio-temporal classification, while Flamingo bridges pretrained vision and language models for multimodal inputs, including videos [2]. MERLOT Reserve [48] aligns video, subtitles, and audio to improve representation learning. VideoBERT [38] quantizes frames into “visual words” and uses a BERT-like model to capture high-level semantics and temporal dynamics for tasks like question answering. VideoAgent [40] adds structured memory of events and objects, supporting complex LLM querying. These models provide strong multimodal encodings and inspire further exploration of temporal structure in video systems.

**Retrieval-Augmented Generation.** With rapid advances in retrieval-augmented generation (RAG) frameworks, knowledge-enhanced generation methods have evolved considerably. RAG [24] uses dense retrieval over flat text, while GraphRAG [13] and LightRAG [18] employ graph-based knowledge for more structured retrieval. Multimodal extensions such as WavRAG [7] and VideoRAG [36] incorporate audio or video, enabling cross-modal fusion, but often lack explicit scene awareness. SceneRAG addresses this gap by combining scene-aware segmentation with graph-based, multimodal knowledge integration for more contextually grounded retrieval. Table 1 summarizes these representative RAG-based methods, highlighting key differences in modality, structure, and scene awareness.

Table 1: Comparison of RAG-based Methods for Knowledge-Enhanced Generation.

| Model    | Modality    | Knowledge Structure | Scene Awareness | Multimodal Fusion |
|----------|-------------|---------------------|-----------------|-------------------|
| RAG      | Text        | Flat                | ✗               | ✗                 |
| GraphRAG | Text        | Graph               | ✗               | ✗                 |
| LightRAG | Text        | Graph               | ✗               | ✗                 |
| WavRAG   | Audio, Text | Flat                | ✗               | ✓                 |
| VideoRAG | Video, Text | Graph               | ✗               | ✓                 |
| SceneRAG | Video, Text | Graph               | ✓               | ✓                 |

**Scene Structuring for Retrieval and Reasoning.** Segmenting videos into shots and scenes has long served as the foundation for video alignment and retrieval [46, 32]. Traditional methods mainly detect shot transitions based on low-level visual features. Recent advances leverage self-supervised learning and contrastive objectives [41, 31, 39] to improve boundary detection and capture richer scene structures, which benefit downstream retrieval tasks. In egocentric and cross-view scenarios, scene-level alignment is also key for connecting first- and third-person perspectives [43]. However, most existing segmentation methods still lack semantic and narrative awareness, often resulting in boundaries that do not align with human-perceived scene changes. With the emergence of large pre-trained vision-language models [25, 26, 50], research has begun to explore human-inspired, cognitively motivated scene segmentation—aiming to partition video into units that better reflect narrative flow and human perception.

### 3 Method

Section 3.1 outlines SceneRAG’s architecture; Section 3.2 introduces hybrid multimodal scene segmentation; Section 3.3 describes scene-based knowledge grounding via a lightweight graph; and Section 3.4 presents the retrieval-augmented generation module.

#### 3.1 Framework Overview

Figure 1 shows the overall architecture of SceneRAG. Given a video  $V$ , we first divide it uniformly into fixed-length clips of duration  $T$ , resulting in a series of segments  $v_1, v_2, \dots, v_n$ , each with transcript snippets and timestamps  $\mathcal{T} = t_1, t_2, \dots, t_n$ . These multimodal inputs are fed into a large language model,  $\text{LLM}(\cdot)$ , which produces an initial set of scene boundaries  $\mathcal{S}^0$ . The boundaries are then refined with heuristic rules to obtain the final scene set  $\mathcal{S} = S_1, S_2, \dots, S_m$ . As each new scene  $S_i$  is identified, the knowledge graph  $G = (N, E)$  is dynamically updated, where nodes  $N$  denote scenes and entities, and edges  $E$  represent relationships such as co-occurrence and temporal adjacency. Given a query  $Q$ , SceneRAG extracts keywords and retrieves candidate segments via multi-hop reasoning on  $G$ , aggregating context to generate a context-aware answer  $A$ .

### 3.2 Automatic Scene Segmentation

**Chunk-wise Processing Pipeline.** To facilitate scalable processing of long videos, we uniformly divide each video into a sequence of overlapping temporal chunks. Each chunk has a fixed length of  $L = 5$  minutes with a 10-second overlap between adjacent segments:

$$\text{Chunks} = \{\text{Chunk}_k\}_{k=1}^K, \quad \text{Chunk}_k \in [t_k^{\text{start}}, t_k^{\text{end}}] \quad (1)$$

where  $K$  denotes the total number of chunks for a given video.

**Chunk-level Audio Transcription (ASR).** For each chunk  $\text{Chunk}_k$ , an automatic speech recognition system [10, 15] is used to obtain a localized transcript  $\mathcal{T}_k$ :

$$\mathcal{T}_k = \text{ASR}(\text{Chunk}_k) \quad (2)$$

**Scene Segmentation within Each Chunk.** Given the chunk-wise transcript  $\mathcal{T}_k$ , we construct a prompt [13]  $P_k$  and apply a language model  $f$  to extract intra-chunk scene boundaries [20]:

$$\hat{S}_k = f(P_k) = \{(t_{i,k}^{\text{start}}, t_{i,k}^{\text{end}}, \hat{s}_{i,k})\}_{i=1}^{N_k} \quad (3)$$

where  $t_{i,k}^{\text{start}}$  and  $t_{i,k}^{\text{end}}$  represent the start and end timestamps of the  $i$ -th scene within chunk  $k$ , and  $N_k$  is the total number of scenes predicted in that chunk. As in the global segmentation process, we enforce constraints on the number, length, and coverage of segments. If any constraints are violated, the system adaptively re-prompts the model and escalates to a stronger one if necessary.

**Silence-Aware Refinement (per chunk).** Silent intervals are extracted using the ASR system’s capabilities, identifying segments where no speech is detected. These silent intervals  $\mathcal{E}_k$  provide important structural cues for refining segmentation. They help detect scene boundaries that may not be captured purely from textual signals. We classify each interval  $e_{j,k}$  based on its duration  $\Delta t_{j,k} = t_{j,k}^{\text{end}} - t_{j,k}^{\text{start}}$  and process it accordingly:

- **Short Silence Assignment:** if  $\Delta t_{j,k} \leq \epsilon$ , set to 10 seconds by default, the silence is evenly split and assigned to adjacent scenes:

$$\text{Assign}(e_{j,k}) = \begin{cases} \text{Previous scene: } [t_{j,k}^{\text{start}}, t_{j,k}^{\text{mid}}] \\ \text{Next scene: } [t_{j,k}^{\text{mid}}, t_{j,k}^{\text{end}}] \end{cases}, \quad t_{j,k}^{\text{mid}} = \frac{t_{j,k}^{\text{start}} + t_{j,k}^{\text{end}}}{2} \quad (4)$$

- **Long Silence Promotion:** if  $\Delta t_{j,k} > \epsilon$ , the silence is promoted as an independent scene:

$$S_{j,k}^{(\text{silent})} = (t_{j,k}^{\text{start}}, t_{j,k}^{\text{end}}, [\text{SILENT}]) \quad (5)$$

These silence-derived segments are retained in downstream stages and later processed using multi-modal models [44] to extract scene-level representations. This prevents information loss from visually or acoustically rich transitions that may lack textual signals.

**Temporal Adjustment and Post-hoc Correction.** To ensure stable segmentation, we refine scene boundaries through two strategies. First, segments shorter than 10 seconds are merged with neighboring ones based on temporal and semantic proximity. Second, all scene boundaries are adjusted to align with sentence-level punctuation in the transcript  $\mathcal{T}_k$ . For each predicted scene  $S_{i,k}$ , we extract its aligned transcript segment:

$$T_{i,k} = \mathcal{T}_k|_{[t_{i,k}^{\text{start}}, t_{i,k}^{\text{end}}]} \quad (6)$$

**Final Output.** The output of the segmentation module is a collection [37] of aligned scene units that encapsulate temporal boundaries and transcripts:

$$\mathcal{S} = \bigcup_{k=1}^K \{(t_{i,k}^{\text{start}}, t_{i,k}^{\text{end}}, T_{i,k})\}_{i=1}^{N_k} \quad (7)$$

This final set  $\mathcal{S}$  provides a unified, structured representation of the video, serving as the foundation for downstream tasks such as summarization, classification, or multimodal grounding.

### 3.3 Scene-based Multimodal Knowledge Grounding

**Multimodal Scene Representation.** For each scene  $S_j = (t_j^{\text{start}}, t_j^{\text{end}}, s_j)$ , we leverage both the transcript  $T_j$  and the visual appearance to construct a comprehensive and unified textual representation. To efficiently represent the visual content of each scene, we uniformly sample  $k$  key frames (typically every 6 seconds, with  $k \leq 10$ ), denoted as  $F_1, F_2, \dots, F_k$ . These sampled frames, together with the corresponding scene transcript  $T_j$ , are fed into a visual language model (VLM) to produce a natural language description  $C_j$  that captures objects, actions, and high-level scene dynamics [33, 30, 28]:

$$C_j = \text{VLM}(T_j, \{F_1, \dots, F_k\}) \quad (8)$$

At this stage, each scene has two complementary textual modalities: the transcript  $T_j$  (audio/text) and the visual description  $C_j$  (visual/text), enabling richer dual-channel knowledge extraction [42]. To maximize cross-modal coverage, we extract entities and relations separately from  $C_j$  and  $T_j$  using large language models (LLMs) [13, 18]:  $(N_j^{(\text{vis})}, E_j^{(\text{vis})})$  from  $C_j$ , and  $(N_j^{(\text{asr})}, E_j^{(\text{asr})})$  from  $T_j$ . This dual-path strategy preserves modality-specific information and effectively alleviates the modality bias issues observed in prior work [36].

**Multimodal Entity and Relation Fusion.** We merge entities and relations from both modalities into a unified scene-level knowledge set, using LLM-assisted disambiguation to align semantically equivalent entities and preserve cross-modal relations [23, 4]:

$$N_j = \text{Fuse}(N_j^{(\text{vis})}, N_j^{(\text{asr})}) \quad (9)$$

$$E_j = \text{Fuse}(E_j^{(\text{vis})}, E_j^{(\text{asr})}) \quad (10)$$

yielding a comprehensive set of entities  $N_j$  and relations  $E_j$  for each scene.

**Knowledge Graph Construction and Integration.** Adopting the incremental graph-building strategy from GraphRAG [13], we assemble all scene-level entities and relations into a unified knowledge graph  $G = (N, E)$ :

$$G = (N, E) = \bigcup_{j=1}^{N'} (N_j, E_j) \quad (11)$$

where  $N'$  is the total number of scenes. For multi-video corpora, we unify semantically equivalent entities and relations across videos for cross-video knowledge aggregation and entity enrichment [23]. LLMs synthesize comprehensive entity descriptions from multiple scene contexts [33].

**Multi-Modal Context Encoding** Each scene’s visual description  $C_j$  and transcript  $T_j$  are concatenated into a unified text chunk  $H_j$ , allowing for richer contextual representation that captures both visual and textual cues. This chunk is then embedded using a pre-trained multimodal encoder [17].

$$H_j = [C_j; T_j] \quad (12)$$

$$e_t^j = \text{TEnc}(H_j) \quad (13)$$

where  $C_j$  denotes the visual description of the scene and  $T_j$  is its transcript. All scene embeddings are aggregated into a matrix  $E_t \in \mathbb{R}^{N' \times d_t}$ , where  $N'$  is the number of scenes and  $d_t$  is the embedding dimension. These embeddings are later utilized for efficient semantic retrieval and downstream integration with external knowledge sources.

