Generative Frame Sampler for Long Video Understanding

Linli Yao[1], Haoning Wu[2], Kun Ouyang[1], Yuanxing Zhang[2],
Caiming Xiong[3], Bei Chen[4], Xu Sun[1]*, Junnan Li[3]*,
[1]National Key Laboratory for Multimedia Information Processing,
School of Computer Science, Peking University
[2]Peking University, [3]Salesforce Research, [4]Independent Researcher
{linliyao,kunouyang}@stu.pku.edu.cn {longo,xusun}@pku.edu.cn
{realtimothyhwu,beibei1019}@gmail.com {cxiong,junnan.li}@salesforce.com

Abstract

Despite recent advances in Video Large Language Models (VideoLLMs), effectively understanding long-form videos remains a significant challenge. Perceiving lengthy videos containing thousands of frames poses substantial computational burden. To mitigate this issue, this paper introduces **Gen**erative Frame **S**ampler (**GenS**), a plug-and-play module integrated with VideoLLMs to facilitate efficient lengthy video perception. Built upon a lightweight VideoLLM, GenS leverages its inherent vision-language capabilities to identify question-relevant frames. To facilitate effective retrieval, we construct **GenS-Video-150K**, a large-scale video instruction dataset with dense frame relevance annotations. Extensive experiments demonstrate that GenS consistently boosts the performance of various VideoLLMs, including open-source models (Qwen2-VL-7B, Aria-25B, VILA-40B, LLaVA-Video-7B/72B) and proprietary assistants (GPT-4o, Gemini). When equipped with GenS, open-source VideoLLMs achieve impressive state-of-the-art results on long-form video benchmarks: LLaVA-Video-72B reaches 66.8 (+4.3) on LongVideoBench and 77.0 (+2.7) on MLVU, while Aria obtains 39.2 on HourVideo surpassing the Gemini-1.5-pro by 1.9 points. We release the code, dataset and models at https://generative-sampler.github.io.

1 Introduction

Recent advances in Large Multimodal Models (LMMs) (Li et al., 2023a; Dai et al., 2023; Wang et al., 2024a; Liu et al., 2023, 2024a; Chen et al., 2023; Tong et al.) have shown remarkable progress, yet understanding long videos remains a significant challenge (Zohar et al., 2024; Li et al., 2024b). Current video-oriented LMMs (Zhang et al., 2024b, 2023; Li et al., 2023b, 2025; Zhang et al., 2024a), known as VideoLLMs, typically employ an image encoder such as CLIP (Radford et al., 2021) or SigLIP (Zhai et al., 2023) to encode individual video frames as an initial step in video perception. When processing hours-long videos containing thousands of frames, a critical challenge emerges: *how to efficiently sample representative frames from the original video sequence?*

Existing VideoLLM assistants primarily employ two approaches for sampling lengthy videos: 1) *uniform sampling* based on the VideoLLM's maximum context length, which leads to significant visual information loss due to limited fixed-interval sampling; 2) *frame-per-second (FPS) sampling*, as implemented in long-context models like Gemini (Gemini Team, 2024), which can capture frames at 1 FPS for comprehensive visual coverage. However, it obtains thousands of frames for hours-long videos, resulting in booming memory consumption and slow inference speed.

Intuitively, for VideoLLM assistants, most frames in long videos are redundant when addressing a specific user instruction (i.e., *query*). To mitigate visual redundancy, several works propose language-guided frame sampling via CLIP to retrieve query-aware frames efficiently (Arefeen et al., 2024; Wang et al., 2024c,b). However, CLIP-based frame samplers have three major limitations. For visual side, its frame-by-frame matching fails to capture temporal relationships implied by successive frames, as depicts in Figure 1 (a). For textual side, it is constrained by limited language capabilities, only able to process concise and simple user queries. Additionally, it embeds frames and textual queries separately to calculate cosine similarity, which hinders sufficient vision-language interaction to achieve complex multi-hop reasoning (Fu et al., 2024; Wu et al., 2024a; Chandrasegaran et al., 2024).

To mitigate these limitations, we present **Gen**erative Frame **S**ampler (**GenS**), a VideoLLM-based approach to retrieve relevant frames through

*Corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2025, pages 17900–17917

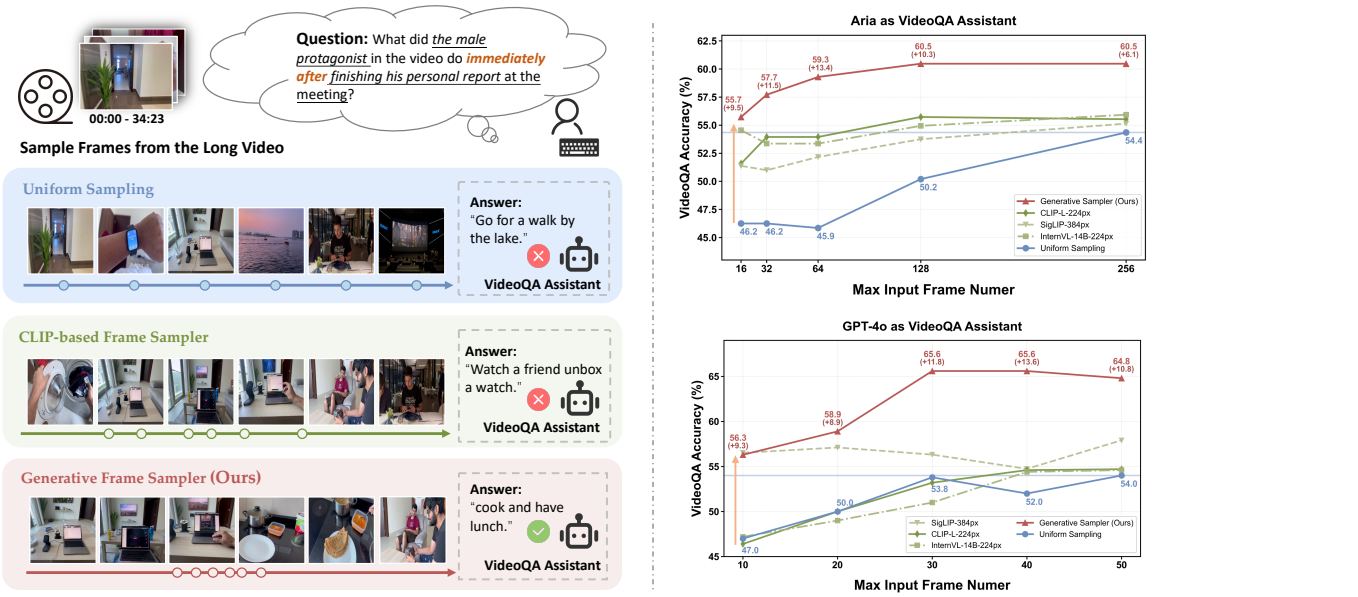

Figure 1: (a) An example of long video question-answering (VideoQA) using different frame samplers. Our Generative Frame Sampler (GenS) accurately identifies relevant frame sequences based on the user question, further enhancing the performance of the downstream VideoQA assistant. (b) VideoQA accuracy results of state-of-the-art VideoQA assistants (Aria (Li et al., 2025) and GPT-4o (OpenAI, 2024)) when equipped with different frame samplers on the *Vision-Centric subset* of LongVideoBench (Wu et al., 2024a) .

flexible user instructions. Built upon an advanced long-context VideoLLM (Li et al., 2025), our approach inherits fundamental video-language perception capabilities. **First**, as Figure 1 (a) illustrates, GenS effectively captures temporal relationships between successive frames, such as "*immediately after*". **Second**, powered by built-in LLMs (Dubey et al., 2024; Chiang et al., 2023), GenS comprehends complex and flexible textual instructions. **Third**, its native multi-modal architecture enables complex multi-hop reasoning by aligning long-range temporal cues with language semantics. As demonstrated in Figure 1 (b), by selecting more relevant visual frames, GenS substantially enhances the performance of VideoQA Assistants across both open-source (Aria (Li et al., 2025)) and proprietary (GPT-4o (OpenAI, 2024)) models. Compared with uniform sampling, GenS improves Aria's accuracy by 13.4 points ($\leq$64 frames) and GPT-4o's accuracy by 13.6 points ($\leq$40 frames) on the challenging long-form video benchmark (Wu et al., 2024a). These significant improvements highlight that efficient video perception is a critical bottleneck for modern VideoQA Assistants, and GenS provides an practical solution to boost their full potential.

