$\mathcal{RTV}$-$Bench$: Benchmarking MLLM Continuous Perception, Understanding and Reasoning through $\mathcal{R}$eal-$\mathcal{T}$ime $\mathcal{V}$ideo

Shuhang Xun[1]*, **Sicheng Tao**[2]*, **Jungang Li**[2,3]*†, **Yibo Shi**[4], **Zhixin Lin**[5], **Zhanhui Zhu**[1], **Yibo Yan**[2,3], **Hanqian Li**[2], **Linghao Zhang**[5], **Shikang Wang**[6], **Yixin Liu**[1], **Hanbo Zhang**[7], **Ying Ma**[1]‡, **Xuming Hu**[2,3]

[1] HIT [2] HKUST (GZ) [3] HKUST [4] XJTU [5] SDU [6] CityU [7] HUST

Project: https://ljungang.github.io/RTV-Bench

Abstract

Multimodal Large Language Models (MLLMs) have made rapid progress in perception, understanding, and reasoning, yet existing benchmarks fall short in evaluating these abilities under continuous and dynamic real-world video streams. Such settings require models to maintain coherent understanding and reasoning as visual scenes evolve over time. We introduce $\mathcal{RTV}$-$Bench$, **a fine-grained benchmark for real-time video analysis with MLLMs**. It is built upon three key principles: multi-timestamp question answering, hierarchical question structures spanning perception and reasoning, and multi-dimensional evaluation of continuous perception, understanding, and reasoning. $\mathcal{RTV}$-$Bench$ comprises 552 diverse videos and 4,608 carefully curated QA pairs covering a wide range of dynamic scenarios. We evaluate a broad range of state-of-the-art MLLMs, including proprietary, open-source offline, and open-source real-time models. Our results show that real-time models generally outperform offline counterparts but still lag behind leading proprietary systems. While scaling model capacity generally yields performance gains, simply increasing the density of sampled input frames does not consistently translate into improved results. These observations suggest inherent limitations in current architectures when handling long-horizon video streams, underscoring the need for models explicitly designed for streaming video processing and analysis.

1 Introduction

The ability to comprehend and respond to complex real-world scenarios in real time remains a fundamental challenge in the pursuit of general artificial intelligence [6, 25, 10]. Motivated by the remarkable success of large language models (LLMs) across a broad spectrum of tasks [26, 11, 2], multimodal large language models (MLLMs) have recently emerged as a promising paradigm for visual scene understanding and reasoning [49, 42, 5]. In particular, the research trajectory of Video-LLMs [22, 28, 4, 36, 34] has evolved from early studies focused on short, vision-centric video clips [16, 14, 23, 20] toward more comprehensive modeling of long-form video content. An increasing body of work integrates omni-modal signals—including video, audio, and subtitles [3, 7, 18, 33, 19, 27]—to support richer contextual understanding and more robust long-horizon reasoning.

*Equal contribution. Emails: Shuhang Xun (24s103400@stu.hit.edu.cn)

†Project Leader. Emails: ljungang.02@gmail.com.

‡Corresponding author. Email: y.ma@hit.edu.cn.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

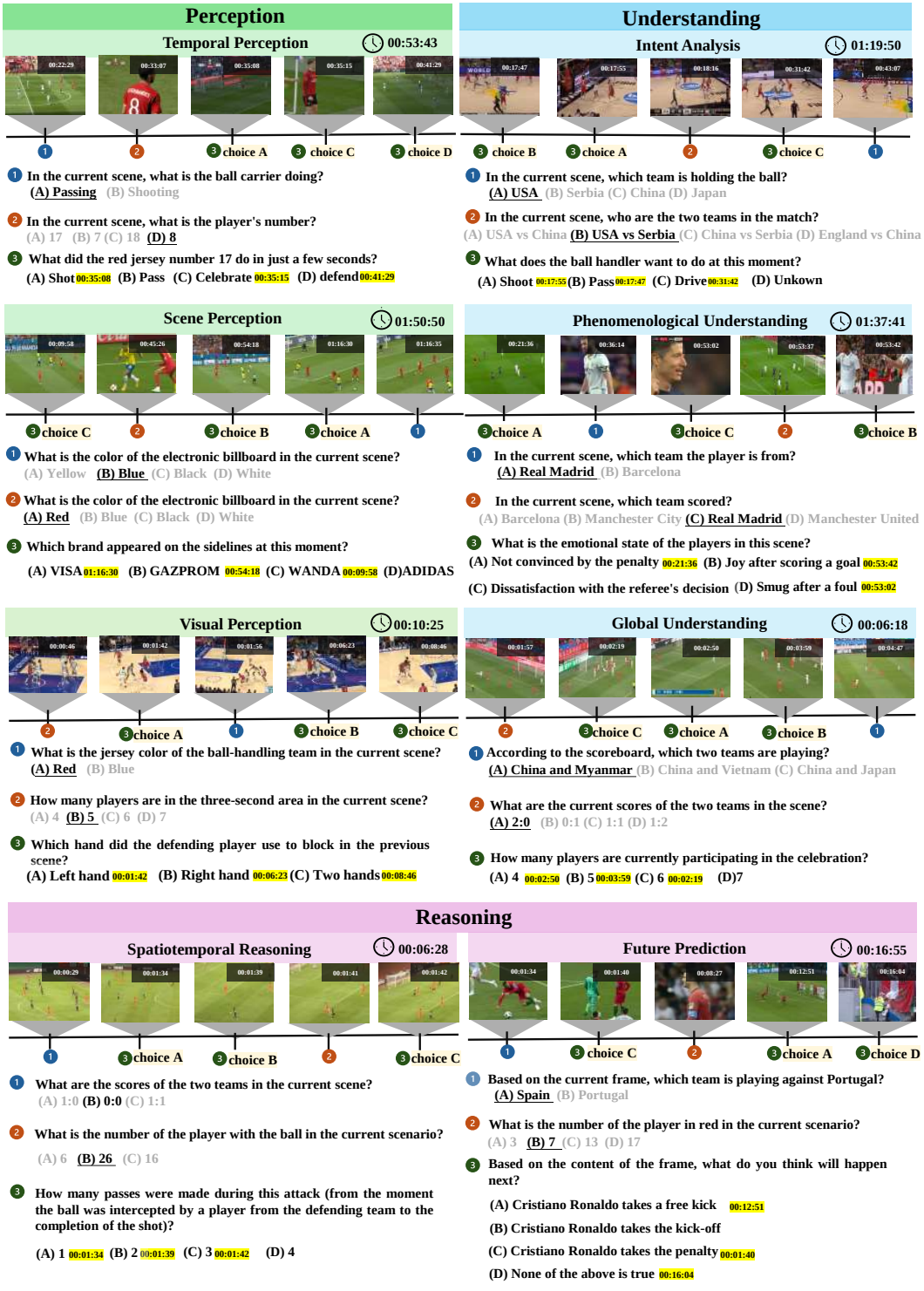


Figure 1: **Representative examples illustrating the diverse task types evaluated in $\mathcal{RTV}$-*Bench*.** 1 and 2 denote fundamental questions within a question group, with their corresponding answers underlined. 3 indicates a dynamically answered question, where the correct response is determined by the query time. As the visual content evolves, the correct answer may change over time; we therefore annotate the appropriate answers corresponding to different query timestamps.