### 3.4 Scene-based Retrieval-Augmented Generation

**Token-Budgeted Scene Retrieval.** Given a query  $q$ , we first encode it as  $e_q$  using the same text encoder as for scene embeddings [34, 17]. The relevance between the query and each scene is measured by cosine similarity  $s_j = \cos(e_q, e_t^j)$ , and we select a subset of scenes  $\mathcal{R}_q^*$  under a total token-length constraint  $\tau$ :

$$\mathcal{R}_q^* = \arg \max_{\mathcal{R} \subseteq \{1, \dots, N'\}} \sum_{j \in \mathcal{R}} s_j \quad \text{s.t.} \quad \sum_{j \in \mathcal{R}} \text{Len}(C_j, T_j) \leq \tau \quad (14)$$

where  $\text{Len}(C_j, T_j)$  is the total token length of the scene’s content and  $\tau$  is the maximum allowed token budget (e.g., 2400 tokens).

Table 2: Comparison between SceneRAG and five baselines (NaiveRAG, GraphRAG<sub>1</sub> (GraphRAG-local), GraphRAG<sub>2</sub> (GraphRAG-global), LightRAG and VideoRAG) on the LongerVideos dataset. Numbers indicate win-rates (%). The “All” column aggregates results from the three domains.

|                   | Lecture               |              | Documentary           |              | Entertainment         |              | All                   |              |
|-------------------|-----------------------|--------------|-----------------------|--------------|-----------------------|--------------|-----------------------|--------------|
|                   | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     |
| Comprehensiveness | 35.9%                 | <b>64.1%</b> | 30.7%                 | <b>69.3%</b> | 35.1%                 | <b>64.9%</b> | 34.8%                 | <b>65.2%</b> |
| Empowerment       | 34.2%                 | <b>65.8%</b> | 26.8%                 | <b>73.2%</b> | 33.9%                 | <b>66.1%</b> | 32.8%                 | <b>67.2%</b> |
| Trustworthiness   | 37.0%                 | <b>63.0%</b> | 28.2%                 | <b>71.8%</b> | 37.5%                 | <b>62.5%</b> | 35.4%                 | <b>64.6%</b> |
| Depth             | 35.3%                 | <b>64.7%</b> | 27.3%                 | <b>72.7%</b> | 34.2%                 | <b>65.8%</b> | 33.6%                 | <b>66.4%</b> |
| Density           | 50.3%                 | <b>49.7%</b> | 50.4%                 | <b>49.6%</b> | 48.4%                 | <b>51.6%</b> | 50.0%                 | <b>50.0%</b> |
| Overall Winner    | 36.2%                 | <b>63.8%</b> | 28.7%                 | <b>71.3%</b> | 34.5%                 | <b>65.5%</b> | 34.5%                 | <b>65.5%</b> |
|                   | GraphRAG <sub>1</sub> | SceneRAG     | GraphRAG <sub>1</sub> | SceneRAG     | GraphRAG <sub>1</sub> | SceneRAG     | GraphRAG <sub>1</sub> | SceneRAG     |
|                   | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     |
| Comprehensiveness | 32.9%                 | <b>67.1%</b> | 37.7%                 | <b>62.3%</b> | 39.4%                 | <b>60.6%</b> | 35.0%                 | <b>65.0%</b> |
| Empowerment       | 28.5%                 | <b>71.5%</b> | 33.6%                 | <b>66.4%</b> | 34.8%                 | <b>65.2%</b> | 30.7%                 | <b>69.3%</b> |
| Trustworthiness   | 31.6%                 | <b>68.4%</b> | 32.9%                 | <b>67.1%</b> | 35.3%                 | <b>64.7%</b> | 32.6%                 | <b>67.4%</b> |
| Depth             | 29.2%                 | <b>70.8%</b> | 33.6%                 | <b>66.4%</b> | 33.6%                 | <b>66.4%</b> | 30.9%                 | <b>69.1%</b> |
| Density           | 36.3%                 | <b>63.7%</b> | 45.8%                 | <b>54.2%</b> | 42.0%                 | <b>58.0%</b> | 39.1%                 | <b>60.9%</b> |
| Overall Winner    | 29.5%                 | <b>70.5%</b> | 35.1%                 | <b>64.9%</b> | 35.0%                 | <b>65.0%</b> | 31.6%                 | <b>68.4%</b> |
|                   | GraphRAG <sub>2</sub> | SceneRAG     | GraphRAG <sub>2</sub> | SceneRAG     | GraphRAG <sub>2</sub> | SceneRAG     | GraphRAG <sub>2</sub> | SceneRAG     |
|                   | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     |
| Comprehensiveness | 29.9%                 | <b>70.1%</b> | 35.7%                 | <b>64.3%</b> | 40.0%                 | <b>60.0%</b> | 32.9%                 | <b>67.1%</b> |
| Empowerment       | 26.1%                 | <b>73.9%</b> | 32.5%                 | <b>67.5%</b> | 37.5%                 | <b>62.5%</b> | 29.4%                 | <b>70.6%</b> |
| Trustworthiness   | 26.2%                 | <b>73.8%</b> | 28.1%                 | <b>71.9%</b> | 31.4%                 | <b>68.6%</b> | 27.5%                 | <b>72.5%</b> |
| Depth             | 25.8%                 | <b>74.2%</b> | 30.1%                 | <b>69.9%</b> | 34.8%                 | <b>65.2%</b> | 28.3%                 | <b>71.7%</b> |
| Density           | 35.5%                 | <b>64.5%</b> | 51.3%                 | <b>48.7%</b> | 49.9%                 | <b>50.1%</b> | 41.2%                 | <b>58.8%</b> |
| Overall Winner    | 26.0%                 | <b>74.0%</b> | 32.7%                 | <b>67.3%</b> | 36.2%                 | <b>63.8%</b> | 29.2%                 | <b>70.8%</b> |
|                   | LightRAG              | SceneRAG     | LightRAG              | SceneRAG     | LightRAG              | SceneRAG     | LightRAG              | SceneRAG     |
|                   | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     |
| Comprehensiveness | 32.4%                 | <b>67.6%</b> | 31.9%                 | <b>68.1%</b> | 33.5%                 | <b>66.5%</b> | 32.5%                 | <b>67.5%</b> |
| Empowerment       | 28.5%                 | <b>71.5%</b> | 30.1%                 | <b>69.9%</b> | 30.4%                 | <b>69.6%</b> | 29.2%                 | <b>70.8%</b> |
| Trustworthiness   | 31.5%                 | <b>68.5%</b> | 28.7%                 | <b>71.3%</b> | 31.8%                 | <b>68.2%</b> | 31.0%                 | <b>69.0%</b> |
| Depth             | 28.7%                 | <b>71.3%</b> | 27.8%                 | <b>72.2%</b> | 29.6%                 | <b>70.4%</b> | 28.7%                 | <b>71.3%</b> |
| Density           | 42.6%                 | <b>57.4%</b> | 50.3%                 | <b>49.7%</b> | 42.9%                 | <b>57.1%</b> | 44.1%                 | <b>55.9%</b> |
| Overall Winner    | 29.8%                 | <b>70.2%</b> | 29.5%                 | <b>70.5%</b> | 31.1%                 | <b>68.9%</b> | 30.0%                 | <b>70.0%</b> |
|                   | VideoRAG              | SceneRAG     | VideoRAG              | SceneRAG     | VideoRAG              | SceneRAG     | VideoRAG              | SceneRAG     |
|                   | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     |
| Comprehensiveness | 43.8%                 | <b>56.2%</b> | 40.2%                 | <b>59.8%</b> | 43.3%                 | <b>56.7%</b> | 43.1%                 | <b>56.9%</b> |
| Empowerment       | 42.6%                 | <b>57.4%</b> | 39.0%                 | <b>61.0%</b> | 42.4%                 | <b>57.6%</b> | 41.9%                 | <b>58.1%</b> |
| Trustworthiness   | 42.4%                 | <b>57.6%</b> | 37.4%                 | <b>62.6%</b> | 40.0%                 | <b>60.0%</b> | 41.0%                 | <b>59.0%</b> |
| Depth             | 42.5%                 | <b>57.5%</b> | 38.3%                 | <b>61.7%</b> | 42.3%                 | <b>57.7%</b> | 41.7%                 | <b>58.3%</b> |
| Density           | 47.7%                 | <b>52.3%</b> | 47.2%                 | <b>52.8%</b> | 50.9%                 | <b>49.1%</b> | 48.2%                 | <b>51.8%</b> |
| Overall Winner    | 42.8%                 | <b>57.2%</b> | 38.9%                 | <b>61.1%</b> | 42.7%                 | <b>57.3%</b> | 42.0%                 | <b>58.0%</b> |

**Scene-aware Answer Generation.** Given a query  $q$ , we first extract salient keywords using an LLM [18], and generate query-focused visual captions for each retrieved scene by prompting a VLM with these keywords, the transcript, and sampled frames. We then aggregate all focused captions, transcripts, and knowledge-graph components from the relevant scenes to construct the context:

$$C_q = \text{Concat}(\{\hat{C}_j, T_j, (N_j, E_j) \mid j \in \mathcal{R}_q^*\}) \quad (15)$$

where  $\hat{C}_j$  is the query-focused visual caption for scene  $j$ ,  $T_j$  is the transcript,  $(N_j, E_j)$  are the node and edge sets of the knowledge graph for  $j$ ,  $\mathcal{R}_q^*$  is the set of relevant scenes retrieved for  $q$ , and  $C_q$  is the constructed context. The answer is then generated by feeding the query and context into an LLM:

$$A_q = \text{LLM}(q, C_q) \quad (16)$$

where  $A_q$  denotes the generated answer for the input query  $q$ , using context  $C_q$ .

## 4 Experiments

### 4.1 Experimental Setup

**Dataset** We evaluate our approach on the LONGERVIDEOS [36] benchmark—a large-scale dataset curated to assess models on extended video content. It contains 164 videos across 22 lists, totaling over 134 hours, and spans three challenging domains: lectures, documentaries, and entertainment. Compared to prior benchmarks focused on short clips or single-topic narratives, LONGERVIDEOS

offers a more realistic and demanding testbed. Its videos often exceed 30 minutes, exhibit multiple semantic shifts, and contain rich multimodal signals, making it well-suited for evaluating scene-level reasoning, retrieval-augmented generation, and temporal comprehension under complex structure.

**Baseline Methods.** To comprehensively evaluate the effectiveness of our proposed method, we compare SceneRAG with a range of baselines across three categories: (1) Text-centric RAG methods, including NaiveRAG [16], GraphRAG (both local and global variants)[13], and LightRAG[18], which operate solely on text transcripts without modeling multimodal or temporal structure; (2) Video-centric RAG frameworks, including VideoRAG [36], which builds a large-scale multimodal graph across videos, performing semantic retrieval based on visual and textual features; (3) General long video understanding models, such as LLaMA-VID [27], VideoAgent [14], and NotebookLM, designed for long-context vision-language understanding but not following the RAG paradigm.

**Evaluation Protocol.** We adopt a two-part LLM-based evaluation protocol for long-context video understanding: *Win-Rate Comparison* and *Quantitative Scoring*. Both use five human-centric dimensions: (1) *Comprehensiveness*—coverage of query intent; (2) *Empowerment*—perceived helpfulness; (3) *Trustworthiness*—factual accuracy; (4) *Depth*—reasoning quality; and (5) *Density*—amount of relevant, non-redundant information. Scores are averaged across three domains (lectures, documentaries, entertainment). Win-rate is based on LLM-judged pairwise comparisons; quantitative scores are LLM-rated against NaiveRAG on a 5-point Likert scale.