To develop the GenS sampler, we address two primary challenges: Firstly, there is a *shortage of training data*, as existing video instruction datasets (Zhang et al., 2024b; Maaz et al., 2024; Liu et al.) lack dense annotations of relevant frames across diverse videos and user instructions. Secondly, the optimal *generative format for relevant frames sampling* remains under-explored. To address the first challenge, we introduce **GenS-Video-150K**, a novel synthetic VideoQA dataset with question-relevant frame annotations via GPT-4o. The relevant frame annotations are: 1) *dense*, with 20% of all frames annotated, and 2) *fine-grained*, with specific confidence scores (levels 1 to 5) assigned to each relevant frame. For the second challenge, we explore different generative formats for indexing relevant frames. Empirical results show that directly appending textual labels ("`Frame Number [N]`") before visual frames is sufficient to distinguish sequential frames. GenS outputs the relevant frame spans with confidence scores as a natural language generation task ({"`Frame` N_{start}-N_{end}`: relevance score`", ...}).

To summarize, our **main contributions** are threefold: 1) We propose GenS, a novel generative frame sampler that leverages VideoLLMs to identify question-aware relevant frames. It serves as a plug-and-play sampler that enhances input frames for VideoQA Assistants. 2) We introduce GenS-Video-150K, a large-scale video instruction dataset that densely annotates relevant frames with fine-grained confidence scores across diverse video

questions. 3) Through extensive experiments, we demonstrate that GenS significantly enhances the performance of both open-source (Qwen2VL, Aria, VILA-v1.5, LLaVA-Video) and proprietary (GPT-4o and Gemini-1.5-pro) VideoQA Assistants. Notably, when equipped with GenS, LLaVA-Video-72B achieves state-of-the-art performance with accuracy scores of 77.0 on MLVU and 66.8 on LongVideoBench, while Gemini1.5-pro attains 40.7 on HourVideo with averaging 45.7 minutes video duration.

2 Method

We introduce the novel GenS method that effectively selects instruction-aware frames from long-form videos. To address the challenge of insufficient training data, we first construct GenS-Video-150K, a video instruction dataset with dense relevant frame annotations (Section 2.1). We then present the GenS architecture, focusing on an efficient generative format for VideoLLM-based frame retrieval (Section 2.2). Finally, we demonstrate how to integrate GenS with existing VideoQA Assistants to enhance long-form video perception (Section 2.3).

2.1 GenS-Video-150K Dataset Collection

Our objective is to construct *(video, user instruction, relevant frames)* samples that enable the GenS to identify salient frames for user instructions. Existing datasets for grounded VideoQA (Bärmann and Waibel, 2022) and event localization (Anne Hendricks et al., 2017; Ren et al., 2024; Liu et al.; Wu et al., 2024b) are limited by their *domain specificity*, *naive instruction*, and *sparse key frame annotations*, make them inadequate for training robust frame samplers in real-world long-form video understanding. To address these limitations, we introduce GenS-Video-150K across diverse video topics and flexible user instructions, with two key features: 1) *dense* frame relevance annotations, with approximately 20% of frames marked as relevant, and 2) *fine-grained* scoring, where each relevant frame is assigned specific confidence scores (1-5).

We observe that even powerful proprietary LMMs like GPT-4o (OpenAI, 2024) struggle to achieve satisfactory retrieval performance when directly processing thousands of frames from lengthy videos (verified in Table 4). To ensure high dataset quality, we decompose the synthetic data creation into a carefully designed four-stage pipeline leveraging GPT-4o. All prompts are provided in the Appendix A.2.

Stage 1: Dense Video Frame Captioning. We first curate a diverse collection of videos from YT-Temporal-1B (Zellers et al., 2022), encompassing a broad range of topics from YouTube [1]. Inspired by prior works (Chen et al., 2024a), we generate differential paragraph captions for each frame at a dense sampling rate (0.2 fps), focusing on distinguishing new visual content from previous frames. This dense frame captioning approach has been widely adopted as a preliminary step in video instruction dataset construction (Chen et al., 2024a; Zhang et al., 2024b).

Stage 2: Construct Video QAs with Grounded Frames. In this stage, we generate 12 distinct types of video question-answer (QA) pairs with grounded frames based on the dense frame captions. Specifically, we prompt GPT-4o to analyze every 50 consecutive frame captions and generate QA pairs of assigned types. Frames referenced during QA generation are marked as grounded frames (detailed in Appendix A.2). To ensure robust generalization, we maintain a balanced distribution between generative and multiple-choice questions (50% each). For multiple-choice questions, we augment the retrieval query by incorporating candidate options alongside the user question. We also strategically include 1% negative samples (questions with no relevant frames) to enhance model robustness against irrelevant queries.

Stage 3: Extend Relevant Frames. We expand the set of relevant frames beyond those strictly grounded obtained in Stage 2. GPT-4o typically references only a small subset of frames during VideoQA generation, resulting in a *low retrieval ratio* $R_f = \frac{N_{grd}}{N_{total}}$, where N_{grd} represents the number of grounded frames and N_{total} denotes the total number of captioned frames. Training with such a small R_f would limit GenS's ability to provide comprehensive frame coverage for long-context video understanding. Therefore, we employ CLIP-based retrieval to increase R_f from 5% to 30% approximately to add candidate relevant frames.

Stage 4: Score Fine-grained Relevant Confidence. Finally, we score the relevance (from 0 to 5, 0 is non-relevant and 5 is the most relevant) among all candidate relevant frames given the video and user question. This stage accurately refines the

[1] https://www.youtube.com/

Table 1: Statistics of GenS-Video-150K dataset.

Property	Value
Format	{video, question, answer, scored relevant frames}
Total Samples	150K
Avg Video Duration	647.5 seconds (10.8 minutes)
Avg Task Number	12
Relevant Frame Rate	around 20%
Relevance Scores	0-5 (0: non-relevant, 5: most relevant)

relevant frames in a global view of the whole video. Moreover, it provides fine-grained supervision to distinguish multi-level relevances to enable a top-K retrieval by confidence score duration inference.

Data Statistics. Our four-stage pipeline yields 150K data samples, with representative examples shown in Appendix Figure 4. Table 1 provides a comprehensive overview of our dataset statistics. Each video has an average duration of 647.5 seconds (10.8 minutes) and contains approximately 129.5 captioned frames. Of these frames, an average of 20% are annotated as relevant with fine-grained confidence scores ranging from 0 to 5 (where 0 indicates non-relevant and 5 indicates most relevant), thereby providing dense supervision.

2.2 GenS Architecture

We design the GenS architecture based on the state-of-the-art Aria model (Li et al., 2025), which offers three key advantages: 1) As a native multimodal model, Aria demonstrates superior capability in understanding interleaved video-language contexts, enabling effective identification of relevant frames with textual indexing. 2) With a context window supporting up to 256 input frames, Aria's architecture excels at modeling temporal relationships implied in diverse user instructions. 3) Through its Mixture-of-Experts (MoE) architecture (Jiang et al., 2024; Krajewski et al., 2024) with 3.9B activated parameters, Aria strikes an optimal balance between inference efficiency and multimodal performance compared to conventional 7B-parameter VideoLLMs (Zhang et al., 2024b; Wang et al., 2024a; Chen et al., 2023; Lin et al., 2023).

2.2.1 Efficient Frame Indexing

Leveraging Aria's advanced capabilities in encoding interleaved visual-textual representations, we implement an efficient frame indexing mechanism by prepending each frame with a textual number `[N]` that denotes the `[N-th Frame]`. This enables GenS to uniquely identify and retrieve relevant frames based on their temporal positions.

