

VSI: Visual–Subtitle Integration for Keyframe Selection to Enhance Long Video Understanding

Supplementary Material

Abstract

001 *This document provides supplementary material for the main*
002 *paper, including additional experiments, derivations, data,*
003 *figures, algorithms, and other relevant content. Please add*
004 *detailed information as needed. This supplementary material*
005 *is submitted together with the main paper to further support*
006 *and complement the main findings.*

007 A. Parametric Analysis

008 The quantitative evaluation of hyperparameter influence con-
009 stitutes a critical component in understanding model behav-
010 ior and optimizing complex AI systems. Delving into the
011 depths of AI model functionality, it becomes evident that hy-
012 perparameters serve as pivotal tuning knobs, shaping the per-
013 formance and adaptability of these intricate systems. Within
014 our proposed framework, the Text Weight coefficient ex-
015 plicitly governs the contribution of textual semantics during
016 multimodal fusion, acting as a decisive factor in harmoniz-
017 ing diverse data streams. By adjusting this parameter, we
018 can fine-tune the balance between text and other modalities,
019 thereby steering the system towards optimal performance.
020 To further explore the impact of this important parameter
021 on our approach, we conducted comparative experiments
022 at LONGVIDEOBENCH-Text-Related-Perception Subsets,
023 meticulously analyzing the interplay between text semantics
024 and overall system efficacy.

025 As demonstrated in Table 2, increasing the Text Weight
026 leads to higher frame searching accuracy on text-related
027 datasets, revealing a profound correlation between text em-
028 phasis and search precision. Notably, when the parameter is
029 set to 1, meaning that only text-related information is utilized
030 for frame searching, our method achieves its highest frame
031 searching accuracy, surpassing previous results by more than
032 double. This remarkable improvement underscores the sig-
033 nificance of text-centric processing within our framework.
034 This experimental pattern indicates that our Subtitle Match
035 Stream effectively captures key information related to the
036 text, showcasing its robust capability to extract and integrate
037 pivotal textual details. Through these insights, we gain a
038 deeper understanding of the dynamic role that text plays in
039 enhancing model performance, paving the way for future
040 advancements in multimodal AI systems.

B. Details of Datasets

041

B.1. Details of VIDEO-MME

042

043 The VIDEO-MME (Video Multi-Modal Evaluation) [?]
044 dataset stands as a pioneering benchmark specifically de-
045 signed to evaluate the capabilities of Vision-Language Mod-
046 els (VLMs) in the realm of video understanding. By address-
047 ing the limitations of existing benchmarks, this dataset sets a
048 new standard in the assessment of VLMs, emphasizing diver-
049 sity, temporal complexity, and multi-modal integration—all
050 while ensuring high-quality human annotations that lend
051 credibility and depth to the evaluation process. With a col-
052 lection of 900 meticulously selected videos, VIDEO-MME
053 covers six major domains—Knowledge, Film and Television,
054 Sports Competition, Artistic Performance, Life Record, and
055 Multilingual—each offering a unique lens through which
056 video content can be analyzed. Within these domains, the
057 dataset further delves into 30 fine-grained subcategories,
058 encompassing areas such as astronomy, esports, and docu-
059 mentaries, to provide a comprehensive scope of video con-
060 tent. The videos in the dataset vary significantly in duration,
061 ranging from brief clips lasting just 11 seconds to extensive
062 long-form content that can stretch up to 1 hour. This wide
063 range of temporal scales allows for a robust evaluation of
064 VLMs, testing their ability to process and understand video
065 content across different lengths and complexities. Through
066 its thoughtful design, VIDEO-MME offers an unparalleled
067 resource for advancing the study of video understanding in
068 AI, fostering innovation and progress in this dynamic field.