**Implementation Details.** We adopt the same experimental setup as VideoRAG, ensuring fair comparisons across models. All methods use GPT-4o-mini for generation, with shared prompt formatting and input encoding pipelines. ASR transcripts are generated using Distil-Whisper [35, 15]. Visual and textual features are extracted using ImageBind [17], which serves as the unified multimodal encoder  $\text{MEnc}(\cdot)$ . All baselines utilize grounded textual knowledge, including visual captions and transcripts from videos, and apply the same chunk-splitting protocol as our method. Entity- and chunk-level retrieval rely on OpenAI’s `text-embedding-3-small` model, while vision-language modeling uses MiniCPM-V [44]. Videos are sampled uniformly at 6-second intervals, and transcript, frame, metadata inputs, as well as all retrieved segments, are processed through a shared normalization pipeline for consistency.

## 4.2 Main Results

We evaluate SceneRAG on the *LongerVideos* dataset using a two-part win-rate study, employing three language models (**GPT-4o-mini**, **GPT-4.1-mini**, and **GPT-4.1-nano**) to ensure robustness and minimize model bias. Results are aggregated across models unless otherwise specified, with per-model details in Appendix C.

**Comparison with Text-Based RAG Baselines.** We compare SceneRAG against five strong text-centric retrieval-augmented generation baselines—NaiveRAG, GraphRAG-local, GraphRAG-global, LightRAG, and VideoRAG—on the *LongerVideos* dataset. Results in Table 2 demonstrate that SceneRAG consistently outperforms all baselines across domains and evaluation dimensions. SceneRAG achieves the highest average overall win-rate of 70.8% when compared to GraphRAG-global, marking it as the strongest model in the comparison. Gains are particularly substantial in dimensions such as Empowerment (up to 73.9%) and Depth (up to 74.2%), underscoring SceneRAG’s ability to integrate temporally dispersed multimodal evidence into coherent, in-depth scene-level outputs. These consistent improvements suggest that scene-aware structuring and multimodal graph-based retrieval provide a significant advantage in understanding long-form videos.

**Comparison with Multimodal RAG Baseline.** We further compare SceneRAG with VideoRAG, a representative multimodal RAG system that retrieves over concatenated video-text representations. SceneRAG outperforms VideoRAG across all five human evaluation metrics, achieving an overall win rate of 56.9% versus 43.1%. While VideoRAG relies on dense video-text embeddings without explicit structural modeling, SceneRAG incorporates scene-aware segmentation and graph-based propagation to improve contextual alignment. This leads to more semantically coherent and temporally grounded outputs, especially in scenarios requiring fine-grained disambiguation or multi-hop reasoning over long-form content. Compared to VideoRAG, which relies on fixed-interval segmentation, our approach aligns better with semantic boundaries, leading to more coherent and relevant outputs.

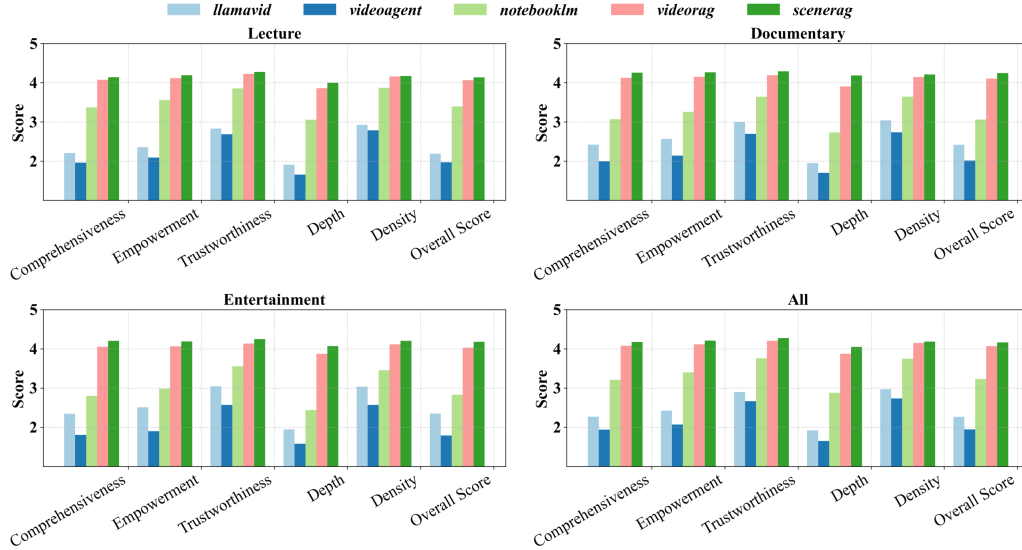


Figure 2: Quantitative Comparison with NaiveRAG Using a 5-Point Likert Scale. The “All” column aggregates results from the three domains.

**Comparison with Large Vision-Language Models.** As shown in Figure 2, SceneRAG consistently outperforms large vision-language models (LVMs) across all four domains, with particularly strong gains in lecture and documentary content that require long-range, multimodal reasoning. This advantage stems from SceneRAG’s dynamic scene segmentation and graph-based retrieval. While LVMs typically process fixed-length video chunks without structural cues, SceneRAG organizes content into semantically coherent scenes, enabling better alignment of audio, visual, and textual information. The graph module further supports cross-scene linking and retrieval, enhancing contextual understanding in complex narratives. These design choices allow SceneRAG to better handle temporal dependencies and topic transitions, especially in genres where meaning builds progressively over time.

### 4.3 Case Study

SceneRAG’s ability to locate and leverage cohesive segments in long-form videos is evaluated via a case study on the “Is This the End of RAG? Anthropic’s NEW Prompt Caching” lecture. For the query “How does prompt caching compare to traditional RAG in cost and efficiency?”, SceneRAG identifies non-contiguous, complementary segments—[23.96–66.08 s] and [213.12–292.48 s]—totaling 122 s. These clips provide a complete, focused answer despite being distributed across the timeline.

**Coherent Segmentation** Traditional methods relying on dense chunking or uniform sliding windows often fragment context, leading to partial or diluted answers. In contrast, SceneRAG’s scene-level segmentation ensures that each retrieved segment preserves a coherent narrative unit. The first retrieved scene introduces the motivation behind prompt caching and frames the cost/latency benefits in contrast to traditional RAG approaches. The second scene dives into concrete quantitative comparisons (e.g., 80–90% reduction in latency and cost), clarifies trade-offs like cache write costs, and juxtaposes Anthropic’s and Gemini’s implementations.

Table 3: Segment Retrieval Comparison. Symbols denote retrieval quality: ✓ indicates all retrieved segments are strongly coherent, relevant, or fully answer the query with key points; ✗ means some segments are fragmented, loosely related, or partially address key information.

| Model    | #Segments | Duration (s) | Coherence | Relevance | Key Points |
|----------|-----------|--------------|-----------|-----------|------------|
| SceneRAG | 2         | 122          | ✓         | ✓         | ✓          |
| VideoRAG | 9         | 270          | ✗         | ✗         | ✓          |



**Precise Retrieval** SceneRAG retrieves only the most relevant, semantically cohesive segments, ensuring responses remain focused. By avoiding irrelevant content, SceneRAG reduces processing time. For example, it retrieved two highly relevant segments totaling 122 seconds. This precision improves efficiency and minimizes distractions. In contrast, VideoRAG retrieved nine segments (270 seconds), including less relevant content, increasing processing time and task complexity. Such over-retrieval is especially problematic for large videos or when VLM resources are limited. The impact is shown in Table 3.

**Domain Adaptability** SceneRAG’s alignment of retrieval granularity with natural discourse boundaries proves especially beneficial in structured content such as technical lectures. Its ability to aggregate temporally distant yet topically connected segments enables accurate, context-rich responses, even when critical information is spread across multiple video parts. This domain adaptability makes SceneRAG an optimal choice for queries spanning multiple scenes in long-form content.

#### 4.4 Ablation

**Ablation Study.** We perform an ablation study to evaluate the impact of components in our scene segmentation pipeline. As shown in Table 4, LLM-based scene-aware segmentation enhances performance across all metrics compared to raw input. Additional rule-based refinements—Short segment allocation and mute processing—yield further improvements, particularly in Depth and Comprehensiveness. This underscores the importance of accurate scene segmentation for enhancing retrieval relevance and generation quality in long-context video understanding.

Table 4: Results of ablation study. “/” denotes fixed segmentation without LLM assistance; “+LLM” adds LLM-guided segmentation; “+LLM+Rules” further incorporates rule-based refinement.

|                   | Lecture |      |                | Documentary |      |                | Entertainment |      |                | All  |      |                |
|-------------------|---------|------|----------------|-------------|------|----------------|---------------|------|----------------|------|------|----------------|
|                   | /       | +LLM | +LLM<br>+Rules | /           | +LLM | +LLM<br>+Rules | /             | +LLM | +LLM<br>+Rules | /    | +LLM | +LLM<br>+Rules |
| Comprehensiveness | 4.35    | 4.36 | 4.51           | 4.33        | 4.44 | 4.56           | 4.30          | 4.39 | 4.52           | 4.35 | 4.38 | 4.52           |
| Empowerment       | 4.49    | 4.45 | 4.57           | 4.38        | 4.56 | 4.64           | 4.29          | 4.42 | 4.56           | 4.43 | 4.47 | 4.58           |
| Trustworthiness   | 4.44    | 4.46 | 4.55           | 4.39        | 4.43 | 4.61           | 4.37          | 4.53 | 4.58           | 4.42 | 4.47 | 4.57           |
| Depth             | 4.22    | 4.22 | 4.44           | 4.10        | 4.31 | 4.56           | 4.03          | 4.29 | 4.45           | 4.16 | 4.25 | 4.47           |
| Density           | 4.51    | 4.53 | 4.60           | 4.53        | 4.68 | 4.62           | 4.46          | 4.46 | 4.62           | 4.50 | 4.55 | 4.61           |
| Overall Score     | 4.37    | 4.35 | 4.49           | 4.31        | 4.40 | 4.54           | 4.24          | 4.38 | 4.51           | 4.33 | 4.37 | 4.50           |

**Effect of Scene-Based Segmentation.** Compared to fixed-length segmentation, scene-based segmentation aligns boundaries with actual semantic shifts, preserving contextual coherence and narrative flow. As shown in Table 5, this approach results in denser knowledge graphs [49], capturing more entities and relations. By structuring information around scenes, relevant entities are more tightly connected, facilitating multi-hop retrieval and improving both retrieval precision and answer generation.

Table 5: Graph Expansion Comparison. Values are normalized to the fixed-segmentation baseline.

|       | / | +LLM | +LLM<br>+Rules |
|-------|---|------|----------------|
| nodes | 1 | 1.13 | 1.29           |
| edges | 1 | 1.22 | 1.34           |

## 5 Conclusion

Long video analysis poses significant challenges in computer vision and machine learning, as existing models struggle to capture complex temporal dependencies and dynamic transitions across lengthy sequences. Inspired by human perception, our model, SceneRAG, addresses these challenges by segmenting video content into evolving, semantically coherent scenes, enhancing understanding of long-term dependencies and context shifts. By focusing on scene-level granularity, SceneRAG efficiently tracks and predicts scene boundaries and transitions, while preserving narrative continuity. In the future, we plan to incorporate advanced spatiotemporal cues and multi-modal data to further improve scene segmentation, enabling more accurate scene prediction, dynamic scene understanding, and flexible handling of diverse video content.