For output representation, we adopt a JSON-based format similar to proprietary video assistants like Gemini and GPT-4o, where GenS generates frame relevance predictions as a language modeling task. The output schema flexibly accommodates both discrete frame annotations (e.g., `{"frame number": relevance score}`) and continuous temporal spans (e.g., `{"start frame - end frame": relevance score}`) based on the retrieval context. Our experiments in Section 3.5 demonstrate that organizing retrieved frames by relevance scores yields better performance compared to temporal ordering.

2.2.2 Adaptation for Various Input FPS

GenS is designed as a plug-and-play frame sampling module that seamlessly integrates with existing VideoQA Assistants (Zhang et al., 2024b; Wang et al., 2024a; Chen et al., 2023; Lin et al., 2023; Wang et al., 2024c). To handle varying candidate frame number and sampling densities across downstream VideoQA models, we implement a flexible frame retrieval mechanism that supports both dense (high FPS) and sparse (fixed-interval) frame sampling patterns. Specifically, we normalize the candidate frame indices to a unified range of 1-256 within each retrieval temporal window, ensuring robust retrieval performance regardless of the original frame sampling rate.

2.3 Training and Inference Paradigm

We train GenS on both our GenS-Video-150K and existing human-annotated event-level video datasets, specifically the E.T. Instruct dataset (Liu et al.), to enhance training data diversity. However, directly mixing E.T. Instruct data degrades GenS's performance due to its sparse grounded frame annotations. Therefore, we integrate and post-process the E.T. Instruct dataset to better align with our frame sampling task (details in Section 3.1).

During inference, GenS processes videos at arbitrary frame rates, retrieving instruction-relevant frames with confidence scores within each temporal window (maximum 256 frames). The retrieval across multiple temporal windows can be parallelized for efficient processing. Output relevant frames are naturally sorted by confidence scores, with the number of relevant frames N_{ret} varying based on the specific question and video content. We select the top K frames for input to a VideoQA model, where $K = \min(N_{ret}, N_{ctx})$, with N_{ctx} being the VideoQA model's maximum context length.

VideoQA Model	Size	Sampled Frames	LongVideoBench$_{val}$ (avg 12min)		MLVU$_{val}$ (avg 12min)	
			Full	*V-Centric*	*Full*	*V-Centric*
Proprietary LMMs						
GPT-4o	-	256/0.5fps	66.7	-	64.6	-
Gemini-1.5-Pro	-	256/256	64.0	-	-	-
Open-source Video LLMs						
LLaVA-Video	7B	64/64	58.9	50.0	70.4	66.9
LLaVA-Video $_{w/\ GenS}$	7B	54/50	63.3 (+4.4)	56.7 (+6.7)	73.4 (+3.0)	70.6 (+3.7)
Qwen2-VL	7B	64/64	56.0	45.9	64.7	62.3
Qwen2-VL $_{w/\ GenS}$	7B	54/50	58.7 (+2.7)	49.2 (+3.3)	66.9 (+2.2)	64.8 (+2.5)
Aria	25B-A3.9B	256/256	62.7	54.4	69.5	62.1
Aria $_{w/\ GenS}$	25B-A3.9B	54/95	66.1 (+3.4)	**59.3** (+4.9)	72.6 (+3.1)	67.5 (+5.4)
VILA-v1.5	40B	14/14	57.4	47.0	57.8	52.5
VILA-v1.5 $_{w/\ GenS}$	40B	14/14	59.6 (+2.2)	50.2 (+3.2)	63.5 (+5.7)	58.3 (+5.8)
LLaVA-Video	72B	64/64	62.5	51.6	74.3	72.5
LLaVA-Video $_{w/\ GenS}$	72B	54/50	**66.8** (+4.3)	58.9 (+7.3)	**77.0** (+2.7)	**74.1** (+1.6)

Table 2: Performance on LongVideoBench (Wu et al., 2024a) and MLVU (Zhou et al., 2024) benchmarks using multiple-choice accuracy metrics. *V-Centric* denotes a vision-centric subset containing questions that explicitly require video understanding rather than language-only reasoning, while filtering short videos. Sampled Frames *N/M* indicates sampled N frames for LongVideoBench and M frames for MLVU separately. Using GenS, we select the K most relevant frames (K <= max frame number of VideoQA models) and report the average number of input frames.

3 Experiments

3.1 Experimental Settings

Evaluation Benchmarks. We evaluate GenS on several long-form video benchmarks including LongVideoBench$_{val}$ (LVB) (Wu et al., 2024a), MLVU$_{Dev}$ (Zhou et al., 2024), and HourVideo (Chandrasegaran et al., 2024). These benchmarks assess multiple-choice question-answering accuracy on videos ranging from minutes to hours in duration. For LVB and MLVU, we construct a more challenging *Vision-Centric subset* by filtering out both questions answerable through pure language reasoning and videos of short duration. Additionally, we evaluate the zero-shot temporal grounding capability of GenS on the Charades-STA (Gao et al., 2017) dataset using mean Intersection over Union (mIoU) and Recall@1 at IoU thresholds of 0.3, 0.5, and 0.7.

Training Dataset. We utilize timestamp-output tasks from E.T.Instruct 164K (Liu et al.), extracting 75K base training samples (denoted as E.T. Instruct-75K). To enhance the density of grounded frame annotations, we post-process these samples through timestamp label aggregation and textual query concatenation within each video, yielding 41K samples (denoted as E.T. Instruct-41K$_{agg.}$). The final training data for GenS combines this aggregated E.T. Instruct dataset with GenS-Video-150K.

Implementation Details. We train the open-source Aria [2] model with a frozen vision encoder. The model supports a maximum sequence length of 32K tokens, accommodating up to 256 frames per sequence. Training consists of 300 iterations with a global batch size of 256, completed in 10 hours using 32 H800 GPUs. We provide complete hyperparameter settings in Appendix A.3. During inference, the MOE architecture of GenS utilizes only 3.9B activated parameters. We obtain original frames from the input video at 1 FPS to ensure comprehensive visual coverage, and then sample frames within each 256-frame interval using a sliding window approach. For multiple-choice questions, we append candidate options to the retrieval query.

3.2 Results on Long-form Video Tasks

GenS functions as a plug-and-play frame sampling module that enhances visual perception capabilities across VideoQA models. We evaluate its effectiveness across three categories of VideoQA models: I) *Advanced proprietary VideoLLMs*, specifically GPT-4o (OpenAI, 2024) and Gemini-1.5-pro (Gemini Team, 2024); II) *Open-source competitive VideoLLMs* with standard context lengths (64 input frames), including LLaVA-Video-7B/72B (Zhang et al., 2024b) and Qwen2-VL-7B (Wang et al., 2024a); III) *Open-source long-context VideoLLMs*, represented by Aria-25B (Li et al., 2025) with 256-frame input capacity.

For MLVU and LongVideoBench, we construct a more challenging *Vision-Centric subset* (also filtering out short videos) based on two observations: 1) several questions in the original datasets can be answered through pure language reasoning without visual context; 2) visual content in short videos can be adequately captured through uniform frame