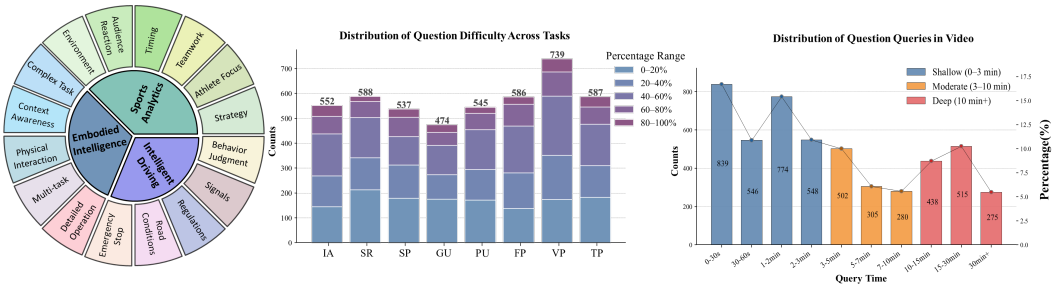


Figure 2: **Video categories and distributions of question difficulty and query characteristics.** (Left) RTV-Bench covers three key domains and 16 video subcategories. (Center) Distribution of question difficulty levels across eight representative task types, quantified by percentage-based performance ranges. (Right) Distribution of question queries with respect to video length, categorized into Shallow, Moderate, and Deep levels. Bar heights indicate query counts, while the overlaid line chart shows the proportion of queries within each duration bucket.

Recently, VStream [46] was the first to attempt to test this capability, with a focus primarily on extending the duration of videos. In addition, both StreamingBench [40] and OVOBench [21] have made varying degrees of improvements in the types and standards of assessment. However, their evaluation of real-time responsiveness is often inadequate, overlooking the capacity to capture transitions and fleeting details from visual input that arrive sequentially, not instantaneously. This limitation highlights the need for a more focused assessment of continuous analysis abilities.

Based on the above considerations, we introduce $\mathcal{RTV}$-$Bench$, featuring three core innovations designed to benchmark the **continuous analysis capabilities**—specifically perception, understanding, and reasoning—of MLLMs within real-time video contexts. First, the **Multi-Timestamp Q&A** mechanism challenges real-time tracking and state update by posing queries whose answers evolve within a video. Crucially, unlike benchmarks like OVO-Bench [21] that typically introduce different questions at different timestamps, RTV-Bench revisits the same conceptual query, where only the correct answer shifts as the scene unfolds. This approach more rigorously tests the model capacities for continuous analysis in real-time scenarios, surpassing single-query-single-answer and static evaluations. Second, the **Hierarchical Question Structure** enforces reliable, sequential reasoning by employing a basic-to-advanced design where higher-order questions logically depend on grasping foundational perceptions and understanding, thus mitigating cognitive shortcuts. Finally, **Multidimensional Evaluation** moves beyond aggregate scores to provide fine-grained diagnostic insights, assessing model performance across eight dimensions that are critical for continuous analysis in dynamic scenarios. This evaluation offers a more informative view of model capabilities and limitations in real-time video understanding; detailed examples are presented in Figure 1.

Through the novel design of $\mathcal{RTV}$-$Bench$ and the comprehensive evaluations presented in this work, we provide systematic insights into the current state of MLLMs for continuous video analysis. Results across all evaluated models reveal substantial bottlenecks in real-time video understanding: ❶ most models achieve accuracies below 50%; ❷ overall performance shows a clear positive correlation with model scale, whereas increasing the number of input frames yields only marginal and non-monotonic gains; and ❸ models explicitly designed for streaming video processing consistently outperform traditional offline video models. Notably, even the lowest-performing real-time model evaluated, VITA-1.5 [9], surpasses a representative offline counterpart, VideoLLaMA2 [4]. Building on these findings, we further discuss promising research directions for online video analysis with MLLMs.

2 Real-Time Video Understanding for MLLMs: RTV-Bench

In this section, we discuss a common challenge in both long video comprehension and real-time video analysis—the continuous analysis capability of MLLMs. To tackle this, we have meticulously curated a benchmark to assess model performance in real-time and continuous scenarios, and developed new metrics to characterize the performance of multimodal large models in this bench.

2.1 Challenge for Real-Time Video Analysis

Existing benchmarks for long video analysis [8, 39] attempt to gauge the true capacity of MLLMs by extending video duration, enhancing the difficulty of QA tasks, and introducing new sub-tasks. Other works [24] focus on real-time video scenarios, evaluating model responsiveness to queries in real-time contexts. However, current models face the risk of memory loss and attention shift, both in long and real-time video contexts. We categorize these challenges as failures in the continuous analysis capability of MLLMs. In video scenarios, MLLMs are considered to possess strong continuous analysis ability only if they can effectively recall prior visual information to accurately respond to user queries (similar to needle-in-a-haystack tasks in natural language settings) and maintain robust perception of streams of new visual data and queries (akin to multimodal multi-turn dialogues).

2.2 Benchmark Overview

The RTV-Bench is designed to assess a model's ability to perceive, understand, and reason in long video and real-time streaming contexts. Models should be adept at recognizing correct temporal query patterns. For example, during live sports queries (*e.g.*, goalkeeper actions), models must prioritize current contexts over historical data. However, it should also retrieve relevant information from memory, such as identifying the goalkeeper from earlier footage. This necessitates that models not only possess exceptionally high fundamental scene comprehension capabilities, but also demonstrate effective understanding of all stages within the temporal flow—namely, comprehensive spatiotemporal comprehension encompassing the past, present, and future. To this end, we propose a series of sub-tasks for video continuous analysis, addressing both spatiotemporal elements within the video and the intrinsic capabilities of the model. The intrinsic model capabilities are divided into eight categories: Temporal Perception (TP), Scene Perception (SP), Visual Perception (VP), Future Prediction (FP), Phenomenological Understanding (PU), Intent Analysis (IA), Global Understanding (GU), and Spatiotemporal Reasoning (SR). An overview of these task categories is provided in Figure 1.

2.3 Benchmark Construction

Terminology The RTV-Bench dataset comprises 552 videos sourced from the internet. A key characteristic is its question structure designed to directly probe temporal dynamics. For each video, questions are organized into sets. Each question set contains about three multiple-choice questions.

Crucially, the core design principle centers on time-varying correct answers through Multi-Timestamp QA. Specifically, within a set of questions, the same underlying conceptual query (*e.g.*, "What is person A holding?" or "Where is the car heading?") is evaluated at multiple points in time throughout the video. As a result, the correct answer may differ depending on the specific timestamp or temporal interval referenced—or implicitly required—by the question context or its earliest inferable timestamp (Figure 1). Rather than merely locating relevant information, models are therefore required to actively track temporal changes and continuously update their understanding as the scene evolves. This design directly targets the evaluation of continuous temporal understanding and a model's sensitivity to dynamic state transitions; representative examples are shown in Figure 4.