## References

- [1] Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdiah Soleymani Baghshah, and Ehsaneddin Asgari. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. *arXiv preprint arXiv:2502.08826*, 2025.
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736, 2022.
- [3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.
- [4] Anurag Arnab, Chen Sun, and Cordelia Schmid. Unified graph structured models for video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8117–8126, 2021.
- [5] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2023.
- [6] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495, 2024.
- [7] Yifu Chen, Shengpeng Ji, Haoxiao Wang, Ziqing Wang, Siyu Chen, Jinzheng He, Jin Xu, and Zhou Zhao. Wavrag: Audio-integrated retrieval augmented generation for spoken dialogue models. *arXiv preprint arXiv:2502.14727*, 2025.
- [8] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.
- [9] W. Chu, J. Lee, P. Xu, et al. Graphrag: A graph-enhanced retrieval-augmented generation framework. In *NeurIPS*, 2024.
- [10] Sampled Chunks. Sscformer: Push the limit of chunk-wise conformer for streaming asr using sequentially sampled chunks and chunked causal convolution.
- [11] James E Cutting. Event segmentation and seven types of narrative discontinuity in popular movies. *Acta psychologica*, 149:69–77, 2014.
- [12] James E Cutting and Kacie L Armstrong. Large-scale narrative events in popular cinema. *Cognitive Research: Principles and Implications*, 4:1–18, 2019.
- [13] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitan, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.
- [14] Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pages 75–92. Springer, 2024.
- [15] Sanchit Gandhi, Patrick von Platen, and Alexander M Rush. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. *arXiv preprint arXiv:2311.00430*, 2023.
- [16] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2:1, 2023.

- [17] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190, 2023.
- [18] Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation. 2024.
- [19] Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. Videorag: Retrieval-augmented generation over video corpus. *arXiv preprint arXiv:2501.05874*, 2025.
- [20] Lei Ji, Chenfei Wu, Daisy Zhou, Kun Yan, Edward Cui, Xilin Chen, and Nan Duan. Learning temporal video procedure segmentation from an automatically collected large dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1506–1515, 2022.
- [21] Peng Jin, Hao Li, Zesen Cheng, Kehan Li, Xiangyang Ji, Chang Liu, Li Yuan, and Jie Chen. Diffusionret: Generative text-video retrieval with diffusion model. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2470–2481, 2023.
- [22] Hyeongjin Kim, Sangwon Kim, Dasom Ahn, Jong Taek Lee, and Byoung Chul Ko. Scene graph generation strategy with co-occurrence knowledge and learnable term frequency. *arXiv preprint arXiv:2405.12648*, 2024.
- [23] Junlin Lee, Yequan Wang, Jing Li, and Min Zhang. Multimodal reasoning with multimodal knowledge graph. *arXiv preprint arXiv:2406.02030*, 2024.
- [24] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.
- [25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.
- [26] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.
- [27] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024.
- [28] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroan Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020.
- [29] Joseph P Magliano and Jeffrey M Zacks. The impact of continuity editing in narrative film on event segmentation. *Cognitive science*, 35(8):1489–1517, 2011.
- [30] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on international conference on multimedia retrieval*, pages 19–27, 2018.
- [31] Jonghwan Mun, Minchul Shin, Gunsoo Han, Sangho Lee, Seongsu Ha, Joonseok Lee, and Eun-Sol Kim. Boundary-aware self-supervised learning for video scene segmentation. *arXiv preprint arXiv:2201.05277*, 2022.
- [32] Gautam Pal, Dwijen Rudrapaul, Suvojit Acharjee, Ruben Ray, Sayan Chakraborty, and Nilanjan Dey. Video shot boundary detection: a review. In *Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India CSI Volume 2*, pages 119–127. Springer, 2015.
- [33] Jinghui Peng, Xinyu Hu, Wenbo Huang, and Jian Yang. What is a multi-modal knowledge graph: A survey. *Big Data Research*, 32:100380, 2023.

- [34] 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*, pages 8748–8763. PmLR, 2021.
- [35] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.
- [36] Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. Videorag: Retrieval-augmented generation with extreme long-context videos. *arXiv preprint arXiv:2502.01549*, 2025.
- [37] Panagiotis Sidiropoulos, Vasileios Mezaris, Ioannis Kompatsiaris, Hugo Meinedo, Miguel Bualho, and Isabel Trancoso. Temporal video segmentation to scenes using high-level audiovisual features. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(8):1163–1177, 2011.
- [38] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473, 2019.
- [39] Jiawei Tan, Pingan Yang, Lu Chen, and Hongxing Wang. Temporal scene montage for self-supervised video scene boundary detection. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(7):1–19, 2024.
- [40] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer, 2024.
- [41] Haoqian Wu, Keyu Chen, Yanan Luo, Ruizhi Qiao, Bo Ren, Haozhe Liu, Weicheng Xie, and Linlin Shen. Scene consistency representation learning for video scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14021–14030, 2022.
- [42] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021.
- [43] Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Retrieval-augmented egocentric video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13525–13536, 2024.
- [44] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.
- [45] Huaying Yuan, Jian Ni, Yueze Wang, Junjie Zhou, Zhengyang Liang, Zheng Liu, Zhao Cao, Zhicheng Dou, and Ji-Rong Wen. Momentseeker: A comprehensive benchmark and a strong baseline for moment retrieval within long videos. *arXiv preprint arXiv:2502.12558*, 2025.
- [46] Jinhui Yuan, Huiyi Wang, Lan Xiao, Wujie Zheng, Jianmin Li, Fuzong Lin, and Bo Zhang. A formal study of shot boundary detection. *IEEE transactions on circuits and systems for video technology*, 17(2):168–186, 2007.
- [47] Jeffrey M Zacks, Nicole K Speer, Khena M Swallow, and Corey J Maley. The brain’s cutting-room floor: Segmentation of narrative cinema. *Frontiers in human neuroscience*, 4:168, 2010.
- [48] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387, 2022.
- [49] Hairong Zhang, Jiaheng Si, Guohang Yan, Boyuan Qi, Pinlong Cai, Song Mao, Ding Wang, and Botian Shi. Rakg: Document-level retrieval augmented knowledge graph construction. *arXiv preprint arXiv:2504.09823*, 2025.

[50] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

## A Implementation Details

SceneRAG adopts a modular architecture inspired by the experimental setup of VideoRAG. Input videos are segmented into overlapping 5-minute chunks using FFmpeg, with a 10-second overlap to preserve contextual continuity. We employ `faster-distil-whisper-large-v3` for automatic speech recognition (ASR), which provides timestamped transcripts and identifies low-energy silence intervals.

Scene segmentation is performed via an LLM-based strategy, as shown in Algorithm 1. The system starts with a lightweight model (GPT-4o-mini) using temperature = 0.7 and top\_p = 0.95. We utilize the prompts illustrated in figure 4 to guide the model’s output. If the returned scene boundaries are inconsistent or violate duration constraints, we iteratively retry up to four times, potentially escalating to a more powerful model (GPT-4o) and adapting the prompts based on the type of failure, such as overly long or short segments.

The output is further refined through post-processing heuristics. Silence-aware adjustment and interpolation are critical, as silent segments are typically omitted from ASR transcripts and manifest as temporal gaps between dialogue-based scene predictions. We detect low-energy silence intervals to inform scene boundaries, promoting long silences (e.g., greater than 10 seconds) as new segment breaks, while redistributing shorter silences across neighboring scenes or filling them through boundary expansion to ensure continuous temporal coverage. Additionally, we merge scene segments that are shorter than 10 seconds with adjacent segments. This merging strategy takes into account both segment duration and textual coherence, favoring merge directions that result in a smoother scene flow. To enhance the accuracy of scene segmentation, we implement a scene correction mechanism, utilizing prompts depicted in figure 3 to improve the model’s understanding and judgment of scene boundaries. For visual understanding, we extract representative video frames and encode them using MiniCPM-V-2\_6-int4. To unify information across video, audio, and text, we employ ImageBind for cross-modal embedding into a shared vector space. The system infrastructure comprises a KV store for caching video paths, transcript blocks, and LLM results, a graph database to represent inter-scene relations and entities, and a vector database for multimodal dense retrieval. All experiments are executed on a single NVIDIA RTX 3090 GPU.

---

**Algorithm 1** LLM-Based Video Scene Segmentation

---

**Require:** Transcript with timestamps  $\mathcal{T}$ , Global config  $\mathcal{C}$ , Video duration  $D$

**Ensure:** Scene time intervals  $\{t_1, t_2, \dots, t_n\}$ , Scene descriptions  $\{s_1, s_2, \dots, s_n\}$

```
1: Format segmentation prompt  $P$  using  $\mathcal{T}$ 
2: Initialize retry count  $r \leftarrow 0$ 
3: repeat
4:    $R \leftarrow \text{LLMFunc}(P, \text{history})$ 
5:    $T \leftarrow \text{ExtractTimeRanges}(R)$ 
6:   if  $\text{CheckTimeRanges}(T, D)$  is valid then
7:     break
8:   else
9:      $P' \leftarrow \text{ChooseFixPrompt}(\text{error})$ 
10:     $P \leftarrow P'$ 
11:     $r \leftarrow r + 1$ 
12:   end if
13: until  $r \geq 4$ 
14: if no valid  $S$  is found then
15:   Return default time intervals and empty description list
16: end if
17:  $S \leftarrow \text{SplitText}(R)$ 
18:  $T \leftarrow \text{FillTimeGaps}(S, T, D)$ 
19:  $T, S \leftarrow \text{MergeIntervals}(T, S)$ 
20: Return  $(T, S)$ 
```

---

### Scene\_Segmentation\_repaired

#### **scene\_segmentation\_too\_little**

Output Error Correction Request:

Too few time ranges. Need at least 3 segments. The previous output has errors. Please verify and correct the following:

1. Ensure each scene has a duration between 15 and 60 seconds.
2. Verify that the scenes are divided correctly based on the content.
3. Ensure each scene starts with a time mark.
4. Ensure each scene contains detailed descriptions, dialogues, or events to form a coherent narrative unit.

Please maintain the required format in your response.

#### **scene\_segmentation\_too\_short**

Output Error Correction Request:

Some scenes have been split with time ranges that are too short. Merge scenes that are too short to meet the minimum duration requirement. Please verify and correct the following:

1. Ensure each scene has a duration between 15 and 60 seconds.
2. Verify that the scenes are divided correctly based on the content.
3. Ensure each scene starts with a time mark.
4. Ensure each scene contains detailed descriptions, dialogues, or events to form a coherent narrative unit.
5. Each scene should follow the previous one in a logical time sequence without gaps or overlaps.

Please maintain the required format in your response.

#### **scene\_segmentation\_too\_long**

Output Error Correction Request:

Some scenes have been split with time ranges that are too long. Split scenes that exceed the maximum duration into smaller segments. Please verify and correct the following:

1. The duration of each scene should ideally not exceed 60 seconds.
2. Verify that the scenes are divided correctly based on the content.
3. Ensure each scene starts with a time mark.
4. Ensure each scene contains detailed descriptions, dialogues, or events to form a coherent narrative unit.
5. Each scene should follow the previous one in a logical time sequence without gaps or overlaps.

Please maintain the required format in your response.

Figure 3: Error correction instructions for scene segmentation.