[2] https://github.com/rhymes-ai/Aria

	Summarization			Perception				Visual Reasoning									Navigation		Avg
	Key Events/ Objects	Temporal Sequencing	Compare/ Contrast	Factual Recall	Sequence Recall	Temporal Distance	Tracking	Relationship	Proximity	Layout	Duration	Frequency	Pre-requisites	Predictive	Causal	Counterfactual	Room-to-Room	Object Retrieval	
Blind LLMs																			
GPT-4	22.7	29.6	24.2	21.9	15.2	20.6	15.8	14.9	21.4	22.2	23.6	19.3	14.7	14.5	18.7	21.2	15.8	18.8	19.6
Socratic Models																			
LLaVA-34B-DPO	34.0	35.5	35.8	30.3	19.3	12.7	34.5	18.3	15.3	26.7	21.3	17.9	23.5	20.9	21.3	22.4	20.8	22.4	22.3
GPT-4	40.5	41.5	43.2	33.1	20.0	20.2	36.7	18.5	21.7	37.8	25.3	22.9	27.1	24.1	24.7	26.5	20.0	26.6	25.7
Multimodal Models																			
Aria (256 frm.)	58.2	53.9	55.8	44.7	33.8	28.1	41.9	26.8	36.9	28.9	42.3	34.9	50.0	54.8	31.3	23.8	15.0	25.0	38.7
Aria $_{\text{w/ GenS}}$ (226 frm.)	56.3	54.6	50.5	44.5	34.7	26.6	45.3	26.8	38.3	26.7	42.7	36.0	50.8	56.8	33.3	29.1	15.8	22.9	39.2
Gemini1.5-Pro (0.5fps)	56.4	59.5	46.7	41.8	33.6	19.7	35.7	27.4	38.2	21.4	37.2	35.4	46.8	46.3	41.0	38.7	19.2	33.9	37.3
Gemini1.5-Pro $_{\text{w/ GenS}}$	57.6	57.9	53.7	45.3	34.7	26.2	45.8	31.7	39.2	15.6	39.8	40.8	48.9	49.4	48.7	37.1	29.2	34.9	**40.7**
(344 frm.)	+1.2	-1.6	+7.0	+3.5	+1.1	+6.5	+10.1	+4.3	+1.0	-5.8	+2.6	+5.4	+2.1	+3.1	+7.7	-1.6	+10.0	+1.0	+3.4

Table 3: Results on HourVideo (Chandrasegaran et al., 2024) benchmark, an extremely challenging video dataset with an average duration of 45.7 minutes, containing 113 videos longer than 60 minutes. *Blind LLMs* perform reasoning without video inputs. *Socratic models* first segment videos into one-minute intervals, generate captions for each segment using LLaVA-34B-DPO or GPT-4, then use GPT-4 to answer questions based on the aggregated captions. *Multimodal Models* directly process video inputs for inference.

sampling. Details are provided in Appendix A.3.

LongVideoBench. As shown in Table 2, GenS demonstrates consistent improvements across different VideoQA models and sizes. For standard-context models (64 frames), GenS enhances LLaVA-Video-7B and Qwen2-VL-7B by 4.4 and 2.7 points, respectively, on the full validation set. Notably, even for long-context models like Aria-25B (256 frames), GenS still brings a significant 3.4-point improvement, highlighting the importance of efficient frame sampling beyond model context length scaling. When equipped with GenS, LLaVA-Video-72B achieves 66.8% accuracy on LongVideoBench, establishing a new state-of-the-art. These gains become more pronounced on the *Vision-Centric subset*, where GenS improves LLaVA-Video-72B by 7.3 points and Aria-25B by 4.9 points, demonstrating its particular effectiveness on questions that demand stronger visual understanding capabilities. Figure 1 (b) indicates that GenS also significantly improves the performance of GPT-4o, achieving a 13.6% accuracy gain with 40 input frames.

MLVU. On the MLVU benchmark, GenS consistently enhances the performance of various VideoQA models. LLaVA-Video-7B's accuracy improves by 3.0 points (from 70.4% to 73.4%), Qwen2-VL-7B shows a 2.2-point increase (from 64.7% to 66.9%), and Aria demonstrates a 3.1-point gain (from 69.5% to 72.6%). Most notably, LLaVA-Video-72B integrated with GenS achieves state-of-the-art performance with 77.0% accuracy.

HourVideo. We further evaluate GenS on HourVideo (Table 3), a particularly challenging benchmark featuring videos with an average duration of 45.7 minutes, including 113 videos that exceed one hour in length. Prior to our work, only Gemini-1.5-Pro could process such extensive videos end-to-end, achieving 37.3% accuracy. With the integration of GenS, both Aria-25B and Gemini-1.5-Pro surpass previous results, reaching 39.2% and 40.7% accuracy respectively, thereby establishing new state-of-the-art performance. These improvements demonstrate GenS's capability to effectively process extremely long videos through its dynamic frame identification.

3.3 Comparison with Sampling Baselines

We evaluate GenS against various frame sampling approaches in Table 4. Specifically, we compare against uniform sampling baseline to assess the effectiveness of these methods.

Image-language matching methods like CLIP-L-224px (Radford et al., 2021) and SigLIP-384px (Zhai et al., 2023) demonstrate significant improvements over uniform sampling only with sparse frame inputs (e.g., 16 frames), as shown in Figure 1 (b). However, when processing 256 frames, they achieve only slight gains (0.7/1.1

	LongVideoBench$_{val}$ *(V-Centric Subset)*
Aria-25B as VideoQA Model (<=256 frames)	
+Uniform Sampling	54.4
Image-Language Matching	
+CLIP-L-224px Sampler	55.5
+SigLIP-384px Sampler	55.1
+InternVL-14B-224px Sampler	55.9
Proprietary LMMs	
+GPT-4o Sampler	55.7
Open-source VideoLLMs	
+GenS (ours)	**60.5**
	MLVU$_{M\text{-}avg}$
VILA-v1.5-40B as VideoQA Model (<=14 frames)	
+Uniform Sampling	57.8
Specialized VideoLLMs for Event Localization	
+TimeChat Sampler$_{[CVPR\ 2024]}$	59.4
Specialized VideoLLMs for Frame Sampling	
+FRAME-VOYAGER Sampler$_{[ICLR\ 2025]}$	61.1
+GenS (ours)	**63.5**

Table 4: Comparison with different frame sampling methods.

points) over uniform sampling, due to their limitations in complex language reasoning and temporal relationship modeling. As Table 4 shows, InternVL-14B-224px (Wang et al., 2024b) outperform CLIP-based approaches, benefiting from their enhanced language understanding capabilities.

Advanced Proprietary LMMs like GPT-4o unexpectedly yield only a 1.3-point improvement over 256-frame uniform sampling. Our empirical observations reveal that GPT-4o struggles with precise frame selection when processing large candidate frame sets, particularly in identifying correct frame indices. While multi-round refinement or step-by-step reasoning could potentially address these limitations, it would be prohibitively expensive.

VideoLLM-based Methods like FRAME-VOYAGER (Yu et al., 2024) specialize in sampling sparse frames from a candidate pool (e.g., 8 from 128 frames). Following their experimental settings, GenS also shows effectiveness for enhancing short-context VideoQA models such as VILA-1.5-40B (14 frames) (Lin et al., 2023), surpassing FRAME-VOYAGER by 2.4 points and uniform sampling by 5.7 points. Event localization methods such as TimeChat (Ren et al., 2024) can identify event timestamps based on textual descriptions, but show only marginal improvements over uniform sampling when used for frame sampling, likely due to the coarse and sparse nature of event localization annotations.

3.4 Results on Temporal Grounding Tasks

Table 5 demonstrates that GenS achieves competitive video grounding performance, surpassing GPT-4o and approaching specialized VideoLLMs like

Grounding Model	Charades-STA			
	R1@0.3	R1@0.5	R1@0.7	mIoU
Temporal Grounding VideoLLMs (7B size)				
VTimeLLM	51.0	27.5	11.4	31.2
HawkEye	50.6	31.4	14.5	33.7
TimeChat$_{[CVPR\ 2024]}$	-	32.2	13.4	30.6
TimeSuite$_{[ICLR\ 2025]}$	69.9	48.7	24.0	-
General VideoLLMs				
GPT-4o	55.0	32.0	11.5	35.4
VideoChat2-7B	9.6	3.4	1.4	-
Qwen2-VL-7B	8.7	5.4	2.4	7.9
LongVA-7B-DPO	22.6	10.1	2.2	14.6
LLaVA-OneVison-7B	31.2	13.5	5.2	-
Aria	39.0	18.6	6.6	26.7
GenS	62.9	38.7	15.2	38.0
GenS $_{w/o\ E.T.Instruct\text{-}41K_{agg.}}$	51.1	28.2	10.4	33.2

Table 5: Results on the Charades-STA (Gao et al., 2017) temporal grounding benchmark.

TimeSuite (Zeng et al., 2024). This highlights its excellence in both long-form video understanding and fine-grained temporal localization. Notably, our model achieves these results without training on any Charades-STA data. Even when excluding *E.T.Instruct-41K$_{agg.}$* that contains temporal grounding data from DiDeMo (8.4K samples), Queryd (661 samples), and TACoS (61 samples) (Anne Hendricks et al., 2017; Oncescu et al., 2021; Regneri et al., 2013), our model's performance remains comparable to GPT-4o.