069 Each video within the VIDEO-MME dataset is accompa-
070 nied by expertly annotated multiple-choice questions, total-
071 ing 2,700 QA pairs, which have been meticulously validated
072 to ensure they are both clear and dependent on visual or
073 multi-modal context. These questions are crafted to cover
074 a diverse array of 12 task types, such as action recogni-
075 tion, temporal reasoning, and domain-specific knowledge,
076 emphasizing situations where answers cannot simply be de-
077 rived from text alone. The design of these questions pushes
078 the boundaries of video analysis, requiring models to draw
079 from rich visual and contextual cues to formulate accurate
080 responses.

081 To quantify temporal complexity, the dataset introduces
082 an innovative approach: certificate length analysis. This
083 method reveals that answering questions often necessi-
084 tates comprehension of extended video segments, adding
085 a layer of challenge that surpasses previous benchmarks like
086 EGOSchema. For instance, the median lengths required for

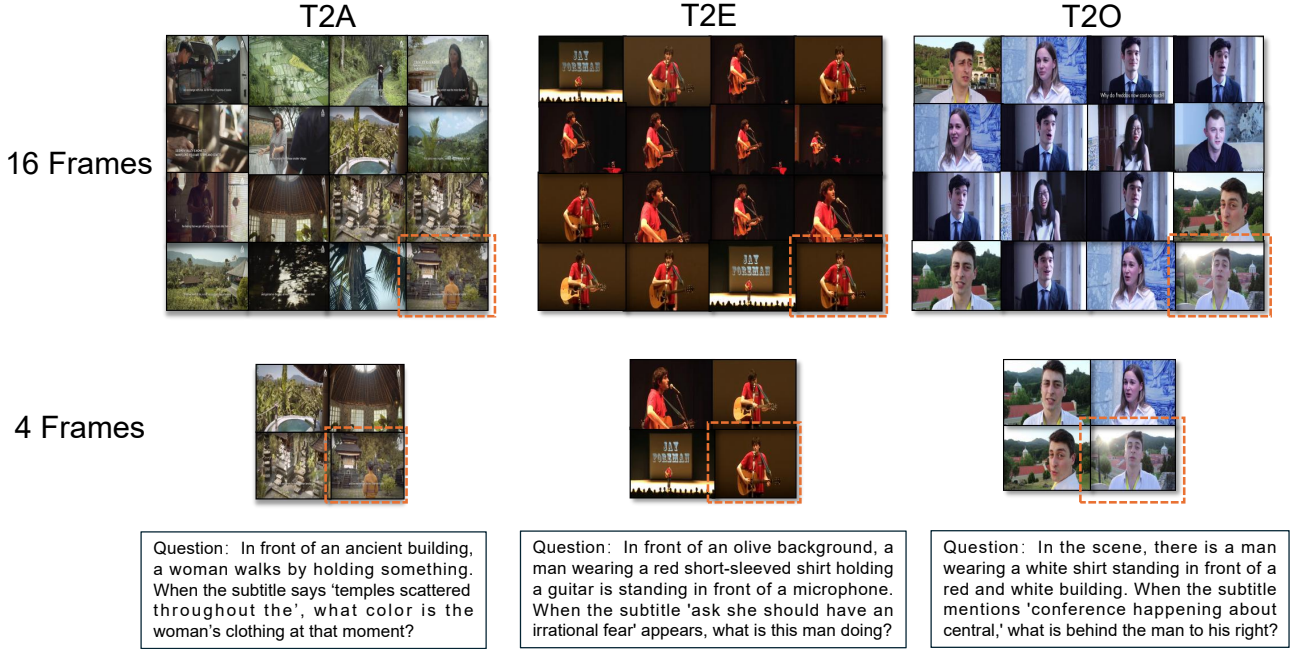


Figure 1. Example diagram of input frame mode in downstream tasks

Method	Searching Modality	Keyframe Source	Text Weight	Frame	Acc \uparrow	LONGVIDEOBENCH	
						Medium	Long
GPT4o+VSI (ours)	Multimodal	Keyframe Searching	0.3	4	14.47	45.61	56.99
GPT4o+VSI (ours)	Multimodal	Keyframe Searching	0.7	4	18.99	55.17	54.95
GPT4o+VSI (ours)	Multimodal	Keyframe Searching	1.0	4	40.00	63.79	68.48
GPT4o+VSI (ours)	Multimodal	Keyframe Searching	0.3	8	20.75	45.61	56.99
GPT4o+VSI (ours)	Multimodal	Keyframe Searching	0.7	8	28.48	56.90	60.44
GPT4o+VSI (ours)	Multimodal	Keyframe Searching	1.0	8	45.00	62.07	69.57

Table 1. Keyframe Search Accuracy and Downstream QA Accuracy on LONGVIDEOBENCH-Text-Related Perception Subsets. Where **Searching Modality** indicates whether the search algorithm uses unimodal or multimodal information, and **Keyframe Source** indicates whether the frames used for VideoQA come from the keyframe search algorithm or uniform sampling. **Text Weight** represents the weight of Text Score in performing score fusion.

087 understanding range from 26 seconds for short videos to a
088 substantial 890.7 seconds for long videos. This emphasis on
089 temporal depth ensures that the evaluation process tests not
090 only the models' ability to process brief clips but also their
091 capacity to maintain coherence and insight over longer du-