## **B Limitations and Future Work**

SceneRAG has two main limitations. First, the scene segmentation pipeline relies heavily on timestamped transcripts from ASR and large language models guided by hand-crafted prompts. This introduces sensitivity to transcription errors and prompt formulation, which can lead to suboptimal or inconsistent scene boundaries—particularly in noisy audio conditions or ambiguous dialogue contexts. Second, the current pipeline underutilizes non-verbal signals: while visual embeddings are incorporated downstream, the segmentation process is primarily text-driven. As a result, visually grounded transitions—such as emotional shifts, scene composition changes, or camera cuts without dialogue—may go undetected or misaligned. To address these challenges, we envision two future directions. One is to incorporate low-level visual and audio cues (e.g., shot boundary detection, background music changes, facial expression shifts) into the segmentation process to capture non-verbal scene transitions. The other is to reduce reliance on handcrafted prompts by leveraging prompt-free or instruction-tuned models that can better generalize across content domains.

## Scene\_Segmentation

### -Goal-

The task is to segment the input text into distinct scenes based on the given criteria. The segmentation should be done purely based on the content provided, without the need for summarization or interpretation.

### -Steps-

#### 1. Scene Identification and Segmentation

- Identify distinct scenes in the text, and need to reflect on why these scenes are segmented. The segmentation should be based solely on the content and structure of the text.
- Ensure each scene contains detailed descriptions, dialogues, or events to form a coherent narrative unit, and must not consist of a single sentence.

#### 2. Time Range and Scene Delimiters

- For each scene, record the time range (if available) at the beginning in the format [start\_time -> end\_time].
- Each scene should follow the previous one in a logical time sequence without gaps or overlaps.
- Add the scene content after the time range.
- The duration of each scene is between 15 and 60 seconds, except for those that you think are special.
- End each scene with {record\_delimiter} (except the last scene).

#### 3. Final Marker

- After all scenes, add {completion\_delimiter} to indicate the end of the task.

#### 4. Output Format

- Return the segmented text as a list of scenes.

- Output format Example:

Scene 1{record\_delimiter}

Scene 2{record\_delimiter}

Scene 3{record\_delimiter}

Scene 4{completion\_delimiter}

- Output only the segmented text without additional explanations. #####

Text: {input\_text}

Figure 4: Instructions for scene segmentation and formatting.

## C Supplementary Results

To ensure the robustness and fairness of our win-rate evaluation, we employ three language models—GPT-4o-mini, GPT-4.1-mini, and GPT-4.1-nano—to independently assess the outputs. For each comparison instance, we conduct two separate evaluations by swapping the order of the candidate answers, thereby mitigating potential position bias. The win-rate results reported for each model are averaged over both answer orders. Detailed per-model win-rate statistics are presented in the accompanying tables/figures. Notably, SceneRAG achieves even better performance when evaluated with the latest GPT-4.1-mini model, consistently outperforming baseline methods across all evaluation metrics and domains. This further demonstrates the robustness and effectiveness of SceneRAG, particularly when assessed by advanced open-domain language models.



Table 6: Comparison of Winning Rates by GPT-4o-mini.

|                   | Lecture               |              | Documentary           |              | Entertainment         |              | All                   |              |
|-------------------|-----------------------|--------------|-----------------------|--------------|-----------------------|--------------|-----------------------|--------------|
|                   | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     |
| Comprehensiveness | 42.2%                 | <b>57.8%</b> | 36.4%                 | <b>63.6%</b> | 37.5%                 | <b>62.5%</b> | 40.2%                 | <b>59.8%</b> |
| Empowerment       | 39.8%                 | <b>60.2%</b> | 30.3%                 | <b>69.7%</b> | 35.7%                 | <b>64.3%</b> | 37.2%                 | <b>62.8%</b> |
| Trustworthiness   | 39.8%                 | <b>60.2%</b> | 31.1%                 | <b>68.9%</b> | 35.7%                 | <b>64.3%</b> | 37.4%                 | <b>62.6%</b> |
| Depth             | 39.8%                 | <b>60.2%</b> | 33.3%                 | <b>66.7%</b> | 37.1%                 | <b>62.9%</b> | 38.0%                 | <b>62.0%</b> |
| Density           | 40.3%                 | <b>59.7%</b> | 37.3%                 | <b>62.7%</b> | 37.1%                 | <b>62.9%</b> | 39.1%                 | <b>60.9%</b> |
| Overall Winner    | 42.0%                 | <b>58.0%</b> | 35.5%                 | <b>64.5%</b> | 37.1%                 | <b>62.9%</b> | 39.9%                 | <b>60.1%</b> |
|                   | GraphRAG <sub>1</sub> | SceneRAG     | GraphRAG <sub>1</sub> | SceneRAG     | GraphRAG <sub>1</sub> | SceneRAG     | GraphRAG <sub>1</sub> | SceneRAG     |
|                   | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     |
| Comprehensiveness | 39.9%                 | <b>60.1%</b> | 42.1%                 | <b>57.9%</b> | 42.4%                 | <b>57.6%</b> | 40.8%                 | <b>59.2%</b> |
| Empowerment       | 37.9%                 | <b>62.1%</b> | 39.5%                 | <b>60.5%</b> | 40.6%                 | <b>59.4%</b> | 38.7%                 | <b>61.3%</b> |
| Trustworthiness   | 37.5%                 | <b>62.5%</b> | 36.8%                 | <b>63.2%</b> | 39.7%                 | <b>60.3%</b> | 37.8%                 | <b>62.2%</b> |
| Depth             | 37.1%                 | <b>62.9%</b> | 40.4%                 | <b>59.6%</b> | 40.6%                 | <b>59.4%</b> | 38.4%                 | <b>61.6%</b> |
| Density           | 34.7%                 | <b>65.3%</b> | 40.8%                 | <b>59.2%</b> | 41.5%                 | <b>58.5%</b> | 37.1%                 | <b>62.9%</b> |
| Overall Winner    | 39.0%                 | <b>61.0%</b> | 42.1%                 | <b>57.9%</b> | 42.0%                 | <b>58.0%</b> | 40.1%                 | <b>59.9%</b> |
|                   | GraphRAG <sub>2</sub> | SceneRAG     | GraphRAG <sub>2</sub> | SceneRAG     | GraphRAG <sub>2</sub> | SceneRAG     | GraphRAG <sub>2</sub> | SceneRAG     |
|                   | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     |
| Comprehensiveness | 34.8%                 | <b>65.2%</b> | 42.1%                 | <b>57.9%</b> | 41.5%                 | <b>58.5%</b> | 37.5%                 | <b>62.5%</b> |
| Empowerment       | 33.1%                 | <b>66.9%</b> | 39.5%                 | <b>60.5%</b> | 40.2%                 | <b>59.8%</b> | 35.6%                 | <b>64.4%</b> |
| Trustworthiness   | 30.7%                 | <b>69.3%</b> | 35.1%                 | <b>64.9%</b> | 32.6%                 | <b>67.4%</b> | 31.9%                 | <b>68.1%</b> |
| Depth             | 34.3%                 | <b>65.7%</b> | 38.6%                 | <b>61.4%</b> | 40.2%                 | <b>59.8%</b> | 36.2%                 | <b>63.8%</b> |
| Density           | 31.2%                 | <b>68.8%</b> | 41.7%                 | <b>58.3%</b> | 40.2%                 | <b>59.8%</b> | 34.9%                 | <b>65.1%</b> |
| Overall Winner    | 34.8%                 | <b>65.2%</b> | 41.7%                 | <b>58.3%</b> | 41.1%                 | <b>58.9%</b> | 37.3%                 | <b>62.7%</b> |
|                   | LightRAG              | SceneRAG     | LightRAG              | SceneRAG     | LightRAG              | SceneRAG     | LightRAG              | SceneRAG     |
|                   | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     |
| Comprehensiveness | 37.9%                 | <b>62.1%</b> | 36.4%                 | <b>63.6%</b> | 37.1%                 | <b>62.9%</b> | 37.5%                 | <b>62.5%</b> |
| Empowerment       | 34.7%                 | <b>65.3%</b> | 33.3%                 | <b>66.7%</b> | 33.9%                 | <b>66.1%</b> | 34.3%                 | <b>65.7%</b> |
| Trustworthiness   | 35.2%                 | <b>64.8%</b> | 30.3%                 | <b>69.7%</b> | 33.9%                 | <b>66.1%</b> | 34.1%                 | <b>65.9%</b> |
| Depth             | 35.2%                 | <b>64.8%</b> | 32.5%                 | <b>67.5%</b> | 36.2%                 | <b>63.8%</b> | 34.9%                 | <b>65.1%</b> |
| Density           | 34.0%                 | <b>66.0%</b> | 38.2%                 | <b>61.8%</b> | 33.0%                 | <b>67.0%</b> | 34.6%                 | <b>65.4%</b> |
| Overall Winner    | 37.5%                 | <b>62.5%</b> | 36.0%                 | <b>64.0%</b> | 37.1%                 | <b>62.9%</b> | 37.1%                 | <b>62.9%</b> |
|                   | VideoRAG              | SceneRAG     | VideoRAG              | SceneRAG     | VideoRAG              | SceneRAG     | VideoRAG              | SceneRAG     |
|                   | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     |
| Comprehensiveness | 46.1%                 | <b>53.9%</b> | 45.2%                 | <b>54.8%</b> | 46.0%                 | <b>54.0%</b> | 45.9%                 | <b>54.1%</b> |
| Empowerment       | 45.3%                 | <b>54.7%</b> | 42.5%                 | <b>57.5%</b> | 43.8%                 | <b>56.2%</b> | 44.5%                 | <b>55.5%</b> |
| Trustworthiness   | 44.5%                 | <b>55.5%</b> | 39.0%                 | <b>61.0%</b> | 41.1%                 | <b>58.9%</b> | 42.9%                 | <b>57.1%</b> |
| Depth             | 45.7%                 | <b>54.3%</b> | 43.9%                 | <b>56.1%</b> | 46.9%                 | <b>53.1%</b> | 45.6%                 | <b>54.4%</b> |
| Density           | 45.2%                 | <b>54.8%</b> | 39.9%                 | <b>60.1%</b> | 45.1%                 | <b>54.9%</b> | 44.2%                 | <b>55.8%</b> |
| Overall Winner    | 45.9%                 | <b>54.1%</b> | 45.2%                 | <b>54.8%</b> | 46.0%                 | <b>54.0%</b> | 45.8%                 | <b>54.2%</b> |

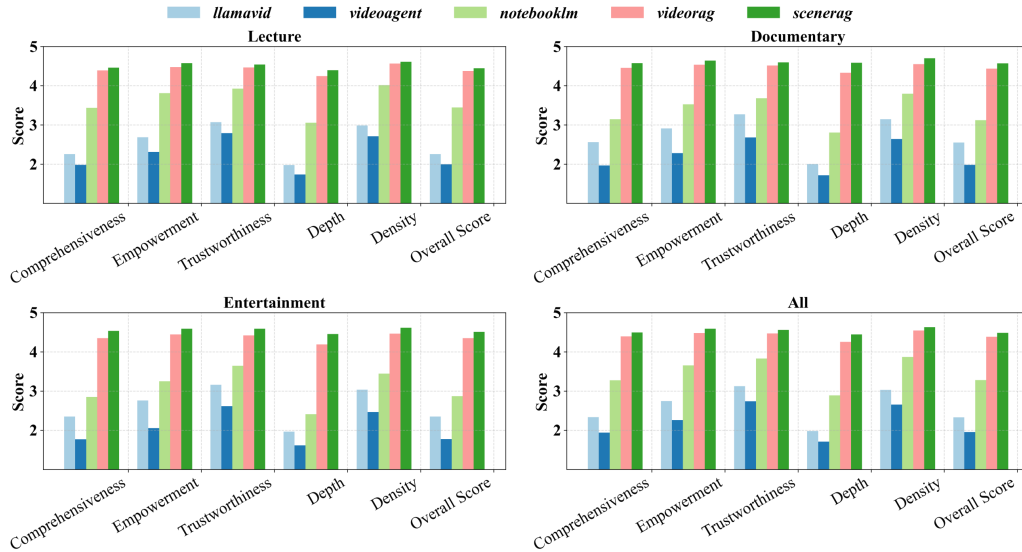


Figure 5: Quantitative Comparison by gpt-4o-mini.