3.5 Analysis

Effectiveness of GenS-Video-150K Dataset. Table 7 presents that adding our GenS-Video-150K brings remarkable improvements over the uniform sampling baseline across two VideoQA models. For GPT-4o with 32 frames input, adding VC-RAG-150K improves accuracy by 10.4 points (from 53.4 to 63.8). For Aria with 256 frames input, the improvement is 3.3 points (from 54.4 to 57.7). We further evaluate the combination of our GenS-Video-150K with temporal grounding data DiDeMo (Anne Hendricks et al., 2017) and time-sensitive video dataset E.T.Instruct (Liu et al.). Experiments reveal that directly combining these datasets degrades GPT-4o's performance due to cross-task inconsistencies. However, applying query aggregation post-processing (described in Section 3.1) leads to improved overall performance. We also compare two prompting strategies: using a single unified prompt as frame-sampling tasks versus using task-specific prompts. Our results show that task-specific prompts perform better, since they allow the model to learn specialized behaviors for each task type while still benefiting from the combined training data.

	Holistic		Multi Detail		Single Detail			M-Avg
	Topic Reason	Anomaly Recognition	Action Order	Action Count	Needle QA	Ego Reason	PlotQA	
Uniform	87.1	69.5	63.7	44.2	77.8	66.2	71.8	69.9
GenS	84.7 (-2.4)	72.0 (+2.5)	73.3 (+9.6)	41.3 (-2.9)	85.4 (+7.6)	66.8 (+0.6)	76.4 (+4.6)	73.2 (+3.3)

Table 6: Breakdown performance analysis on MLVU (val) using Aria as the VideoQA model with 256 input frames.

VideoQA Models	GPT-4o (<=32 frm.)	Aria (<=256 frm.)
Uniform Sampling	53.4	54.4
Frame Sampler		
+ GenS-Video-150K	63.8	57.7
Directly Mixing Dataset		
+ GenS-Video-150K + DiDeMo-40K	61.5	57.7
+ GenS-Video-150K + E.T.Instruct-75K	62.7	59.5
With Query Aggregation		
+ GenS-Video-150K + E.T.Instruct-41K$_{agg.}$ *(unified task prompts)*	63.8	**60.9**
+ GenS-Video-150K + E.T.Instruct-41K$_{agg.}$ *(distinct task prompts)*	**64.2**	60.5

Table 7: Ablation study on different training datasets and combination strategies. Results are accuracy (%) on LongVideoBench$_{val}$ *(V-Centric Subset)*.

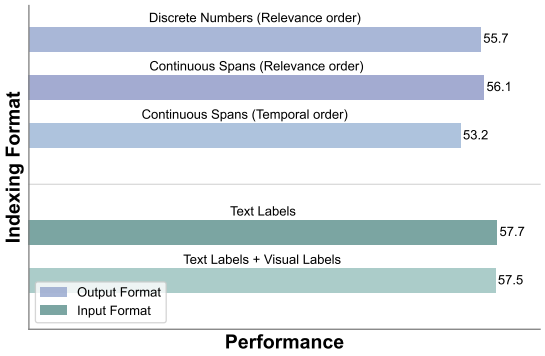

Figure 2: Ablation study on different input and output frame indexing formats.

Input and Output Indexing Format. For output formats, we evaluate two key aspects: (1) use discrete index numbers versus integrate successive frames into continuous spans, and (2) order frames chronologically or by relevance. Figure 2 depicts that *continuous spans* with confidence scores *ordered by relevance* achieve the best performance (56.1). For input formats, we compare two strategies: (1) *textual indexing alone* that prepending each frame with a textual number `[N]` and (2) *combining textual and visual indexing* which additionally overlays visual numerical indices directly onto each frame at the pixel level (Wu et al., 2024c). Our results show that *textual indexing alone* performs marginally better than *combining textual and visual indexing*, indicating that GenS can effectively process interleaved visual-textual sequences.

Breakdown results on different question types. Table 6 shows that GenS achieves significant improvements on question types that require precise temporal understanding and localization. Specifically, GenS brings notable gains on Needle QA (+7.6) and Action Order (+9.6) tasks, where identifying specific moments or temporal relationships between actions is crucial. However, for Topic Reason tasks that require holistic video understanding, uniform sampling provides better coverage of the overall video content.

4 Extension Applications

Coarse-to-Fine Hybrid Sampling. We propose a coarse-to-fine hybrid approach that combines a lightweight CLIP (Radford et al., 2021) sampler with our GenS to improve sampling efficiency for extremely long videos. Specifically, we first adopt CLIP to densely sample frames from the 1 fps candidate pool and return top 256 most relevant frames, then apply GenS to re-sample the most informative frames within a single 256-frame temporal window.

Table 8 demonstrates that this hybrid approach consistently outperforms both uniform sampling and standalone CLIP sampling across various frame count constraints. We also provide the GenS sampled at 1fps as an upper bound for the frame sampling efficiency-performance tradeoff. This demonstrates that a simple hybrid approach can effectively improve the efficiency of the GenS sampler while maintaining competitive performance, which provides a more practical solution for real-world applications.

GenS Implementation on Qwen2.5VL-3B. Our design of *generative frame sampling* is not limited to a specific VideoLLM (e.g., Aria) as the base model. To verify the generalizability of our approach, we implemented GenS on Qwen2.5VL-3B using low-resolution inputs (112×112 pixels) for frame sampling.

Results in Table 9 demonstrate that GenS based on Qwen2.5VL-3B achieves remarkable performance compared to both uniform sampling and CLIP-based samplers. The model shows consistent improvements across all frame count configurations, with gains of up to 7.51 points when using

#Sampled Frames	≤10 frm	≤20 frm	≤30 frm	≤40 frm	≤50 frm
Uniform	47.04	50.00	53.75	51.98	53.95
CLIP-L-224px	46.44	50.00	53.16	54.55	54.74
CLIP-L-224px + GenS	**53.16** (+6.12)	**56.13** (+6.13)	**60.47** (+6.72)	**61.26** (+6.71)	**60.86** (+6.12)
GenS (Upper Bound)	56.32	58.89	65.61	65.61	64.82

Table 8: Performance of a hybrid sampling approach on LongVideoBench$_{val}$ *(V-Centric Subset)* using GPT-4o as VideoQA model. The hybrid approach first retrieves the top 256 relevant frames from a 1 FPS candidate pool using CLIP-L-224px, then applies GenS to select the most informative frames within a single 256-frame temporal window.

#Sampled Frames	≤10 frm	≤20 frm	≤30 frm	≤40 frm	≤50 frm
Uniform	47.04	50.00	53.75	51.98	53.95
InternVL-14B-224px	47.23	49.01	50.99	54.35	54.55
CLIP-L-224px	46.44	50.00	53.16	54.55	54.74
GenS$_{Qwen2.5VL-3B}$	**54.74** (+7.51)	**55.93** (+5.93)	**57.11** (+3.36)	**59.68** (+5.13)	**58.69** (+3.95)

Table 9: Performance of GenS with Qwen2.5VL-3B as the base VideoLLM using low-resolution input (112×112 pixels) on LongVideoBench$_{val}$ *(V-Centric Subset)*, with GPT-4o as the VideoQA model.

just 10 frames. The successful adaptation of GenS to the distinctly different Qwen2.5VL-3B architecture validates the broad generalizability of our approach. Our method can be integrated with various advanced VideoLLMs without requiring architectural modifications, enabling it to leverage ongoing advancements in the field.