092 rations. Through this thorough and challenging framework,
093 the VIDEO-MME dataset sets a new benchmark in video
094 understanding, pushing the capabilities of Vision-Language
095 Models to new heights and encouraging advancements in
096 multi-modal AI research.

097 VIDEO-MME serves as a universal benchmark, offer-
098 ing valuable applicability to both image- and video-focused
099 Multimodal Large Language Models (MLLMs), and it high-

lights pivotal challenges that could shape the trajectory of
future research. Among these challenges is the need to re-
fine architectures to better handle long-sequence processing,
a necessity for capturing the intricacies of extended video
content. Additionally, the creation of datasets that facilitate
complex temporal reasoning is crucial, as they enable mod-
els to grasp the nuanced progression of events over time.
Enhancing cross-modal alignment is another critical area,
ensuring that models can seamlessly integrate information
from diverse sources to form cohesive understanding.

By providing a rigorous evaluation framework, VIDEO-
MME is poised to drive significant progress toward the
development of MLLMs that are adept at comprehending

Method	Searching Modality	Keyframe Source	Text Weight	Frame	Acc \uparrow	LONGVIDEOBENCH	
						Medium	Long
GPT4o+VSI (ours)	Multimodal	Keyframe Searching	0.7	4	28.70	52.83	48.97
GPT4o+VSI (ours)	Multimodal	Keyframe Searching	0.7	8	37.38	55.53	46.54
GPT4o+VSI (ours)	Multimodal	Keyframe Searching	0.7	16	47.44	55.28	49.72
GPT4o+VSI (ours)	Multimodal	Keyframe Searching	0.7	32	57.50	56.02	50.65

Table 2. Keyframe Search Accuracy and Downstream QA Accuracy on LONGVIDEOBENCH. Where **Searching Modality** indicates whether the search algorithm uses unimodal or multimodal information, and **Keyframe Source** indicates whether the frames used for VideoQA come from the keyframe search algorithm or uniform sampling. **Text Weight** represents the weight of Text Score in performing score fusion.

dynamic, real-world scenarios. It encourages researchers to push the boundaries of current technologies, exploring innovative solutions that elevate the capabilities of AI models in processing and interpreting rich, multifaceted media. Through its comprehensive approach, VIDEO-MME not only sets a high standard for evaluation but also inspires advancements that could lead to more sophisticated AI systems capable of navigating the complexities of real-world environments with greater accuracy and insight.