Table 7: Comparison of Winning Rates by GPT-4.1-mini.

|                   | Lecture               |              | Documentary           |              | Entertainment         |              | All                   |              |
|-------------------|-----------------------|--------------|-----------------------|--------------|-----------------------|--------------|-----------------------|--------------|
|                   | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     |
| Comprehensiveness | 26.1%                 | <b>73.9%</b> | 23.2%                 | <b>76.8%</b> | 29.9%                 | <b>70.1%</b> | 26.2%                 | <b>73.8%</b> |
| Empowerment       | 23.7%                 | <b>76.3%</b> | 19.7%                 | <b>80.3%</b> | 29.0%                 | <b>71.0%</b> | 23.9%                 | <b>76.1%</b> |
| Trustworthiness   | 31.5%                 | <b>68.5%</b> | 25.4%                 | <b>74.6%</b> | 37.9%                 | <b>62.1%</b> | 31.6%                 | <b>68.4%</b> |
| Depth             | 28.1%                 | <b>71.9%</b> | 22.4%                 | <b>77.6%</b> | 30.8%                 | <b>69.2%</b> | 27.5%                 | <b>72.5%</b> |
| Density           | 60.6%                 | <b>39.4%</b> | 57.9%                 | <b>42.1%</b> | 54.5%                 | <b>45.5%</b> | 59.0%                 | <b>41.0%</b> |
| Overall Winner    | 27.7%                 | <b>72.3%</b> | 21.1%                 | <b>78.9%</b> | 30.4%                 | <b>69.6%</b> | 26.9%                 | <b>73.1%</b> |
|                   | GraphRAG <sub>1</sub> | SceneRAG     | GraphRAG <sub>1</sub> | SceneRAG     | GraphRAG <sub>1</sub> | SceneRAG     | GraphRAG <sub>1</sub> | SceneRAG     |
|                   | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     |
| Comprehensiveness | 20.1%                 | <b>79.9%</b> | 29.4%                 | <b>70.6%</b> | 32.6%                 | <b>67.4%</b> | 24.2%                 | <b>75.8%</b> |
| Empowerment       | 11.6%                 | <b>88.4%</b> | 23.2%                 | <b>76.8%</b> | 24.6%                 | <b>75.4%</b> | 16.2%                 | <b>83.8%</b> |
| Trustworthiness   | 23.3%                 | <b>76.7%</b> | 23.7%                 | <b>76.3%</b> | 28.6%                 | <b>71.4%</b> | 24.3%                 | <b>75.7%</b> |
| Depth             | 17.8%                 | <b>82.2%</b> | 23.2%                 | <b>76.8%</b> | 25.0%                 | <b>75.0%</b> | 20.2%                 | <b>79.8%</b> |
| Density           | 35.0%                 | <b>65.0%</b> | 47.8%                 | <b>52.2%</b> | 38.8%                 | <b>61.2%</b> | 38.1%                 | <b>61.9%</b> |
| Overall Winner    | 16.1%                 | <b>83.9%</b> | 25.4%                 | <b>74.6%</b> | 26.3%                 | <b>73.7%</b> | 19.8%                 | <b>80.2%</b> |
|                   | GraphRAG <sub>2</sub> | SceneRAG     | GraphRAG <sub>2</sub> | SceneRAG     | GraphRAG <sub>2</sub> | SceneRAG     | GraphRAG <sub>2</sub> | SceneRAG     |
|                   | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     |
| Comprehensiveness | 18.5%                 | <b>81.5%</b> | 26.8%                 | <b>73.2%</b> | 34.8%                 | <b>65.2%</b> | 23.1%                 | <b>76.9%</b> |
| Empowerment       | 11.7%                 | <b>88.3%</b> | 21.5%                 | <b>78.5%</b> | 29.5%                 | <b>70.5%</b> | 16.9%                 | <b>83.1%</b> |
| Trustworthiness   | 20.5%                 | <b>79.5%</b> | 18.0%                 | <b>82.0%</b> | 30.8%                 | <b>69.2%</b> | 21.9%                 | <b>78.1%</b> |
| Depth             | 16.4%                 | <b>83.6%</b> | 20.2%                 | <b>79.8%</b> | 30.8%                 | <b>69.2%</b> | 19.8%                 | <b>80.2%</b> |
| Density           | 39.0%                 | <b>61.0%</b> | 58.8%                 | <b>41.2%</b> | 60.7%                 | <b>39.3%</b> | 46.8%                 | <b>53.2%</b> |
| Overall Winner    | 14.4%                 | <b>85.6%</b> | 20.6%                 | <b>79.4%</b> | 30.4%                 | <b>69.6%</b> | 18.5%                 | <b>81.5%</b> |
|                   | LightRAG              | SceneRAG     | LightRAG              | SceneRAG     | LightRAG              | SceneRAG     | LightRAG              | SceneRAG     |
|                   | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     |
| Comprehensiveness | 22.6%                 | <b>77.4%</b> | 24.1%                 | <b>75.9%</b> | 25.9%                 | <b>74.1%</b> | 23.5%                 | <b>76.5%</b> |
| Empowerment       | 17.2%                 | <b>82.8%</b> | 23.7%                 | <b>76.3%</b> | 21.4%                 | <b>78.6%</b> | 19.2%                 | <b>80.8%</b> |
| Trustworthiness   | 27.5%                 | <b>72.5%</b> | 22.8%                 | <b>77.2%</b> | 29.5%                 | <b>70.5%</b> | 27.0%                 | <b>73.0%</b> |
| Depth             | 20.1%                 | <b>79.9%</b> | 21.1%                 | <b>78.9%</b> | 22.8%                 | <b>77.2%</b> | 20.8%                 | <b>79.2%</b> |
| Density           | 47.1%                 | <b>52.9%</b> | 56.6%                 | <b>43.4%</b> | 48.2%                 | <b>51.8%</b> | 49.1%                 | <b>50.9%</b> |
| Overall Winner    | 20.2%                 | <b>79.8%</b> | 20.6%                 | <b>79.4%</b> | 23.2%                 | <b>76.8%</b> | 20.8%                 | <b>79.2%</b> |
|                   | VideoRAG              | SceneRAG     | VideoRAG              | SceneRAG     | VideoRAG              | SceneRAG     | VideoRAG              | SceneRAG     |
|                   | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     |
| Comprehensiveness | 38.7%                 | <b>61.3%</b> | 36.8%                 | <b>63.2%</b> | 42.0%                 | <b>58.0%</b> | 39.0%                 | <b>61.0%</b> |
| Empowerment       | 36.2%                 | <b>63.8%</b> | 34.2%                 | <b>65.8%</b> | 39.7%                 | <b>60.3%</b> | 36.5%                 | <b>63.5%</b> |
| Trustworthiness   | 37.8%                 | <b>62.2%</b> | 34.2%                 | <b>65.8%</b> | 38.4%                 | <b>61.6%</b> | 37.2%                 | <b>62.8%</b> |
| Depth             | 37.1%                 | <b>62.9%</b> | 32.0%                 | <b>68.0%</b> | 38.8%                 | <b>61.2%</b> | 36.5%                 | <b>63.5%</b> |
| Density           | 50.9%                 | <b>49.1%</b> | 53.1%                 | <b>46.9%</b> | 58.0%                 | <b>42.0%</b> | 52.7%                 | <b>47.3%</b> |
| Overall Winner    | 37.2%                 | <b>62.8%</b> | 32.5%                 | <b>67.5%</b> | 40.6%                 | <b>59.4%</b> | 37.0%                 | <b>63.0%</b> |

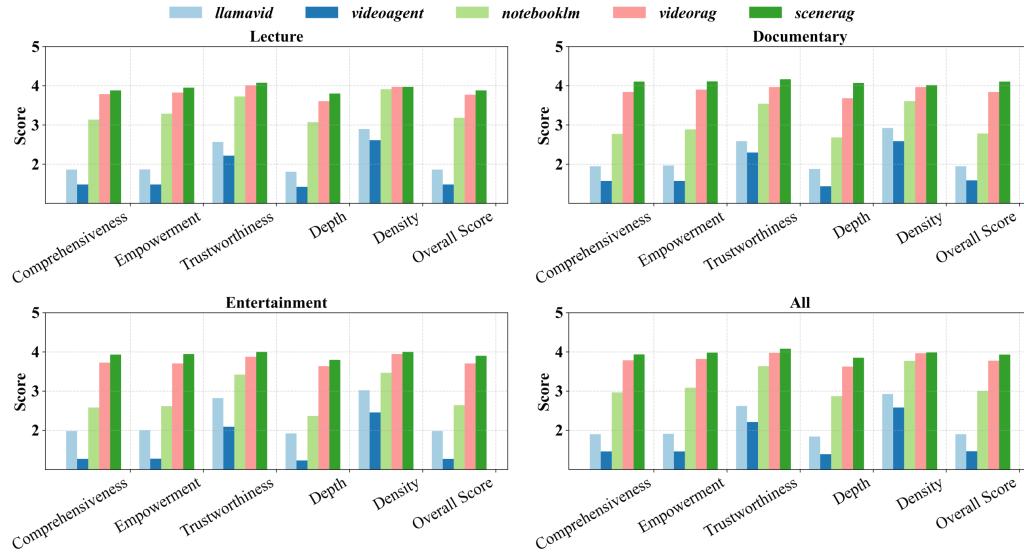


Figure 6: Quantitative Comparison by gpt-4.1-mini.