5 Related Work

5.1 Long-form Video Understanding

Current video assistants (Li et al., 2023b; Zhang et al., 2024b; Lin et al., 2023; Wang et al., 2024a; Chen et al., 2024b; Li et al., 2024a) have demonstrated impressive capabilities in video-language understanding. However, processing hours-long videos (e.g., 3600 frames per hour at 1 fps) for comprehensive visual coverage remains computationally prohibitive. Recent approaches address this limitation through either: 1) extending model context length to accommodate more frames (e.g., 256-512) (Zhang et al., 2024a; Liu et al., 2024b; Fei et al., 2024; Wang et al., 2024d, 2025; Xue et al., 2024), or 2) performing visual token compression within the model (Li et al., 2024d; Yao et al., 2024; Zhang et al., 2025; Li et al., 2024c). In contrast, we propose a more efficient paradigm - incorporating a frame sampler prior to model input, thus eliminating redundant visual processing inside the large-scale video assistants.

5.2 VideoLLMs with Retrieval-Augmented Generation

To enhance video-language interaction, recent works have equipped VideoLLM assistants with Retrieval-Augmented Generation (RAG). Unlike text-based retrieval methods, i.e., Video-RAG (Luo et al., 2024), Q-ViD (Romero and Solorio, 2024), and R2A (Pan et al., 2023), we propose a visual-centric approach that directly retrieves relevant frames. Compared to CLIP-based retrieval (Wang et al., 2024b; Arefeen et al., 2024; Wang et al., 2024c; Xu et al., 2024), our method built on a VideoLLM excels at capturing long-range temporal perception and complex language understanding. While similar frame samplers (Yu et al., 2024; Sun et al., 2025) are limited to sparse sampling (e.g., 8 from 128 frames), our approach can efficiently retrieve thousands of frames with adaptive sampling rates, substantially enhancing long-context video assistants on hours-long video perception.

6 Conclusion

This paper presents GenS, a novel generative frame sampling method and a high-quality video instruction dataset GenS-Video-150K. Our extensive experiments show that GenS brings consistent improvements across different VideoQA models' architectures and sizes and achieve new state-of-the-art results on LongVideoBench (66.9), MLVU (77.0), and HourVideo (40.7). It suggests that efficient frame sampling is a promising direction for advancing long-form video understanding.

Limitations

Leveraging GenS for key frame retrieval in long-form videos incurs additional computational overhead compared to naive uniform sampling. Specifically, while uniform sampling processes N frames (where N is the context length of the video question-answering model), our approach needs to analyze M frames (M=256) within each retrieval window. However, this computational cost can be mitigated through parallel processing of multiple segment windows, making the overall inference time practically manageable. Meanwhile, for large-scale advanced VideoQA Assistants like LLaVA-Video-72B, sampling few relevant frames via GenS (3.9B activated parameters) is more efficient than substantially extending the model context length of a 72B VideoQA Assistant. The performance of GenS could be further enhanced through multi-round retrieval iterations and integration with Video Agent systems for refined frame selection.

Acknowledgements

This research was partially supported by the National Natural Science Foundation of China under Grant No. 92470205 and No. 62176002. Xu Sun and Junnan Li are the corresponding authors.

References

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812.

Md Adnan Arefeen, Biplob Debnath, Md Yusuf Sarwar Uddin, and Srimat Chakradhar. 2024. Vita: An efficient video-to-text algorithm using vlm for rag-based video analysis system. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2266–2274.

Leonard Bärmann and Alex Waibel. 2022. Where did i leave my keys?-episodic-memory-based question answering on egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1560–1568.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. 2024. Hourvideo: 1-hour video-language understanding. *arXiv preprint arXiv:2411.04998*.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. 2024a. Sharegpt4video: Improving video understanding and generation with better captions. *ArXiv preprint*, abs/2406.04325.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Preprint*, arXiv:2404.16821.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *ArXiv preprint*, abs/2312.14238.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar,

Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *ArXiv preprint*, abs/2407.21783.

Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. 2024. Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos. *arXiv preprint arXiv:2408.14023*.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *ArXiv preprint*, abs/2405.21075.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275.

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv preprint*, abs/2403.05530.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Jakub Krajewski, Jan Ludziejewski, Kamil Adamczewski, Maciej Pióro, Michał Krutul, Szymon Antoniak, Kamil Ciebiera, Krystian Król, Tomasz Odrzygóźdź, Piotr Sankowski, et al. 2024. Scaling laws for fine-grained mixture of experts. *arXiv preprint arXiv:2402.07871*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *ArXiv preprint*, abs/2408.03326.

Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Fan Zhou, Chengen Huang, Yanpeng Li, Chongyan Zhu, Xiaoyi Ren, Chao Li, Yifan Ye, Peng Liu, Lihuan Zhang, Hanshu Yan, Guoyin Wang, Bei Chen, and Junnan Li. 2025. Aria: An Open Multimodal Native Mixture-of-Experts Model.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023b. Videochat: Chat-centric video understanding. *ArXiv preprint*, abs/2305.06355.

Lei Li, Yuanxin Liu, Linli Yao, Peiyuan Zhang, Chenxin An, Lean Wang, Xu Sun, Lingpeng Kong, and Qi Liu. 2024b. Temporal reasoning transfer from text to video. *arXiv preprint arXiv:2410.06166*.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. 2024c. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2024d. Llama-vid: An image is worth 2 tokens in large language models.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. 2023. Vila: On pretraining for visual language models. *Preprint*, arXiv:2312.07533.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. Et bench: Towards open-ended event-level video-language understanding. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. 2024b. Nvila: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*.

Yongdong Luo, Xiawu Zheng, Xiao Yang, Guilin Li, Haojia Lin, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. 2024. Video-rag: Visually-aligned retrieval-augmented long video comprehension. *arXiv preprint arXiv:2411.13093*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*.

Andreea-Maria Oncescu, Joao F Henriques, Yang Liu, Andrew Zisserman, and Samuel Albanie. 2021. Queryd: A video dataset with high-quality text and audio narrations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2265–2269. IEEE.

OpenAI. 2024. Gpt-4o system card.

Junting Pan, Ziyi Lin, Yuying Ge, Xiatian Zhu, Renrui Zhang, Yi Wang, Yu Qiao, and Hongsheng Li. 2023. Retrieving-to-answer: Zero-shot video question answering with frozen large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 272–283.

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

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2024. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323.

David Romero and Thamar Solorio. 2024. Question-instructed visual descriptions for zero-shot video question answering. *arXiv preprint arXiv:2402.10698*.

Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. 2025. Mdp3: A training-free approach for list-wise frame selection in video-llms. *arXiv preprint arXiv:2501.02885*.

Shengbang Tong, Ellis L Brown II, Penghao Wu, Sanghyun Woo, ADITHYA JAIRAM IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, et al. 2024b. Needle in a multimodal haystack. *arXiv preprint arXiv:2406.07230*.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. 2024c. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer.

Xidong Wang, Dingjie Song, Shunian Chen, Chen Zhang, and Benyou Wang. 2024d. Longllava: Scaling multi-modal llms to 1000 images efficiently via a hybrid architecture. *arXiv preprint arXiv:2409.02889*.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, et al. 2025. Internvideo2. 5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024a. Longvideobench: A benchmark for long-context interleaved video-language understanding. *ArXiv preprint*, abs/2407.15754.

Yongliang Wu, Xinting Hu, Yuyang Sun, Yizhou Zhou, Wenbo Zhu, Fengyun Rao, Bernt Schiele, and Xu Yang. 2024b. Number it: Temporal grounding videos like flipping manga. *arXiv preprint arXiv:2411.10332*.

Yongliang Wu, Xinting Hu, Yuyang Sun, Yizhou Zhou, Wenbo Zhu, Fengyun Rao, Bernt Schiele, and Xu Yang. 2024c. Number it: Temporal grounding videos like flipping manga. *arXiv preprint arXiv:2411.10332*.

Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. 2024. Retrieval-augmented egocentric video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13525–13536.

Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. 2024. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*.

Linli Yao, Lei Li, Shuhuai Ren, Lean Wang, Yuanxin Liu, Xu Sun, and Lu Hou. 2024. Deco: Decoupling token compression from semantic abstraction in multimodal large language models. *arXiv preprint arXiv:2405.20985*.

Sicheng Yu, Chengkai Jin, Huanyu Wang, Zhenghao Chen, Sheng Jin, Zhongrong Zuo, Xiaolei Xu, Zhenbang Sun, Bingni Zhang, Jiawei Wu, et al. 2024. Frame-voyager: Learning to query frames for video large language models. *arXiv preprint arXiv:2410.03226*.