The question set features three questions of escalating difficulty. The first two are simpler, while the final question is significantly more complex, demanding integration of broader context or clues. The information and reasoning needed for the final question generally encompass those required for the first two. Successfully answering the complex third question strongly suggests the capability to also answer the simpler ones. This structure evaluates the model's ability to handle increasing complexity, synthesize comprehensive context, and perform robust analysis.

In addition, we emphasize diversity through the richness of specific sub-scenes and the distribution of video lengths. The RTV-Bench primarily encompasses intelligent driving, sports events, and egocentric videos—categories rich in dynamic information and real-time contextual relevance. Each category includes a variety of sub-categories, as illustrated in Figure 2.

Data Statistic RTV-Bench comprises 552 videos with a total duration of 167.2 hours (average 18.2 minutes per video) and contains 4,631 QA pairs, for detailed information, refer to Figure 2 and Table 1. The benchmark features diverse question scenarios and evenly distributed video durations, thereby establishing a comprehensive framework for video continuous understanding tasks with rich sub-scenarios and structured problem dimensions.

Table 1: **Comparison of video QA benchmarks.** MT denotes multi-timestamp questions. Labels: A (automatic), M (manual), A+M (both).

Benchmark	MT	#QA (k)	Avg. Duration (min)	Total Duration (h)	Labels
Video-MME [8]	✗	2.7	17.00	254.0	M
MSRVTT [41]	✗	73.0	0.25	12.5	A
MSVD-QA [1]	✗	13.0	0.17	1.4	A
MovieChat-1k [32]	✗	13.0	9.40	156.0	M
MVBench [17]	✗	4.0	0.25	34.5	A+M
ActivityNet-QA [44]	✗	58.0	1.87	25.0	M
Vstream-Q [46]	✗	3.5	40.00	21.0	–
StreamBench [38]	✗	1.8	4.50	25.0	M
OVO-Bench [21]	✗	2.8	6.00	66.0	A+M
OVBench [12]	✓	7.0	–	–	A
RTV-Bench (Ours)	✓	4.6	18.00	167.2	M

Video Collection and Filtering Our video data sources include EgoSchema [29] and publicly available online videos. Unlike most existing benchmarks, we incorporated manual review during the collection phase. Three data collectors sourced videos from various domains and manually excluded highly similar videos, focusing on long videos with high dynamics and real-time needs. These sources ensure targeted and scientific evaluation of video models' real-time capabilities.

Manual Annotation Rigorous manual annotation by qualified experts underpins the reliability of RTV-Bench for evaluating continuous video understanding. We leverage an LLM (DeepSeek [26]) only to produce initial question templates, and human annotators then refine every question to better reflect dynamic scenes and the demands of temporal reasoning. In particular, annotators intentionally craft questions whose correct answers evolve over time and systematically determine the earliest valid timestamp associated with each answer option in the MTQA setting. This human-centric, multi-annotator protocol strengthens annotation robustness and enables RTV-Bench to explicitly evaluate models' sensitivity to temporal dynamics in video.

Quality Control To ensure the benchmark's quality, each video and its corresponding Q&A pairs underwent multiple rounds of review. We manually filtered videos based on length distribution and the presence of sub-scenes examining real-time event changes, resulting in high-quality videos focused on real-time analysis tasks. We conducted manual video-question alignment and precise timestamp checks, utilizing GPT-4 and human review to verify annotation format and sensitive information.

3 Experiments

3.1 Experiment Setup

Our experiments were conducted on two NVIDIA A800 GPUs to comprehensively evaluate the performance of mainstream multimodal large language models (MLLMs) on our benchmark. We consider a diverse set of representative models, including VideoLLaMA2 [4], VideoLLaMA3 [45], GPT-4o [13], InternLM-XComposer2.5-OmniLive (IXC2.5-OL) [47], VITA1.5 [9], VideoChat-Online (4B) [12], LLaVA-Video [22], LLaVA-OneVision [15], and Qwen2.5-VL [43].

To ensure a fair comparison under comparable computational budgets, most models are evaluated using configurations around the 7B scale when available, while smaller models (e.g., VideoChat-Online at 4B) are evaluated at their native parameter size. All models are tested under a unified uniform frame sampling protocol. Specifically, for models that support variable frame inputs (e.g., Qwen2.5-VL), we evaluate multiple sampling settings with 8, 16, 32, and 64 uniformly sampled frames. For each model, we report in the main tables the best-performing configuration across different frame counts, following the same evaluation protocol for all baselines. Other models are evaluated analogously using their supported frame sampling ranges, and their strongest results are reported for comparison.

Real-Time Video Model vs. Offline Video Model We compare two model categories: traditional offline models (M_{offline}) and novel real-time online models (M_{online}). These categories differ significantly in architecture, training, and data requirements. Architecturally, M_{offline} typically employ

sequential vision encoder-decoders, often limited by fixed context windows and incurring higher processing latency. In contrast, M_{online} are designed for continuous, low-latency ingestion of the video stream V. They prioritize maintaining an internal state S_t that summarizes the video information processed up to time t (*i.e.*, $V[0, t]$), enabling real-time responsiveness. Models like IXC2.5-OL [47] implement this using techniques like modular parallelism and dedicated long-term memory. These architectural distinctions lead to different training paradigms and data needs. M_{offline} usually rely on end-to-end fine-tuning using standard annotated video datasets. M_{online}, however, often require specialized training strategies (*e.g.*, VITA-1.5's staged fusion, IXC2.5-OL's targeted training for memory and interaction) and benefit most from specialized corpora designed for long-duration, interactive streaming scenarios. The ability of M_{online} to maintain and update state S_t is particularly relevant for RTV-Bench's Multi-Timestamp QA (MTQA) challenge, where the correct answer $A^*(Q, t_q)$ to a query Q depends on the specific query time t_q. To evaluate both model types on this benchmark, we adapt the testing procedure. For M_{online}, queries Q are presented at their timestamp t_q, and the models leverage their continuously updated state S_{t_q} to generate the answer $A_{M_{\text{online}}} = M_{\text{online}}(Q, t_q | S_{t_q})$. Since M_{offline} lack this inherent streaming capability, we simulate real-time interaction for them: when a query Q_i is posed at time $t_{q,i}$, we extract and provide only the relevant video segment V_i (corresponding to the query's context). The offline model's answer is thus based solely on this isolated segment: $A_{M_{\text{offline}},i} = M_{\text{offline}}(Q_i, V_i)$.

To evaluate the performance of these models, we employed two metrics: **Accuracy** and **Score**.

Accuracy. The accuracy metric measures the proportion of correct answers provided by the model compared to the ground truth.