Table 8: Comparison of Winning Rates by GPT-4.1-nano.

|                   | Lecture               |              | Documentary           |              | Entertainment         |              | All                   |              |
|-------------------|-----------------------|--------------|-----------------------|--------------|-----------------------|--------------|-----------------------|--------------|
|                   | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     | NaiveRAG              | SceneRAG     |
| Comprehensiveness | 39.5%                 | <b>60.5%</b> | 32.5%                 | <b>67.5%</b> | 37.9%                 | <b>62.1%</b> | 37.9%                 | <b>62.1%</b> |
| Empowerment       | 39.2%                 | <b>60.8%</b> | 30.3%                 | <b>69.7%</b> | 37.1%                 | <b>62.9%</b> | 37.1%                 | <b>62.9%</b> |
| Trustworthiness   | 39.8%                 | <b>60.2%</b> | 28.1%                 | <b>71.9%</b> | 38.8%                 | <b>61.2%</b> | 37.4%                 | <b>62.6%</b> |
| Depth             | 38.0%                 | <b>62.0%</b> | 26.3%                 | <b>73.7%</b> | 34.8%                 | <b>65.2%</b> | 35.2%                 | <b>64.8%</b> |
| Density           | 50.0%                 | <b>50.0%</b> | 56.1%                 | <b>43.9%</b> | 53.6%                 | <b>46.4%</b> | 51.8%                 | <b>48.2%</b> |
| Overall Winner    | 39.0%                 | <b>61.0%</b> | 29.4%                 | <b>70.6%</b> | 36.2%                 | <b>63.8%</b> | 36.6%                 | <b>63.4%</b> |
|                   | GraphRAG <sub>1</sub> | SceneRAG     | GraphRAG <sub>1</sub> | SceneRAG     | GraphRAG <sub>1</sub> | SceneRAG     | GraphRAG <sub>1</sub> | SceneRAG     |
|                   |                       |              |                       |              |                       |              |                       |              |
| Comprehensiveness | 38.8%                 | <b>61.2%</b> | 41.7%                 | <b>58.3%</b> | 43.3%                 | <b>56.7%</b> | 40.2%                 | <b>59.8%</b> |
| Empowerment       | 36.2%                 | <b>63.8%</b> | 38.2%                 | <b>61.8%</b> | 39.3%                 | <b>60.7%</b> | 37.1%                 | <b>62.9%</b> |
| Trustworthiness   | 34.2%                 | <b>65.8%</b> | 38.2%                 | <b>61.8%</b> | 37.5%                 | <b>62.5%</b> | 35.5%                 | <b>64.5%</b> |
| Depth             | 32.7%                 | <b>67.3%</b> | 37.3%                 | <b>62.7%</b> | 35.3%                 | <b>64.7%</b> | 34.1%                 | <b>65.9%</b> |
| Density           | 39.2%                 | <b>60.8%</b> | 48.7%                 | <b>51.3%</b> | 45.5%                 | <b>54.5%</b> | 42.2%                 | <b>57.8%</b> |
| Overall Winner    | 33.5%                 | <b>66.5%</b> | 37.7%                 | <b>62.3%</b> | 36.6%                 | <b>63.4%</b> | 34.9%                 | <b>65.1%</b> |
|                   | GraphRAG <sub>2</sub> | SceneRAG     | GraphRAG <sub>2</sub> | SceneRAG     | GraphRAG <sub>2</sub> | SceneRAG     | GraphRAG <sub>2</sub> | SceneRAG     |
|                   |                       |              |                       |              |                       |              |                       |              |
| Comprehensiveness | 36.4%                 | <b>63.6%</b> | 38.2%                 | <b>61.8%</b> | 43.8%                 | <b>56.2%</b> | 38.1%                 | <b>61.9%</b> |
| Empowerment       | 33.5%                 | <b>66.5%</b> | 36.4%                 | <b>63.6%</b> | 42.9%                 | <b>57.1%</b> | 35.8%                 | <b>64.2%</b> |
| Trustworthiness   | 27.5%                 | <b>72.5%</b> | 31.1%                 | <b>68.9%</b> | 30.8%                 | <b>69.2%</b> | 28.8%                 | <b>71.2%</b> |
| Depth             | 26.9%                 | <b>73.1%</b> | 31.6%                 | <b>68.4%</b> | 33.5%                 | <b>66.5%</b> | 29.0%                 | <b>71.0%</b> |
| Density           | 36.4%                 | <b>63.6%</b> | 53.5%                 | <b>46.5%</b> | 48.7%                 | <b>51.3%</b> | 41.9%                 | <b>58.1%</b> |
| Overall Winner    | 28.9%                 | <b>71.1%</b> | 36.0%                 | <b>64.0%</b> | 37.1%                 | <b>62.9%</b> | 31.7%                 | <b>68.3%</b> |
|                   | LightRAG              | SceneRAG     | LightRAG              | SceneRAG     | LightRAG              | SceneRAG     | LightRAG              | SceneRAG     |
|                   |                       |              |                       |              |                       |              |                       |              |
| Comprehensiveness | 36.7%                 | <b>63.3%</b> | 35.1%                 | <b>64.9%</b> | 37.5%                 | <b>62.5%</b> | 36.5%                 | <b>63.5%</b> |
| Empowerment       | 33.6%                 | <b>66.4%</b> | 33.3%                 | <b>66.7%</b> | 35.7%                 | <b>64.3%</b> | 34.0%                 | <b>66.0%</b> |
| Trustworthiness   | 31.8%                 | <b>68.2%</b> | 32.9%                 | <b>67.1%</b> | 32.1%                 | <b>67.9%</b> | 32.1%                 | <b>67.9%</b> |
| Depth             | 30.9%                 | <b>69.1%</b> | 29.8%                 | <b>70.2%</b> | 29.9%                 | <b>70.1%</b> | 30.5%                 | <b>69.5%</b> |
| Density           | 46.5%                 | <b>53.5%</b> | 56.1%                 | <b>43.9%</b> | 47.3%                 | <b>52.7%</b> | 48.5%                 | <b>51.5%</b> |
| Overall Winner    | 31.6%                 | <b>68.4%</b> | 32.0%                 | <b>68.0%</b> | 33.0%                 | <b>67.0%</b> | 32.0%                 | <b>68.0%</b> |
|                   | VideoRAG              | SceneRAG     | VideoRAG              | SceneRAG     | VideoRAG              | SceneRAG     | VideoRAG              | SceneRAG     |
|                   |                       |              |                       |              |                       |              |                       |              |
| Comprehensiveness | 46.7%                 | <b>53.3%</b> | 38.6%                 | <b>61.4%</b> | 42.0%                 | <b>58.0%</b> | 44.3%                 | <b>55.7%</b> |
| Empowerment       | 46.3%                 | <b>53.7%</b> | 40.4%                 | <b>59.6%</b> | 43.8%                 | <b>56.2%</b> | 44.7%                 | <b>55.3%</b> |
| Trustworthiness   | 44.9%                 | <b>55.1%</b> | 39.0%                 | <b>61.0%</b> | 40.6%                 | <b>59.4%</b> | 43.0%                 | <b>57.0%</b> |
| Depth             | 44.7%                 | <b>55.3%</b> | 39.0%                 | <b>61.0%</b> | 41.1%                 | <b>58.9%</b> | 42.9%                 | <b>57.1%</b> |
| Density           | 46.9%                 | <b>53.1%</b> | 48.7%                 | <b>51.3%</b> | 49.6%                 | <b>50.4%</b> | 47.8%                 | <b>52.2%</b> |
| Overall Winner    | 45.2%                 | <b>54.8%</b> | 39.0%                 | <b>61.0%</b> | 41.5%                 | <b>58.5%</b> | 43.4%                 | <b>56.6%</b> |

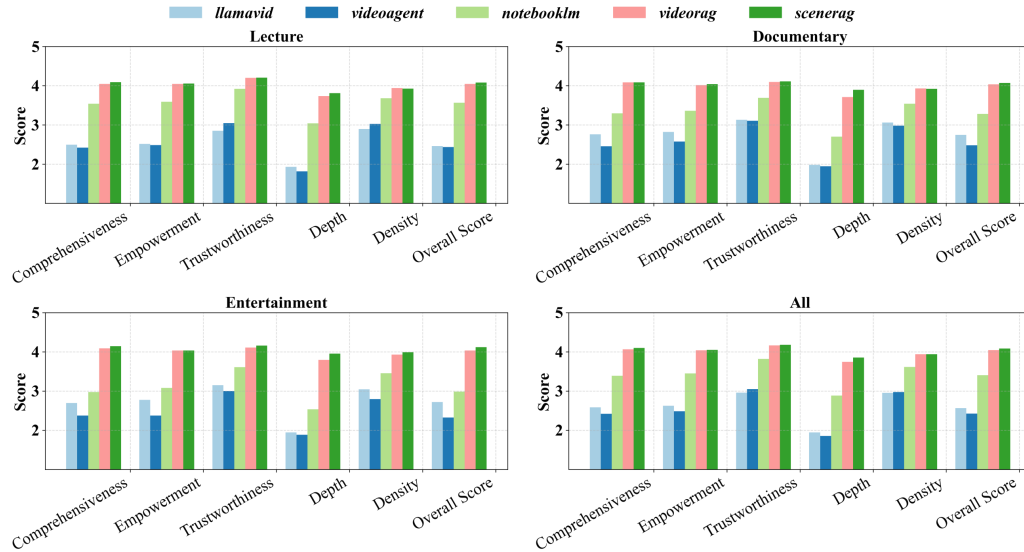


Figure 7: Quantitative Comparison by gpt-4.1-nano.

## D Dataset Descriptions

The LongerVideos dataset [36] is designed to evaluate models on long-form video comprehension and knowledge extraction. It comprises diverse video lists sourced primarily from YouTube, including online courses and thematic compilations, with durations ranging from just a few minutes to several hours. Each list is paired with a set of open-ended questions that require integrating information across multiple videos. Videos were collected using yt-dlp and questions were curated with assistance from advanced multi-video QA tools. In total, the dataset contains 22 curated video lists, facilitating comprehensive assessment of a model’s ability to synthesize, reason, and provide accurate responses over extended video content. Detailed dataset statistics can be found in Table 9.

Table 9: Detailed statistics of the *LongerVideos* dataset.

| Video Type    | video list name                        | #video | #query | overall duration |
|---------------|--|--------|--------|------------------|
| Lecture       | climate-week-at-columbia-engineering   | 4      | 26     | 5.91 hours       |
|               | rag-lecture                            | 19     | 38     | 5.34 hours       |
|               | ai-agent-lecture                       | 39     | 45     | 9.35 hours       |
|               | daubechies-wavelet-lecture             | 4      | 25     | 8.97 hours       |
|               | daubechies-art-and-mathematics-lecture | 4      | 21     | 4.87 hours       |
|               | tech-ceo-lecture                       | 4      | 31     | 4.83 hours       |
|               | dspy-lecture                           | 9      | 38     | 4.22 hours       |
|               | trading-for-beginners                  | 2      | 23     | 4.11 hours       |
|               | ahp-superdecision                      | 11     | 24     | 2.40 hours       |
|               | decision-making-science                | 4      | 26     | 2.20 hours       |
|               | 12-days-of-openai                      | 12     | 35     | 3.43 hours       |
|               | autogen                                | 23     | 44     | 8.70 hours       |
| Documentary   | fights-in-animal-kingdom               | 1      | 11     | 3.00 hours       |
|               | nature-scenes                          | 1      | 17     | 3.98 hours       |
|               | education-united-nations               | 6      | 39     | 8.41 hours       |
|               | elon-musk                              | 1      | 13     | 8.63 hours       |
|               | jeff-bezos                             | 3      | 34     | 4.47 hours       |
| Entertainment | black-myth-wukong                      | 10     | 23     | 21.36 hours      |
|               | primetime-emmy-awards                  | 3      | 17     | 7.31 hours       |
|               | journey-through-china                  | 1      | 27     | 3.37 hours       |
|               | fia-awards                             | 1      | 27     | 3.02 hours       |
|               | game-awards                            | 2      | 18     | 6.73 hours       |

## E Details of Case Study

To evaluate SceneRAG’s effectiveness in segmented retrieval and synthesis, we conducted a case study using the long-form technical lecture “*Is This the End of RAG? Anthropic’s NEW Prompt Caching*”. The core query—“How does prompt caching compare to traditional RAG in terms of cost and efficiency?”—requires aggregating evidence distributed across the video timeline. SceneRAG’s discourse-aware segmentation algorithm identified two non-contiguous yet complementary scenes ([23.96–66.08,s] and [213.12–292.48,s]), which together serve as a comprehensive and focused basis for answering the query, as summarized in Table 11. The first segment introduces the motivation and cost/latency benefits of prompt caching, while the second presents quantitative comparisons (e.g., “up to 80–90% reduction in both latency and cost”) and a trade-off analysis between Anthropic and Gemini implementations, detailed in Table 12.