B.2. Details of LONGVIDEOBENCH

The LONGVIDEOBENCH benchmark [?] stands as a trailblazer in the evaluation of long-context interleaved video-language understanding within Vision-Language Models (VLMs), effectively addressing critical gaps found in existing benchmarks. By focusing on detailed retrieval and temporal reasoning over hour-long multimodal inputs, LONGVIDEOBENCH sets out to challenge models in ways previously unexplored. It is designed to counteract the "single-frame bias" that has been a limitation in earlier video benchmarks, thereby pushing models to engage with content across extended sequences rather than isolated frames. The innovative referring reasoning paradigm introduced by LONGVIDEOBENCH empowers models to locate and analyze specific contexts within these lengthy sequences, fostering a deeper understanding of the intricate interplay between video and language.

The dataset is composed of 3,763 web-sourced videos that span a diverse array of themes, including movies, news, life vlogs, and various knowledge domains such as art, history, and STEM. These videos are carefully categorized into four progressive duration levels: 8-15 seconds, 15-60 seconds, 3-10 minutes, and 15-60 minutes. This gradation allows for a comprehensive assessment of VLMs across different temporal scales, from brief clips to extensive content. Each video is paired with aligned subtitles, creating interleaved multimodal inputs that closely mimic real-world viewing scenarios. This pairing ensures that models must not only process visual information but also integrate linguistic context, simulating the complex environment in which humans typically consume video content. Through its groundbreak-

ing approach, the LONGVIDEOBENCH benchmark lays the groundwork for advancing VLM capabilities, challenging them to achieve higher levels of comprehension and insight in the dynamic realm of video-language interaction.

The LONGVIDEOBENCH benchmark is distinguished by its inclusion of 6,678 human-annotated multiple-choice questions, meticulously categorized into 17 fine-grained task types distributed across two primary levels: Perception and Relation. The Perception level focuses on tasks that require object and attribute recognition within single scenes, while the Relation level challenges models to engage in temporal and causal reasoning across multiple scenes. This dual-level approach ensures comprehensive evaluation of a model's ability to understand both immediate visual details and broader narrative contexts.

The questions are crafted to include explicit referring queries, such as "When the woman descends the rocky hill...", which serve to anchor the reasoning process to specific moments within the videos. This specificity is crucial for driving models to accurately pinpoint and interpret relevant scenes. With an average question length of 43.5 words, these queries are designed to ensure precision and clarity, demanding a high level of comprehension and detail-oriented analysis from the models.

Temporal complexity within the benchmark is quantified through duration-grouped analysis, pushing models to process up to 256 frames (at a rate of 1 frame per second) for hour-long videos. This requirement significantly exceeds the demands of previous benchmarks, setting a new standard for evaluating the capability of models to handle extensive and interleaved video content. By challenging models in this way, the LONGVIDEOBENCH benchmark aims to foster advancements in VLMs, encouraging the development of systems that can navigate and interpret complex, real-world video scenarios with enhanced accuracy and depth.

C. Frame count analysis

The approach depicted in Fig.1 involves concatenating multiple images into a single composite image before inputting it into a large model. This technique is designed to optimize

token consumption, which is particularly beneficial for processing within models that have constraints on input size or token limits. By merging several images into one, the model can handle more visual information simultaneously without exceeding these limitations, potentially improving efficiency and resource utilization.

This method allows the model to perceive and analyze relationships between different visual elements within a single input, which might enhance its ability to understand context and draw connections across the concatenated images. However, it also necessitates careful consideration of image arrangement and scaling to ensure that important details are preserved and accessible to the model during analysis. Overall, this strategy aims to balance the need for comprehensive visual input with the practical constraints of model processing capabilities, contributing to more effective and streamlined utilization of large models in image-based tasks.

Table 2 illustrates a key observation regarding the relationship between the number of top-k frames selected and the accuracy of key frame search. The data indicates that as the number of frames increases, the accuracy of identifying key frames also improves. Notably, when employing 32 frames for answering downstream Video-QA tasks, the accuracy achieved state-of-the-art (SOTA) levels, highlighting a positive correlation between the number of input frames and the accuracy of question answering.