Score. The score metric evaluates the model's ability to correctly answer advanced-level questions (type `q2`), contingent upon its demonstrated mastery of prerequisite basic questions (types `q0` and `q1`) within the same question group, thus emphasizing reliable advanced reasoning built upon a solid foundation. Calculation involves a prerequisite check for each group i: if all basic questions are correct ($B_i = 1$), the group contributes points equal to the number of correctly answered `q2` questions ($N_{q2,i}^{\text{correct}}$); otherwise ($B_i = 0$), it contributes zero points. The final score is the ratio of total conditionally awarded points to the total number of `q2` questions across all N valid groups (those containing `q2`). Formula:

$$\text{Score} = \frac{\sum_{i=1}^{N} B_i \cdot N_{q2,i}^{\text{correct}}}{\sum_{i=1}^{N} N_{q2,i}^{\text{total}}}$$

where N is the number of valid groups, B_i is the prerequisite indicator (1 if basics are correct, 0 otherwise) for group i, $N_{q2,i}^{\text{correct}}$ is the count of correct `q2` answers in group i, and $N_{q2,i}^{\text{total}}$ is the total count of `q2` questions in group i. Advantages of this metric include: ensuring foundational accuracy by rewarding advanced correctness only when basics are mastered; reflecting model robustness by penalizing superficial success on complex tasks without fundamental understanding; and aligning with hierarchical learning principles where complex skills build upon simpler ones.

Table 2: Evaluation results on RTV-Bench. **Perception**, **Understanding**, and **Reasoning** denote different task categories. **FQA** refers to foundational video question answering without multi-timestamp supervision. **MTQA** refers to multi-timestamp question answering with time-varying correct answers. Scores are computed using group-aware Q2 evaluation.

Model	**#Size**	**Perception** Acc (%) / Score	**Understanding** Acc (%) / Score	**Reasoning** Acc (%) / Score	**FQA** Acc (%)	**MTQA** Acc (%)	**Overall** Acc (%) / Score
Open-Source Offline Video Models							
Qwen2.5-VL [43]	7B	42.30 / 7.70	39.85 / 7.00	38.16 / 6.90	44.07	37.46	40.41 / 7.13
VideoLLaMA2 [4]	7B	40.62 / 8.67	39.85 / 7.77	37.49 / 6.75	45.77	34.95	39.55 / 7.90
VideoLLaMA3 [45]	7B	37.98 / 5.83	35.29 / 5.73	35.78 / 6.80	38.62	34.91	36.42 / 6.10
LLaVA-OneVision [15]	7B	35.38 / 3.97	34.21 / 4.63	33.57 / 4.95	35.80	33.58	34.49 / 4.40
LLaVA-Video [22]	7B	35.83 / 5.03	33.81 / 3.77	35.15 / 5.75	36.28	34.17	34.90 / 4.80
Open-Source Online Models							
VITA-1.5 [9]	7B	45.66 / 12.80	44.12 / 11.83	43.37 / 10.15	55.06	36.32	44.51 / 11.80
IXC2.5-OL [47]	7B	47.21 / 15.87	48.22 / 15.23	46.18 / 14.45	**59.05**	38.21	47.33 / 15.40
VideoChat-Online [12]	4B	46.86 / 12.30	46.34 / 12.80	43.53 / 11.00	55.16	38.21	45.83 / 12.10
Closed-Source Business Models							
GPT-4o [13]	–	**51.61 / 21.90**	**49.31 / 20.76**	**48.71 / 23.95**	56.53	**44.73**	**50.02 / 22.10**
Gemini 2.0 Flash [35]	–	41.67 / 11.00	42.71 / 12.73	41.44 / 12.05	47.49	38.64	42.00 / 12.00

3.2 Experiment Results

Online vs. Offline Models. As shown in Table 2, online models optimized for real-time processing—particularly IXC2.5-OL—surpass offline counterparts in overall performance metrics. IXC2.5-OL achieves 47.33% Accuracy and 15.40 Score, significantly outperforming offline models like VideoLLaMA2 (39.55% Accuracy / 7.90 Score). Furthermore, when comparing online models, IXC2.5-OL demonstrates a clear advantage over VITA-1.5 with improvements of 2.82% in Accuracy and 3.6 points in Score. A notable performance gap emerges in temporal analysis tasks: IXC2.5-OL attains 38.21% Accuracy in Multi-Timestamp Question Answering (MTQA), substantially higher than the 33–35% range typical of offline models. This notable discrepancy suggests that current online models may be promising avenues toward continuous analysis capabilities.

Table 3: Detailed evaluation results on the category of **Perception**. Temporal Perception (TP), Visual Perception (VP) and Scene Perception (SP).

Method	#Size	TP Acc (%) / Score	VP Acc (%) / Score	SP Acc (%) / Score	Overall Acc (%) / Score
Open-Source Offline Video Models					
Qwen2.5-VL [43]	7B	39.35 / 6.7	45.47 / 9.1	41.15 / 6.9	42.30 / 7.7
VideoLLaMA2 [4]	7B	39.52 / 7.9	42.49 / 9.4	39.85 / 8.7	40.62 / 8.67
VideoLLaMA3 [45]	7B	37.82 / 6.1	39.24 / 7.6	36.87 / 3.8	37.98 / 5.83
LLaVA-OneVision [15]	7B	35.09 / 3.9	35.86 / 3.8	35.20 / 4.2	35.38 / 3.97
LLaVA-Video [22]	7B	34.07 / 4.8	38.97 / 5.8	34.45 / 4.5	35.83 / 5.03
Open-Source Online Models					
VITA-1.5 [9]	–	46.51 / 12.1	47.09 / 13.2	43.39 / 13.1	45.66 / 12.8
IXC2.5-OL [47]	7B	**49.57** / 17.6	49.80 / 16.5	42.27 / 13.5	47.21 / 15.87
VideoChat-Online [12]	4B	48.55 / 13.3	48.58 / 14.43	42.64 / 8.3	46.86 / 12.3
Closed-Source Business Models					
GPT-4o [13]	–	48.60 / **18.2**	**53.59 / 23.4**	**52.63 / 24.1**	**51.61 / 21.90**
Gemini 2.0 Flash [35]	–	40.49 / 9.5	45.19 / 16.1	39.34 / 7.4	41.67 / 11.00

Table 4: Detailed evaluation results on the category of **Understanding**. Phenomenological Understanding (PU), Global Understanding (GU) and Intent Analysis (IA).