For transparency, SceneRAG highlights key evidence at specific timestamps—24,s, 40,s, 214,s, and 275,s—which directly correspond to the main arguments in the retrieved scenes. This approach ensures that the synthesized answer is both contextually rich and easily traceable to the original material. The case study thus demonstrates SceneRAG’s advantage in precise, context-preserving retrieval for complex technical queries.

Table 10: A case study evaluating SceneRAG’s segmented retrieval and synthesis capabilities on long-form video. The system is tested on the lecture “Is This the End of RAG? Anthropic’s NEW Prompt Caching,” using a query focused on cost and efficiency comparisons between prompt caching and traditional RAG. Color legend: **green** = Cost-related; **Blue** = Efficiency/latency-related.

**Query: How does prompt caching compare to traditional RAG in terms of cost and efficiency?**

**Input Video: Is This the End of RAG? Anthropic’s NEW Prompt Caching (Lecture)**

### SceneRAG Response:

#### ### Overview of Prompt Caching vs. Traditional RAG

Prompt caching and traditional Retrieval-Augmented Generation (RAG) are both techniques designed to improve the efficiency and effectiveness of language models. However, they differ significantly in terms of cost efficiency, processing time, and usage scenarios.

#### Cost Efficiency

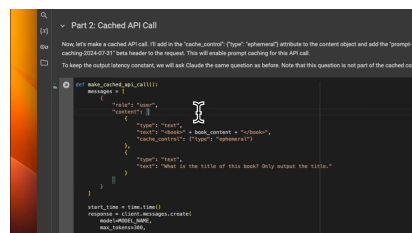
One of the primary advantages of prompt caching is its remarkable cost-effectiveness. For instance, prompt caching can reduce costs by up to 90% compared to traditional methods when dealing with extensive prompts or multiple interactions. This reduction is attributed to the fact that cached tokens cost significantly less—approximately 10% of the base input token price—thus leading to substantial savings during multiple API calls (Source: Video on Prompt Caching). In comparison, traditional RAG systems can incur higher costs, especially when a large number of tokens or documents are fetched repeatedly from a database. Although RAG systems allow for the retrieval of only the most relevant information based on specific queries, the contextual embeddings and processing involved can lead to increased operational expenses over time, mainly due to the need for constant storage and retrieval of data (Source: Video on RAG systems).

#### Efficiency and Processing Time

When it comes to latency and processing times, prompt caching also shows impressive results. For example, when handling prompts of 100000 tokens, the response time can drop significantly from about 12 seconds without caching to approximately 2.4 seconds with caching—representing an 80% reduction in processing time. Similarly, even in many-shot prompting scenarios, prompt caching demonstrates a latency reduction of about 31% (Source: Videos detailing prompt caching benefits). Traditional RAG methods tend to have longer response times. RAG typically requires an initial retrieval step, which can introduce delays, particularly when large datasets are involved. The complexity involved in retrieving, processing, and generating coherent responses from various chunks of data can result in prolonged wait times for end-users (Source: Video on RAG systems). (The remaining content are omitted for brevity.)

SceneRAG precisely retrieved two non-contiguous, highly relevant video segments from the lecture *Is This the End of RAG? Anthropic’s NEW Prompt Caching*, covering both the motivation and technical comparisons between prompt caching and traditional RAG. Here, we highlight the main evidence aligned with the answer at key timestamps: **24 s**, **40 s**, **214 s**, and **275 s**.

From left to right, these correspond to: the introduction of prompt caching’s cost/latency benefits (24 s, 40 s), and concrete quantitative comparisons, including 80–90% cost and latency reduction and cross-system trade-offs (214 s, 275 s). These moments collectively supply a focused, complete answer to the cost and efficiency comparison query.



| Use case  | Latency w/o caching (time to first token) | Latency w/ caching (time to first token) | Cost reduction |
|---|---|--|----------------|
| Chat with a book (100,000 token cached prompt) (1)                    | 11.5s                                     | 2.4s (-79%)                              | -90%           |
| Many-shot prompting (10,000 token prompt) (1)                         | 1.6s                                      | 1.1s (-31%)                              | -86%           |
| Multi-turn conversation (10-turn convo with a long system prompt) (2) | ~10s                                      | ~2.5s (-75%)                             | -53%           |

How we price cached prompts

Table 11: Segmented Scenes and Transcripts in the Video

| Scene and Transcript Content  |
|---|
| <p><b>[0.00s,23.96s]</b></p> <p>So Anthropic just introduced prompt caching with Cloud. That can reduce cost up to by 90% and latency up to by 85, which is huge. And did they just kill Rag with this new feature? Now Google was the first one who introduced context caching with their Gemini models. There are some similarities, but some major differences as well between these two approaches. We will discuss them later in the video. I'll show you how to get started. And what type of performance difference you can expect. Before looking at the quote example, let's see what was released today. Prompt caching enables developers to cache frequently used contacts between API calls.</p> <p><b>[23.96s,66.08s]</b></p> <p>Anthropic models have a huge context window of 200,000 tokens. However, if you're chatting with long documents, you'll have to send them with each prompt. So that becomes very expensive. And hence, the prompt caching is going to be extremely helpful. So now customers can provide Cloud with more background information and example outputs or few short prompting. Reducing cost by up to 90% and latency up to 85%. Now, these numbers are great, but they are not going to be consistent based on the example use cases. And we are going to look at some of them. This feature is available both for Cloud 3.5 Sonnet and Cloud 3 Haiku. Support for Cloud 3 opus is coming soon.</p> <p><b>[66.08s,134.22s]</b></p> <p>As I said in the beginning, context caching has been available for the Gemini models. And there are some interesting differences between the two, which I'm going to highlight throughout the video. So what are going to be some use cases for prompt caching? Well, the first one is conversational agents. So if you're having a long-form conversation and there is a substantial chat history, you can put that chat history in the cache and just ask questions from that. Another example, use case is coding assistants. Usually code bases are pretty huge. So you can put them in the prompt cache and then use your subsequent messages for question answer. Also, launch document processing or detailed instruction sets. This specifically will apply if you have highly detailed system prompt with a lot of few short examples. So this is going to be very helpful that you can just send those ones and then you can have subsequent conversations while this is cached.</p> <p><b>[134.22s,213.12s]</b></p> <p>A genetic search and tool usage is another example, especially if you have to define your tools, what inputs are to different tools so you can put them in your prompt cache and then send that once and that will save you a lot of money. And that example is going to be talked to books, papers, documentation, podcast, transcripts and other long form content. So this is a very enticing application for Rack and with these long context models, especially with prompt caching or context caching, now it becomes viable to just put these documents in the context rather than chunking them, computing embedding, and then doing retrieval on the documents. Now here's a table that shows what type of reduction in cost and latency you can expect for different applications.</p> <p><b>[213.12s,300.00s]</b></p> <p>If you're chatting with your documents and you're sending 100,000 tokens without caching, that would take about 12 seconds to generate a response, but with caching, that's about only 2.4 or 2.5 seconds, which is 80% reduction in processing time on latency and 90% reduction in the cost. If you're doing a few shot prompting with 10,000 tokens, you expect about 31% reduction in latency and about 86% reduction in cost. Whereas if you're doing multi-turn conversation, a 10-turn conversation, you're expecting about 75% reduction in latency, but only about 53% reduction in cost. Now, the way the cash to tokens are charged versus the input output tokens are different and that's why you see these reductions in the cost as well. Now, we saw the cost reduction because the cash tokens are costing only 10% of the base input token price, which is a huge reduction of 90%. However, you also need to keep in mind that writing to the cash costs about 25% more than the base input token price for any given model. So there is an overhead when you have a writing to the cash for the first time, but then there is a substantial reduction in cost. Now, the Gemini models do it in a different way. There is no cost associated with the actual cash token, but there is a storage cost of \$1 per million tokens per hour. Okay, so here's what the reduction is going to look like.</p> |

Table 12: Scene Transcripts and Captions of the Video

| Scene and Transcript Content   | Scene Description (Caption)  |
|--|--|
| <p><b>[23.96s,66.08s]</b></p> <p>Anthropic models have a huge context window of 200,000 tokens. However, if you're chatting with long documents, you'll have to send them with each prompt. So that becomes very expensive. And hence, the prompt caching is going to be extremely helpful. So now customers can provide Cloud with more background information and example outputs or few short prompting. Reducing cost by up to 90% and latency up to 85%. Now, these numbers are great, but they are not going to be consistent based on the example use cases. And we are going to look at some of them. This feature is available both for Cloud 3.5 Sonnet and Cloud 3 Haiku. Support for Cloud 3 opus is coming soon.</p>  | <p>The video presents a tutorial on using prompt caching with Anthropic's AI models, focusing particularly on reducing costs and improving response latency. It begins by showcasing the integration of prompt caching into an API call script within a development environment, highlighting the use of <code>cache_control</code> headers to enable this feature. The video then transitions to a web page detailing how prompt caching works, its benefits, and its availability in public beta for specific versions of the models (Claude 3 Sonnet and Claude 3 Haiku). As the narration explains these points, the webpage is displayed with clear visuals such as icons and text that support the explanation. The video emphasizes the practical applications of prompt caching, including its effectiveness in situations requiring large amounts of context, enhancing background knowledge, and optimizing performance through reduced costs and faster responses. Throughout, the focus remains on educating viewers about the implementation and advantages of prompt caching without showing any human subjects or interactions beyond the digital interface being demonstrated.</p>   |
| <p><b>[213.12s,300.00s]</b></p> <p>If you're chatting with your documents and you're sending 100,000 tokens without caching, that would take about 12 seconds to generate a response, but with caching, that's about only 2.4 or 2.5 seconds, which is 80% reduction in processing time on latency and 90% reduction in the cost. If you're doing a few shot prompting with 10,000 tokens, you expect about 31% reduction in latency and about 86% reduction in cost. Whereas if you're doing multi-turn conversation, a 10-turn conversation, you're expecting about 75% reduction in latency, but only about 53% reduction in cost. Now, the way the cash to tokens are charged versus the input output tokens are different and that's why you see these reductions in the cost as well. Now, we saw the cost reduction because the cash tokens are costing only 10% of the base input token price, which is a huge reduction of 90%. However, you also need to keep in mind that writing to the cash costs about 25% more than the base input token price for any given model. So there is an overhead when you have a writing to the cash for the first time, but then there is a substantial reduction in cost. Now, the Gemini models do it in a different way. There is no cost associated with the actual cash token, but there is a storage cost of \$1 per million tokens per hour. Okay, so here's what the reduction is going to look like.</p> | <p>The video presents a detailed comparison of prompt caching in natural language processing, focusing on the reduction in latency and cost when using cached content. It begins by displaying a table with three use cases: chatting with documents, many-shot prompting, and multi-turn conversations, each showing reduced latency times ranging from 12 seconds to over 30 seconds without caching, down to approximately 2.5 seconds with caching, resulting in up to an 86% cost reduction. The visual transitions smoothly through various sections of text, highlighting key points about pricing based on token counts, input/output tokens, and storage costs. The video explains that writing to cache incurs additional charges, but using cached content significantly lowers these costs, often costing only 10% of the base input token price. This information is reinforced through a dark-themed interface showcasing rate limits, context caching, and output prices for different models like Claude 3.5 Sonnet, Claude 3 Opus, and Gemini models. Each model's details are presented with specific pricing structures, emphasizing how the cost varies depending on factors such as the number of tokens, input/output prompts, and storage requirements. Throughout the video, the cursor moves across the screen, pointing out significant figures and changes in the text, ensuring viewers understand the financial implications of prompt caching. The consistent transition between frames keeps the focus on the educational content, providing a clear overview of how prompt caching can optimize performance while minimizing expenses.</p> |