Despite this positive correlation, the need to balance efficiency and accuracy led to a focus on experiments involving data sets with 4 and 8 frames in the main text. This approach allows for faster processing and reduced computational load while still aiming to maintain a reasonable level of accuracy. By concentrating on fewer frames, the experiments seek to identify the optimal trade-off that maximizes efficiency without significantly compromising the accuracy of the results. This strategy is essential for ensuring practical application in scenarios where computational resources and processing time are limited, while still striving to achieve high-performance outcomes in Video-QA tasks.

D. Ablation study of iterative strategies

Epsilon-Greedy Algorithm

The Epsilon-Greedy algorithm is a fundamental exploration-exploitation strategy widely used in multi-armed bandit problems and reinforcement learning. Its core idea is to balance the trade-off between exploiting the currently best-known action and exploring potentially better but untested actions. Specifically, with probability $1 - \epsilon$ (where $\epsilon \in (0, 1)$ is a small constant), the algorithm selects the action with the highest estimated reward (exploitation); with probability ϵ , it randomly selects any action from the available set (exploration). This mechanism ensures that non-optimal actions are still occasionally tried, preventing premature convergence to

suboptimal solutions.

Mathematically, the action selection rule can be formulated as:

$$a_t = \begin{cases} \arg \max_{a \in \mathcal{A}} \hat{\mu}_a(t) & \text{with probability } 1 - \epsilon, \\ \text{uniformly random from } \mathcal{A} & \text{with probability } \epsilon, \end{cases}$$

where a_t denotes the action chosen at time t , \mathcal{A} is the set of all possible actions, and $\hat{\mu}_a(t)$ represents the estimated mean reward of action a up to time t .

UCB1 Algorithm

The UCB1 (Upper Confidence Bound 1) algorithm is another influential exploration-exploitation strategy, distinguished by its use of confidence intervals to guide action selection. It operates under the principle that an action’s value should be estimated not only by its observed average reward but also by the uncertainty in that estimate—actions with higher uncertainty (i.e., less frequently selected) are prioritized to reduce this uncertainty.

The key formula for UCB1’s action selection is:

$$a_t = \arg \max_{a \in \mathcal{A}} \left(\hat{\mu}_a(t) + \sqrt{\frac{2 \ln t}{n_a(t)}} \right),$$

where $\hat{\mu}_a(t)$ is the average reward of action a up to time t , t is the total number of steps taken so far, and $n_a(t)$ is the number of times action a has been selected by time t . The term $\sqrt{\frac{2 \ln t}{n_a(t)}}$ represents the upper confidence bound, which decreases as the action is selected more frequently (reducing uncertainty) and increases with the total number of steps (accounting for cumulative information).

D.1. Comparative Experiments on Iterative Strategies

As shown in the table 3, the original iteration strategy significantly outperforms the other compared strategies in terms of task completion efficiency under the same task objectives and test environment conditions. Specifically, this strategy requires the least number of iterations and has the shortest task completion delay; combining the core evaluation dimensions (iteration efficiency, latency) set in this study, the initial iterative strategy shows the optimal comprehensive performance under the current experimental scenarios and constraints.

Method	Strategy	Iteration ↓	Latency(sec) ↓
VSI	original	13.65	18.28
VSI	epsilon_greedy	15.28	26.17
VSI	UCB	16.25	19.18

Table 3. Comparative Experiments on Iterative Strategies.

280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330

E. Case Study of VSI Keyframe Selection

To delve deeper into the algorithmic intricacies, we'll employ a specific example from the LONGVIDEOBENCH-TOS task to illustrate the process. This example highlights the question: *"When the subtitle mentions 'conference happening about central,' what is behind the man to his right?"* This query necessitates utilizing text information for spatial localization, a task that unimodal approaches struggle with. Our method initiates by performing similarity matching between the question and the subtitles, meticulously evaluating the correspondence between textual elements. Subsequently, frames associated with the subtitles are scored according to their matching degree, allowing us to determine their relevance to the query.