Method	#Size	GU Acc (%) / Score	PU Acc (%) / Score	IA Acc (%) / Score	Overall Acc (%) / Score
Open-Source Offline Video Models					
Qwen2.5-VL [43]	7B	36.92 / 5.8	42.02 / 6.9	40.22 / 8.2	39.85 / 7.00
VideoLLaMA2 [4]	7B	37.34 / 7.6	42.21 / 9.5	40.92 / 6.2	39.85 / 7.77
VideoLLaMA3 [45]	7B	33.54 / 5.8	39.13 / 5.9	33.39 / 4.3	35.35 / 5.33
LLaVA-OneVision [15]	7B	32.07 / 4.3	33.51 / 3.0	37.06 / 6.6	34.21 / 4.63
LLaVA-Video [22]	7B	29.42 / 2.5	35.69 / 3.9	36.33 / 4.9	33.81 / 3.77
Open-Source Online Models					
VITA-1.5 [9]	7B	40.30 / 7.2	46.01 / 15.1	46.06 / 13.2	44.12 / 11.83
IXC2.5-OL [47]	7B	43.88 / 11.9	52.17 / 18.7	**48.62** / 15.1	48.22 / 15.23
VideoChat-Online [12]	4B	42.19 / 8.3	48.99 / 13.82	47.28/ 15.7	46.34 / 12.8
Closed-Source Business Models					
GPT-4o [13]	–	**45.02 / 15.7**	**54.32 / 25.8**	48.58 / **20.8**	**49.31 / 20.76**
Gemini 2.0 Flash [35]	–	35.70 / 10.6	45.65 / 11.3	46.78 / 16.3	42.71 / 12.73

Open-Source vs. Close-Source Models. While a performance gap persists compared to leading closed-source models like GPT-4o, state-of-the-art online architectures demonstrate remarkable progress. The online model IXC2.5-OL achieves near-top-tier performance with 47.33% Overall Accuracy and 15.40 Score, substantially outperforming mid-range closed-source systems like Gemini 2.0 Flash. Notably, IXC2.5-OL closes the accuracy gap with GPT-4o to 4.4% in perception tasks (Tables 3) and 1.1% in video understanding, demonstrating competitive performance in multimodal domains. However, limitations emerge in complex reasoning where GPT-4o maintains decisive advantages, especially on complex tasks like **Understanding** and **Reasoning** (Tables 4 and 5). This pattern highlights that while modern online models have approached entry-level commercial systems and even challenged premium models in specific competencies, structural innovations remain critical to bridge gaps in advanced cognitive tasks like multi-step reasoning and temporal analysis.

Table 5: Detailed evaluation results on the category of **Reasoning**. Future Prediction (FP) and Spatiotemporal Reasoning (SR).

Method	#Size	FP Acc (%) / Score	SR Acc (%) / Score	Overall Acc (%) / Score
Open-Source Offline Video Models				
Qwen2.5-VL [43]	7B	42.49 / 9.7	33.84 / 4.2	38.16 / 6.90
VideoLLaMA2 [4]	7B	41.47 / 7.5	33.50 / 6.0	37.49 / 6.75
VideoLLaMA3 [45]	7B	38.05 / 6.9	33.84 / 3.9	35.95 / 5.40
LLaVA-OneVision [15]	7B	38.23 / 7.2	28.91 / 2.7	33.57 / 4.95
LLaVA-Video [22]	7B	39.08 / 9.1	31.22 / 2.4	35.15 / 5.75
Open-Source Online Models				
VITA-1.5 [9]	7B	47.95 / 12.2	38.78 / 8.1	43.37 / 10.15
IXC2.5-OL [47]	7B	51.88 / 18.1	40.48 / 10.8	46.18 / 14.45
VideoChat-Online [12]	4B	48.12 / 14.69	38.95 / 7.5	43.53 / 11.0
Closed-Source Business Models				
GPT-4o [13]	–	**54.67 / 27.1**	**42.75 / 20.8**	**48.71 / 23.95**
Gemini [35] 2.0 Flash	–	44.42 / 13.6	38.46 / 10.5	41.44 / 12.05

Impact of Model Scales. We analyze the effect of model scale by evaluating Qwen2.5-VL from 3B to 72B parameters under different frame sampling budgets (8–64 frames), as shown in Figure 3(c). Overall accuracy exhibits a clear and largely monotonic improvement with increasing model size. Specifically, the 72B model consistently achieves the highest performance across all frame settings, reaching up to 40.78% with 64 frames, while smaller models (3B–32B) remain below 40%.

Notably, scaling benefits are consistent but moderate, with absolute gains from 3B to 72B on the order of $\sim$2–3 points, suggesting diminishing returns at larger scales. In addition, model scale interacts with temporal resolution: larger models benefit more reliably from increased frame counts, whereas smaller models show non-uniform or even fluctuating trends when additional frames are introduced. These observations indicate that while parameter scaling remains beneficial for real-time video understanding, its effectiveness is increasingly constrained by architectural and temporal modeling

capacities, highlighting the importance of improving temporal representation efficiency beyond naive model enlargement.

Impact of Frame Numbers. Figure 3(c,d) jointly examine the effect of frame sampling density across model scales and architectures. From Figure 3(c), increasing the number of frames from 8 to 64 does not yield consistent accuracy gains across model sizes. While larger models (*e.g.*, 72B) show modest improvements with more frames, smaller and medium-scale models exhibit non-monotonic or saturated trends, indicating limited benefit from denser temporal sampling. In some cases (*e.g.*, 3B and 32B), additional frames lead to marginal gains or even slight regressions.

This phenomenon becomes more pronounced in Figure 3(d), which aggregates performance across models. Average accuracy remains largely stable as frame count increases, while the total score—reflecting global understanding—often declines. Notably, IXC2.5-OL suffers a clear performance drop with more frames, suggesting that excessive temporal inputs may overwhelm the model's effective processing capacity. Together, these results indicate that simply increasing frame numbers is insufficient for improving real-time video understanding and may instead introduce redundancy or attention dilution. This highlights the need for temporally selective, adaptive frame utilization strategies rather than uniform increases in sampling density.