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. 2022. 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.

Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong

Shi, Zhengrong Yue, Yi Wang, et al. 2024. Timesuite: Improving mllms for long video understanding via grounded tuning. *arXiv preprint arXiv:2410.19702*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. 2025. Video-ollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 543–553.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2024a. Long context transfer from language to vision. *ArXiv preprint*, abs/2406.16852.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024b. Video instruction tuning with synthetic data. *Preprint*, arXiv:2410.02713.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2024. Mlvu: A comprehensive benchmark for multi-task long video understanding. *ArXiv preprint*, abs/2406.04264.

Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, et al. 2024. Apollo: An exploration of video understanding in large multimodal models. *arXiv preprint arXiv:2412.10360*.

A Appendix

A.1 Additional Results and Analysis

Leveraging Video Subtitles for Frame Sampling. As shown in Table 10, incorporating video subtitle information during frame sampling improves the quality of selected frames and enhances downstream VideoQA model performance. Specifically, across different input scales ranging from 16 to 256 frames, adding subtitle information yields consistent improvements of 0.6-2.77 points. This demonstrates GenS's ability to effectively integrate visual and textual cues while maintaining efficient frame sampling. Notably, GenS achieves this performance without explicit training on frame-subtitle sequences. This zero-shot capability stems from GenS's base VideoLLM architecture, which inherently supports processing interleaved vision-text inputs.

Sampling	Max Input Frames				
	≤ 16	≤ 32	≤ 64	≤ 128	≤ 256
w/o subtitle	54.74	54.74	55.34	55.53	54.74
w/ subtitle	55.34	54.74	55.34	56.72	57.51
	+0.60	+0.00	+0.00	+1.19	+2.77

Table 10: Impact of incorporating video subtitles during frame sampling, evaluated on LongVideoBench (V-Centric) using Aria as the VideoQA model. The subtitles are only used by the frame sampler GenS, not by the VideoQA model. Results show accuracy (%) for different maximum frame budgets.

Frame Sampling Parameters during Inference. We investigate two key parameters in our frame sampling approach: sampling ratio and temporal window size. The sampling ratio determines how densely we sample candidate frames from the original video (measured in frames per second), while the temporal window size (temp_win_size) controls how many consecutive frames are considered simultaneously during sampling. As shown in Table 11, increasing the sampling ratio from 0.2 to 1.0 fps with temp_win_size=128 significantly improves performance from 52.37 to 55.13, as denser sampling provides more comprehensive video coverage. With the sampling ratio fixed at 1.0 fps, expanding the temporal window from 128 to 256 frames yields a further improvement to 56.13, demonstrating the benefit of longer-range temporal perception. Notably, during training, we use an average sampling ratio of 0.2 fps and temp_win_size of 129 frames. The superior performance achieved with different parameters during inference suggests that GenS generalizes well beyond its training configuration rather than overfitting to the training settings.

Sampling Configuration	Accuracy
sample_ratio=0.2, temp_win_size=128	52.37
sample_ratio=1.0, temp_win_size=128	55.13
sample_ratio=1.0, temp_win_size=256	**56.13**

Table 11: Impact of sampling ratio and temporal window size on LongVideoBench (V-Centric). Higher sampling ratio enables denser frame coverage, while larger temporal window allows longer-range temporal perception.

A.2 GenS-Video-150K Dataset Details

Prompts for GenS-Video-150K Annotation. We provide detailed prompts for each stage of the GenS-Video-150K annotation process via GPT-4o. Table 14 depicts the prompts for **Stage 2 Construct Grounded Video QAs**, while Table 15 shows the prompts for **Stage 4 Score Frame Relevance**.

When constructing Video QAs with grounded frames (Stage 2), we define 12 specific question types to comprehensively cover different aspects of video understanding capabilities:

- **Reasoning Tasks:** Object, Action, Spatial, and Temporal Reasoning questions test the model's ability to make logical inferences about relationships and changes in the video.
- **Perception Tasks:** Object, Action, Attribute, and Spatial Perception questions focus on basic visual understanding of scenes, actions, and object properties.
- **Specialized Tasks:** Video Detail Referring requires fine-grained visual attention, Counting tests quantitative understanding, OCR evaluates text recognition, and Temporal Perception assesses understanding of event sequences.

A.3 Training Hyper-parameters and Evaluation Details

We provide training hyper-parameters in Table 12.

Video-Centric Subset. We use GPT-4o to filter questions that can be answered by purely textual reasoning on LongVideoBench (LVB) and MLVU. The filtered dataset contains 506 samples for LVB (excluding 159 non-vision-centric questions and 672 videos shorter than 10 minutes) and 879 samples for MLVU (excluding 200 non-vision-centric

Hyper-parameter	Value
Visual Encoder	
Frame Sampling Rate	Varied FPS (0.2-1.0)
Input Resolution	490
Visual Tokens per Image	128
Max Image per Sequence	256
Patch Size	14x14
Large Language Model (MOE)	
Number of Layers	28
Hidden Size	2560
FFN Hidden Size	13568
MOE FFN Dimension	1664
Number of Attention Heads	20
Number of KV Heads	20
Number of Experts	64
Top-k Experts	6
Number of Shared Experts	2
Model Training	
Max Context Length	32768
Batch Size	256
Learning Rate	1e-5
Min Learning Rate	1e-8
Warmup Ratio	0.0
Training Iterations	300
Z Loss	1e-5
EP ST Load Balancing Loss	1e-3
LR Scheduler Type	Cosine

Table 12: Training hyper-parameters for GenS.

questions and 1,095 videos shorter than 8 minutes). All remaining questions in this subset explicitly require long-range visual understanding capabilities.

A.4 Frame Sampling Baseline Implementation

CLIP / SigLIP / InternVL. For image-language models like CLIP, SigLIP and InternVL, we implement frame sampling through similarity-based retrieval. We first densely sample frames from the original video at 1 FPS and extract visual features for each frame. We then encode the input question into text features and compute cosine similarity scores between each frame and the question embedding. Finally, we select the top-K frames with highest similarity scores as key frames, where K is determined by the maximum input frame capacity of the VideoQA model.

TimeChat. For event localization VideoLLMs like TimeChat, we use the question as a textual query to identify relevant event timestamps in the video. We then uniformly sample K frames from these identified temporal segments as key frames for downstream processing.

Prompt as a Frame Retrieval Assistant:
You are an advanced AI visual assistant tasked with assessing frame relevance for question answering. Please retrieve the video frames relevant to the question (maybe with options) to answer it correctly, output the frame timestamp, exactly in format [XX:XX], [XX:XX, XX:XX], or [XX:XX, XX:XX, XX:XX], etc. If no matching frames are found, output [None].
Video Frames:

```
[00:00] <image_placeholder>
[00:05] <image_placeholder>
[00:10] <image_placeholder>
...
```

Question:

```
<question_placeholder>
```

Output Relevance Frames:

```
[17:07, 17:26, 18:24]
```

Table 13: The prompt used by GPT-4o to retrieve relevant video frames for question answering.

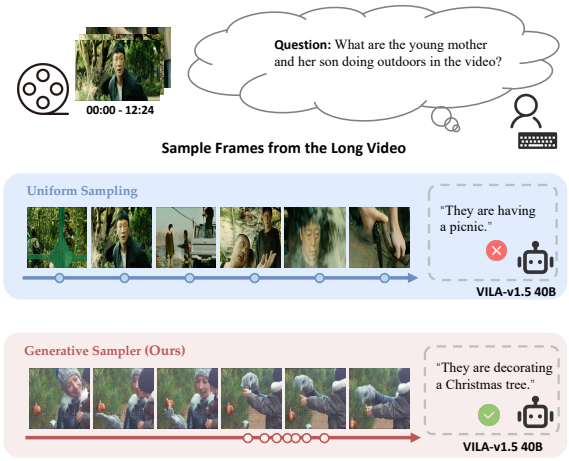

Figure 3: Visualization of GenS integrated with VILA-v1.5-40B (<=14frames) on MLVU dataset.