These scores, once combined with those derived from the Video Search Stream, are instrumental in updating the sampling probability of the object detection model. This fusion of information empowers the model to swiftly locate all pertinent objects. In this instance, the sentence exhibiting the highest similarity, as calculated through text similarity analysis, is *"conference happening about central."* Utilizing this information facilitates pinpointing the target key frame with precision. When the text weight is judiciously calibrated, the ground truth key frame emerges within the top k keyframes selected based on the final score distribution, demonstrating the efficacy of the approach in capturing and utilizing nuanced textual cues for accurate localization.

Through this example, we gain insight into how **VSI** adeptly processes text-related information, elucidating the reasons behind its impressive performance in text-related tasks. By leveraging advanced similarity matching techniques, **VSI** can effectively correlate textual queries with subtitles, extracting pertinent contextual information that guides the identification and localization of key frames.

The approach demonstrates a nuanced understanding of text, allowing it to discern subtleties and nuances that unimodal methods often overlook. By integrating text similarity scores with visual cues from the Video Search Stream, **VSI** creates a comprehensive framework that enhances its ability to identify key objects and frames with remarkable precision. This multi-faceted strategy ensures that the model is not only responsive to visual elements but is also attuned to the intricate details embedded within the textual narrative.

Ultimately, the example illustrates **VSI**'s proficiency in harmonizing text and visual information, showcasing its capability to navigate complex scenarios where textual data plays a critical role. This synergy between modalities is pivotal in achieving superior results in tasks that require a deep understanding of both language and imagery, underscoring the robustness and versatility of **VSI** in handling text-related challenges effectively.

Listing 1. Case Study

```
1 "video_id": "5qMcDQd17Y4",
2 "video_path": "5qMcDQd17Y4.mp4",
3 "question": "In the scene, there is a man
  wearing a white shirt standing in front
  of a red and white building. When the
  subtitle mentions 'conference happening
  about central,' what is behind the man to
  his right?",
4 "options": "A) cell phone\nB) car\nC) computer
  \nD) street light",
5 "answer": "D",
6 "duration_group": 600,
7 "position": [1290],
8 "question_category": "T2O",
9 "grounding_objects": {
10   "target_objects": ["man", "white shirt", "
    building"],
11   "cue_objects": ["red", "white", "right"]
12 }
```

F. Prompt Design

In this section, we include the prompts designed for environment representation, focusing on query grounding and question answering tasks.

F.1. Prompt for Query Grounding

The following is the prompt used by our system for query grounding:

Listing 2. Prompt Template for Query Grounding

```
1 <system prompt>
2 Here is a video:
3 <image>
4 <image>
5 <image>
6 ...
7 Here is a question about the video:
8 Question: <Question>
9 Options: <Options>
10
11 When answering this question about the video:
12 1. What key objects to locate the answer?
13   - List potential key objects (short
    sentences, separated by commas).
14 2. What cue objects might be near the key
    objects and might appear in the scenes?
15   - List potential cue objects (short
    sentences, separated by commas).
16
17 Please provide your answer in two lines,
    directly listing the key and cue objects,
    separated by commas.
```

This prompt is designed to generate a representation of the environment that facilitates the grounding of queries in a structured, object-centric manner.

341 F.2. Prompt for Question Answering

342 The following prompt is used to answer questions based on
343 the embodied environment representation. This design is
344 adapted from :

Listing 3. Prompt Template for Question Answering

```
1 <system prompt>
2 Select the best answer to the following
   multiple-choice question based on the
   video.
3 <image>
4 <image>
5 <image>
6 ...
7 Question: <Question>
8 Options: <Options>
9
10 Answer with the option letter from the given
   choices directly.
```