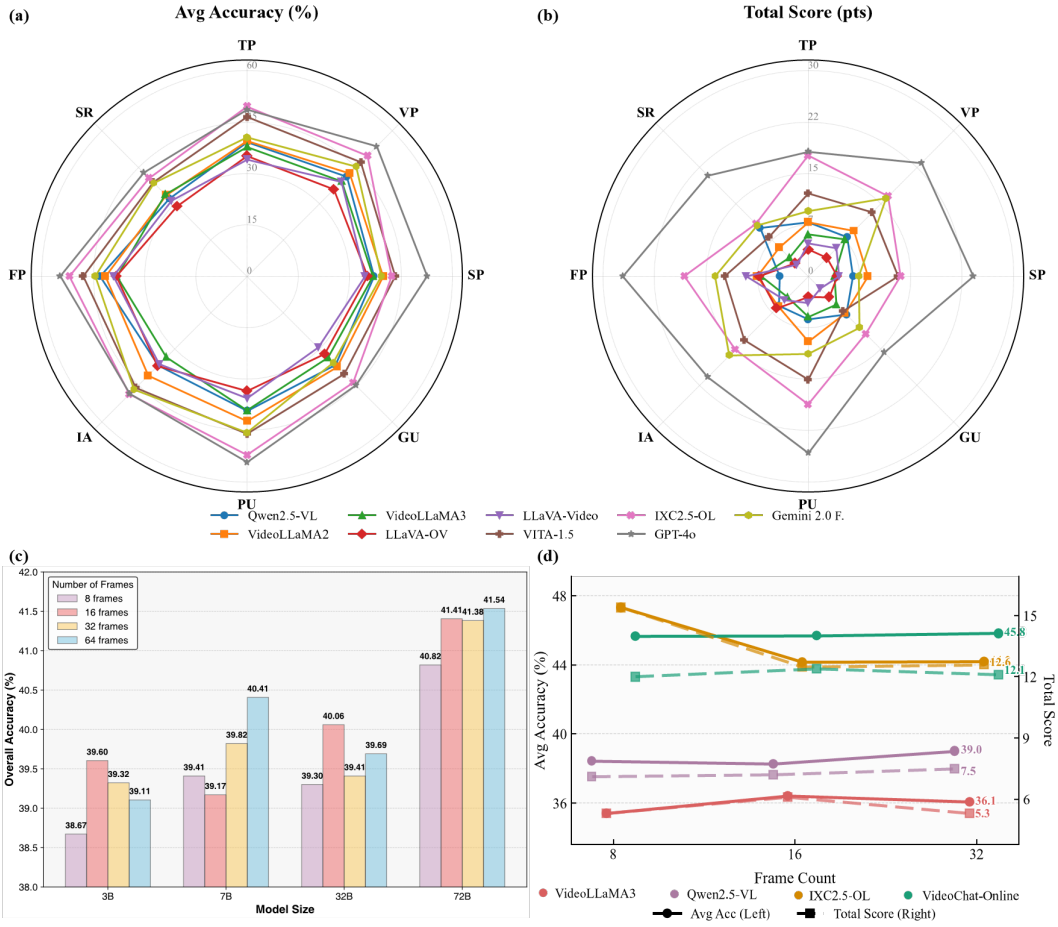


Figure 3: **Performance visualization and analysis on RTV-Bench**: (a) Visualization of overall **Accuracy** results; (b) Visualization of overall **Score** results; (c) Performance impact of varying input frame counts; (d) Performance comparison across different Qwen2.5-VL model scales.

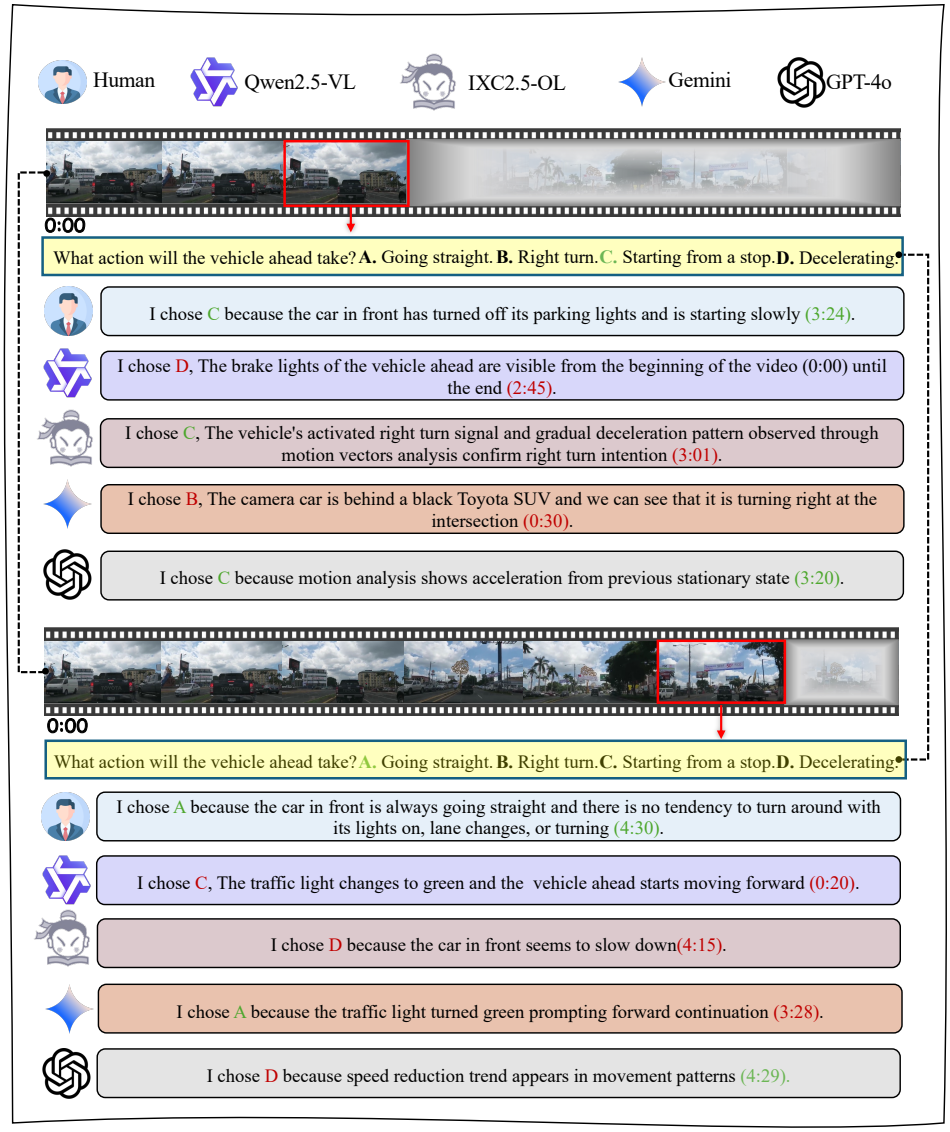


Figure 4: **Comparison of response from different models for the same question on the same video.** Green indicates correct answers or timestamps; red indicates incorrect answers or timestamps. This case demonstrates that even current high-performance models struggle to provide high-quality responses (where both the answer and the corresponding timestamp are accurate).

4 Related Work

4.1 Advanced MLLMs towards Real-time Video Analysis

Video-oriented large-scale models represent a highly promising domain with significant real-world potential, giving rise to numerous valuable applications [28, 8]. Video-LLMs have rapidly advanced, moving from analyzing short videos [16, 22] to longer ones [31, 36, 48] using various techniques. However, most research targets offline video analysis. Real-world applications necessitate real-time, continuous perception and reasoning on unfolding video streams. While models like InternLM-XComposer2.5-OmniLive [47], VITA-1.5 [9], and Dispider [30] are exploring real-time capabilities, rigorously evaluating their ability to continuously track and understand dynamic events remains a critical challenge.

4.2 Existing Video Benchmarks

Existing benchmarks primarily focus on offline evaluation using offline videos [8, 37, 39]. These are less suited for assessing how MLLMs track dynamically changing states, as they often use static question-answer pairs. Newer benchmarks tackle real-time and streaming aspects [24, 21], evaluating responsiveness and contextual understanding in online settings. However, RTV-Bench specifically addresses the gap in evaluating continuous perception, understanding, and reasoning. Its key distinction is the use of dynamic question answering, where the correct answer evolves with the video stream, directly probing the MLLM's ability to maintain and update its understanding of complex, unfolding events over time, complemented by a multi-dimensional evaluation structure.