GPT-4o. We leverage GPT-4o's vision capabilities to score frame relevance based on the prompt template in Table 13. Since GPT-4o has limitations in processing massive frames at once, we first sample frames at 1 FPS from the original video and divide them into windows of 50 frames each. GPT-4o then processes each window independently to identify relevant frames. The relevant frames from all windows are aggregated to obtain K candidate frames. If K exceeds the VideoQA model's maximum input capacity N, we randomly sample N frames from the candidate set to form the final input.

A.5 Visualization Cases

We visualize a case in Figure 3.

Videos: lVNV1qXnGb0.mp4

Questions: What event occurs immediately before the scene transitions to the outdoor environment with the cave entrance?

Answer: A detailed view of an individual examining a substance through a microscope in the laboratory.

Relevant Frames with Fine-grained Scores: {"30s-35s": score 5, "25s": score 4, "40s": score 4, "265s-270s": score 3, "305s-310s": score 3, "285s": score 2, "330s": score 2, ..., "720s": score 1}

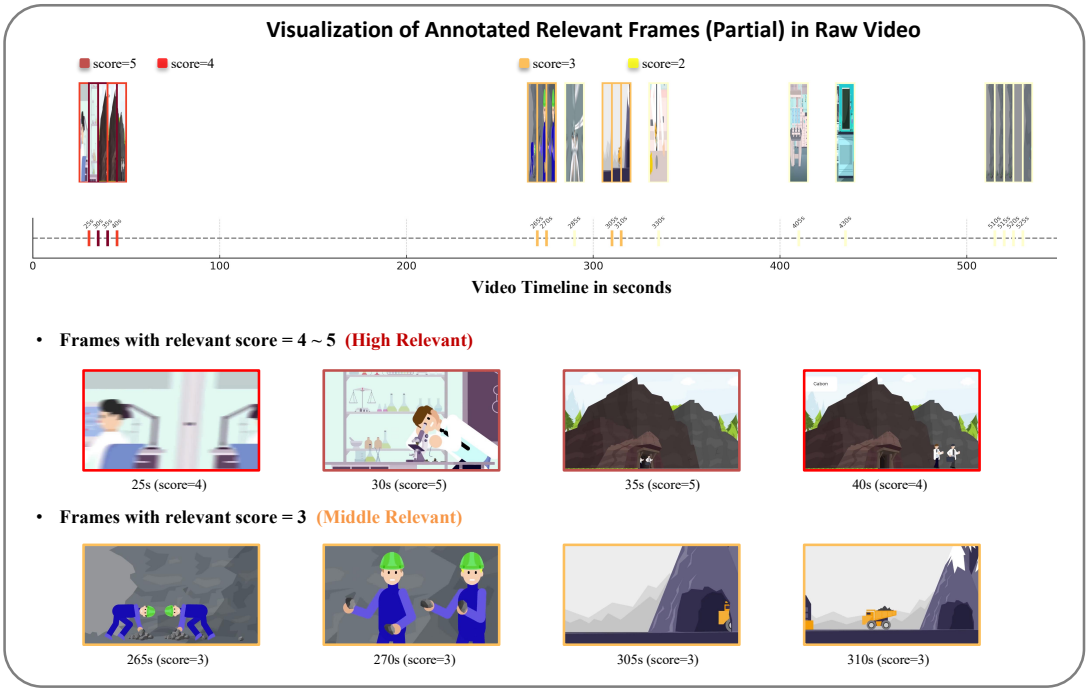

Figure 4: Visualization of annotated data sample from GenS-Video-150K.

GPT-4 Prompt for Grounded Video Question Generation (Stage 2):
You are a teacher designing challenging questions for a "Long-term Video Understanding" class. Your task is to create questions that test students' ability to comprehend and analyze long video content. I will provide video frame narrations with timestamps, and you should generate questions following the criteria below:

Question Types:

- **Specific requirements for a question type:**

  ```
  <Question Type Placeholder>
  ```

- **Multiple Choice:** Create questions with five options (A-E), where only one is correct. All options should be closely related to the video content but clearly distinguishable.
- **Open-ended:** Create questions requiring concise, specific answers that can be directly supported by the video content.

Key Requirements:

1. **Video-Centric:** Questions must be answerable solely through careful analysis of the video content. Avoid requiring external knowledge.
2. **Temporal Reasoning:** Questions should require understanding relationships between events across different timestamps.
3. **Clear Answers:** Ensure answers are concise, accurate, and directly supported by video evidence.
4. **Difficulty:** Make questions challenging by requiring careful analysis of multiple video segments.
5. **Format:** Do not reference scene indices or specific timestamps in questions.

Input Format:

```
[Timestamp] Description of video frame
[04:30] Person walks into room
[04:35] Person picks up book
...
```

Output Format:
For Multiple Choice Questions:

```
{
    "question": "...",
    "options": {
        "A": "...", "B": "...", "C": "...", "D": "...", "E": "..."
    },
    "correct_option": "A/B/C/D/E",
    "rationale_timestamps": ["04:30", "04:35", ...]
}
```

For Open-ended Questions:

```
{
    "question": "...",
    "answer": "...",
    "rationale_timestamps": ["04:30", "04:35", ...]
}
```

Additional Guidelines:

- Make wrong options slightly longer than correct ones
- Distribute correct answers evenly across options A-E
- Include only timestamps directly relevant to the question
- Ensure answers compress information from multiple timestamps

Table 14: The prompt template used for **Stage 2 - Construct Grounded Video QAs** of GenS-Video-150K.

GPT-4o Prompt for Scoring Relevance (Stage 4):
You are a helpful and precise assistant designed to evaluate the relevance of each video frame to a given textual question. I will provide video frames (or their narrations) along with their timestamps as input. Your task is to assign an overall relevance score on a scale from 1 to 5 for each video frame, where higher scores indicate better alignment with the question. Use the criteria below to guide your scoring:

Scoring Criteria:

- **5 (Highly Relevant):** The video frame contains unique visual cues critical to accurately answering the question. Without this frame, it would be challenging to reason or provide an accurate answer.
- **4 (Directly Relevant):** The video frame is directly related to the question, but its visual cues can be partially replaced or supplemented by other important frames.
- **3 (Moderately Relevant):** The video frame is important for addressing the question, helping identify related scenes, activities, actions, individuals, or other elements, or aiding in ruling out incorrect options.
- **2 (Somewhat Relevant):** The video frame indirectly relates to the question, providing supporting context that aids in reasoning or finding the correct answer.
- **1 (Minimally Relevant):** The video frame has minimal relevance to the question. While it may involve the same person, activity, or action as the question, it does not contribute meaningfully to answering it.
- **0 (Irrelevant):** The video frame has no relevance to the question.

Additional Notes:

1. Some textual questions may require multi-hop reasoning, necessitating the combination of visual cues from multiple frames to arrive at the correct answer.
2. Some questions may ask about the global information of the video, such as identifying its main focus or summarizing the content. In these cases, assign higher scores to frames with unique and non-redundant visual information to ensure the selected frames collectively provide a comprehensive summary of the video while minimizing redundancy.
3. Most input video frames will have some relevance to the question, so prioritize scoring between 1 and 5. Use a score of 0 only for entirely irrelevant frames.
4. If the question is: (1) Ambiguous, such that none of the input frames can provide an answer, or (2) Contains logical issues (e.g., contradictions or nonsensical reasoning), then the question should be flagged as low quality, and the output should be "the question has low quality".

Input Frames:

```
[04:40] <image_placeholder>
[04:45] <image_placeholder>
[04:50] <image_placeholder>
...
```

Question:

```
<question_placeholder>
```

Hint:
Frames at timestamps

```
<timestamp_placeholder>
```

Output Format:
Provide an explanation for the assigned scores to justify your reasoning. Return the results in the following JSON format:

```
{
    "[04:40]": score 0-5,
    "[04:45]": score 0-5,
    ...
}
``Explain why each score was assigned, detailing the relevance of the frames to the question...''
```

Table 15: The prompt template used for **Stage 4 - Score Frame Relevance** of GenS-Video-150K.