5 Limitations and Future Work

Our benchmark reveals counter-intuitive findings, such as the limited impact of model scale and input frame count on performance, suggesting that MLLM mechanisms for processing continuous video are poorly understood and effective analysis tools are lacking. Furthermore, the current evaluation is primarily limited to the visual modality. Future work will focus on investigating the underlying causes of these phenomena and developing more adapted analytical methods. Concurrently, a key direction involves incorporating important modalities like audio into RTV-Bench to enable a more comprehensive evaluation of continuous perception, understanding, and reasoning in realistic multimodal scenarios.

6 Conclusion

In this work, we introduce **RTV-Bench**, a benchmark designed to systematically evaluate the continuous video understanding and real-time reasoning capabilities of multimodal large language models (MLLMs). RTV-Bench comprises 552 long-form and streaming-style videos paired with 4,608 carefully constructed QA instances, targeting time-varying perception, understanding, and reasoning under realistic online settings.

Extensive experiments across a diverse set of models reveal two key findings. First, models explicitly designed for online or streaming video processing consistently demonstrate stronger continuous understanding capabilities than general-purpose counterparts. Second, both parameter scaling and uniformly increasing the number of sampled frames during training or inference yield only limited and sometimes inconsistent performance gains. In particular, denser temporal sampling often leads to performance saturation or degradation, highlighting inherent limitations in current architectures when handling long or high-density visual streams. These observations underscore that effective real-time video understanding requires principled temporal modeling and selective information aggregation, rather than relying on naive increases in model size or frame counts.

References

[1] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, pages 190–200, 2011.

[2] Junhao Chen, Yu Huang, Siyuan Li, Rui Yao, Hanqian Li, Hanyu Zhang, Jungang Li, Jian Chen, Bowen Wang, and Xuming Hu. Knowmt-bench: Benchmarking knowledge-intensive long-form question answering in multi-turn dialogues. arXiv preprint arXiv:2509.21856, 2025.

[3] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. Advances in Neural Information Processing Systems, 36:72842–72866, 2023.

[4] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. CoRR, 2024.

[5] Song Dai, Yibo Yan, Jiamin Su, Dongfang Zihao, Yubo Gao, Yonghua Hei, Jungang Li, Junyan Zhang, Sicheng Tao, Zhuoran Gao, et al. Physicsarena: The first multimodal physics reasoning benchmark exploring variable, process, and solution dimensions. arXiv preprint arXiv:2505.15472, 2025.

[6] Yunkai Dang, Mengxi Gao, Yibo Yan, Xin Zou, Yanggan Gu, Aiwei Liu, and Xuming Hu. Exploring response uncertainty in mllms: An empirical evaluation under misleading scenarios. arXiv preprint arXiv:2411.02708, 2024.

[7] Haoyi Duan, Yan Xia, Zhou Mingze, Li Tang, Jieming Zhu, and Zhou Zhao. Cross-modal prompts: Adapting large pre-trained models for audio-visual downstream tasks. Advances in Neural Information Processing Systems, 36:56075–56094, 2023.

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

[9] Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Yangze Li, Zuwei Long, Heting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. arXiv preprint arXiv:2501.01957, 2025.

[10] Xuming Hu, Hanqian Li, Jungang Li, and Aiwei Liu. Videomark: A distortion-free robust watermarking framework for video diffusion models. arXiv preprint arXiv:2504.16359, 2025.

[11] Sirui Huang, Hanqian Li, Yanggan Gu, Xuming Hu, Qing Li, and Guandong Xu. Hyperg: Hypergraph-enhanced llms for structured knowledge. arXiv preprint arXiv:2502.18125, 2025.

[12] Zhenpeng Huang, Xinhao Li, Jiaqi Li, Jing Wang, Xiangyu Zeng, Cheng Liang, Tao Wu, Xi Chen, Liang Li, and Limin Wang. Online video understanding: Ovbench and videochat-online. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 3328–3338, 2025.

[13] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. arXiv preprint arXiv:2410.21276, 2024.

[14] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13700–13710, 2024.

[15] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326, 2024.

[16] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355, 2023.

[17] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 22195–22206, 2024.

[18] Linjun Li, Tao Jin, Wang Lin, Hao Jiang, Wenwen Pan, Jian Wang, Shuwen Xiao, Yan Xia, Weihao Jiang, and Zhou Zhao. Multi-granularity relational attention network for audio-visual question answering. IEEE Transactions on Circuits and Systems for Video Technology, 2023.

[19] Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, Bowen Ding, Wei Song, Zhenglin Cheng, Yuqi Huo, et al. Baichuan-omni technical report. arXiv preprint arXiv:2410.08565, 3(7), 2024.

[20] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. arXiv preprint arXiv:2403.18814, 2024.

[21] Yifei Li, Junbo Niu, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, et al. Ovo-bench: How far is your video-llms from real-world online video understanding? arXiv preprint arXiv:2501.05510, 2025.

[22] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. arXiv preprint arXiv:2311.10122, 2023.

[23] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 26689–26699, 2024.

[24] Junming Lin, Zheng Fang, Chi Chen, Zihao Wan, Fuwen Luo, Peng Li, Yang Liu, and Maosong Sun. Streamingbench: Assessing the gap for mllms to achieve streaming video understanding. arXiv preprint arXiv:2411.03628, 2024.

[25] Zhixin Lin, Jungang Li, Shidong Pan, Yibo Shi, Yue Yao, and Dongliang Xu. Mind the third eye! benchmarking privacy awareness in mllm-powered smartphone agents. arXiv preprint arXiv:2508.19493, 2025.

[26] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437, 2024.

[27] Kai Liu, Jungang Li, Yuchong Sun, Shengqiong Wu, Jianzhang Gao, Daoan Zhang, Wei Zhang, Sheng Jin, Sicheng Yu, Geng Zhan, Jiayi Ji, Fan Zhou, Liang Zheng, Shuicheng YAN, Hao Fei, and Tat-Seng Chua. Javisgpt: A unified multi-modal llm for sounding-video comprehension and generation. In Conference on Neural Information Processing Systems [Spotlight], November 2025.

[28] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 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 (Volume 1: Long Papers), pages 12585–12602, 2024.

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

[30] Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction. arXiv preprint arXiv:2501.03218, 2025.

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

[32] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18221–18232, 2024.

[33] Reuben Tan, Ximeng Sun, Ping Hu, Jui-hsien Wang, Hanieh Deilamsalehy, Bryan A Plummer, Bryan Russell, and Kate Saenko. Koala: Key frame-conditioned long video-llm. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13581–13591, 2024.

[34] Sicheng Tao, Jungang Li, Yibo Yan, Junyan Zhang, Yubo Gao, Hanqian Li, ShuHang Xun, Yuxuan Fan, Hong Chen, Jianxiang He, et al. Moss-chatv: Reinforcement learning with process reasoning reward for video temporal reasoning. arXiv preprint arXiv:2509.21113, 2025.

[35] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530, 2024.

[36] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191, 2024.

[37] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. Lvbench: An extreme long video understanding benchmark. arXiv preprint arXiv:2406.08035, 2024.

[38] Cheng-Kuang Wu, Zhi Rui Tam, Chieh-Yen Lin, Yun-Nung Chen, and Hung-yi Lee. Streambench: Towards benchmarking continuous improvement of language agents. arXiv preprint arXiv:2406.08747, 2024.

[39] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. Advances in Neural Information Processing Systems, 37:28828–28857, 2025.

[40] Haomiao Xiong, Zongxin Yang, Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Jiawen Zhu, and Huchuan Lu. Streaming video understanding and multi-round interaction with memory-enhanced knowledge. arXiv preprint arXiv:2501.13468, 2025.

[41] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5288–5296, 2016.

[42] Yibo Yan, Guangwei Xu, Xin Zou, Shuliang Liu, James Kwok, and Xuming Hu. Docpruner: A storage-efficient framework for multi-vector visual document retrieval via adaptive patch-level embedding pruning. arXiv preprint arXiv:2509.23883, 2025.

[43] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. arXiv preprint arXiv:2412.15115, 2024.

[44] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In AAAI, pages 9127–9134, 2019.

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

[46] Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-vstream: Memory-based real-time understanding for long video streams. arXiv preprint arXiv:2406.08085, 2024.

[47] Pan Zhang, Xiaoyi Dong, Yuhang Cao, Yuhang Zang, Rui Qian, Xilin Wei, Lin Chen, Yifei Li, Junbo Niu, Shuangrui Ding, et al. Internlm-xcomposer2. 5-omnilive: A comprehensive multimodal system for long-term streaming video and audio interactions. arXiv preprint arXiv:2412.09596, 2024.

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

[49] Xin Zou, Di Lu, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Xu Zheng, Linfeng Zhang, and Xuming Hu. Don't just chase" highlighted tokens" in mllms: Revisiting visual holistic context retention. arXiv preprint arXiv:2510.02912, 2025.

Appendix

The appendix includes the following sections:

- **Section A: Experimental Analysis Supplement.**
- **Section B: Framework Design and Application Extensions.**
- **Section C: Case.**

A Experimental Analysis Supplement

This section supplements the materials on Score Design and OAE Design, including related content and partial experimental data.

Score VS Accuracy In Section 3.1, to comprehensively evaluate the performance of these models, we design two metrics: **Accuracy** for tack-specific correctness and **Score** for dynamic reasoning consistency. The Score metric is crucial for revealing reliable, hierarchical reasoning beyond simple Accuracy. While GPT-4o's Accuracy gain over IXC2.5-OL is moderate (50.02% vs. 47.33%), its substantially higher Score (22.10 vs. 15.40) indicates a more robust reasoning process, reliably building upon foundational understanding to address complex queries. Conversely, lower Scores relative to Accuracy, common among open-source models, suggest instability in multi-step reasoning. Thus, as shown in Figure 3, the Score metric effectively quantifies the deeper, more reliable comprehension capabilities demonstrated by models like GPT-4o on RTV-Bench's challenging tasks and it can distinguish model performance more clearly.

With the aim of validating the discriminative power of the Score metric, we conducted statistical analysis on 1,527 multi-timespan QA-triples, as shown in A.1. The result revealed a significantly positive correlation between foundational reasoning(Q1-Q2) and subsequent conplex reasoning performance(Q3). This demonstrates that models achieving higher accuracy on elementary visual perception tasks exhibit proportionally stronger performance on advanced temporal reasoning tasks.

Such alignment shows our core design philosophy for RTV-Bench: **hierarchical reasoning capabilities** depend on robust foundational perception, mirroring human cognitive processes in real-time video understanding. Furthermore, it also illustrates the rationality of **Score** as a real-time video metric, which addresses critical limitations of conventional single-answer accuracy metrics in assessing continuous reasoning dynamics.

Table A.1: Accuracy Distribution of Question 3 Conditioned on Preceding Question Performance: Q3 Accuracy when at least one of the preceding questions (Q1 or Q2) was answered correctly, versus Q3 accuracy when both preceding questions (Q1 and Q2) were answered incorrectly.

Model	#Size	Q3 | Q1/Q2≥1	Q3 | Q1&Q2=0
GPT-4o [13]	–	297	64
IXC2.5-OL [47]	7B	499	81
LLaVA-OneVision [15]	7B	309	182
LLaVA-Video [22]	7B	308	195

OAE Design Overview During evaluation, we also incorporated the Object-Action-Event framework(Table B.1) as part of our analytical scope, designed to assess video comprehension from multiple agent-centric perspectives. For example, in live sports scenarios, we evaluate three perspectives: objects (e.g., players appearing or disappearing during an offensive play), actions (e.g., dynamic maneuvers by offensive players), and events (e.g., real-time offensive strategies deployed in the midfield).

Table A.2: Detailed evaluation results on the category of **OAE**. Object, Action and Event.

Method	#Size	Object Acc (%) / Score	Action Acc (%) / Score	Event Acc (%) / Score
Open-Source Offline Video Models				
Qwen2.5-VL [43]	7B	39.67 / 8.1	38.12 / 6.7	37.18 / 6.8
VideoLLaMA2 [4]	7B	40.39 / 8.2	40.25 / 8.6	38.69 / 6.8
VideoLLaMA3 [45]	7B	34.31 / 4.5	37.77 / 7.4	34.21 / 3.9
LLaVA-OneVision [15]	7B	34.31 / 5.0	35.82 / 4.9	33.42 / 3.4
LLaVA-Video [22]	7B	36.10 / 6.1	35.40 / 4.7	33.88 / 3.8
Open-Source Online Models				
VITA-1.5 [9]	7B	47.39 / 13.7	44.09 / 11.6	42.85 / 10.3
IXC2.5-OL [47]	7B	49.89 / 17.2	46.34 / 14.6	46.61 / 15.4
Closed-Source Business Models				
GPT-4o [13]	–	**50.63 / 23**	**50.97 / 22.3**	**49.01 / 20.9**
Gemini [35] 2.0 Flash	–	42.66 / 12.2	43.34 / 11.5	40.24 / 12.4

Table A.3: Analytical Object Taxonomy: RTV-Bench systematically formulates an object-action-event framework, with explicit definitions of core characteristics and discriminative criteria for each category.

Category	Dimensions
Spatiotemporal Elements	Objects: Physical entities appearing in video frames. Actions: Dynamic behaviors performed by objects. Events: Complex occurrences combining objects and actions.

OAE Accuracy and Score Analysis In Table B.1, GPT-4o maintains the leading positions in all three dimensions of the OAE , and all online models demonstrate significantly higher accuracy and scores compared to offline models, particularly in the Object dimension, this highlights that the perspective design of OAE requires exceptionally strong capabilities in continuous analysis tasks. Overall, the models show no significant gaps in accuracy and scores across the three dimensions. Most offline models achieve their best performance in the Action aspect, while online models excel particularly in Object. Notably, nearly all models exhibit the lowest accuracy and scores in the Event dimension, indicating that continuous analysis tasks with higher complexity remain a formidable challenge for current systems.

B Framework Design and Application Extensions

This section supplements the materials on the rationale for and necessity of evaluation dimension design, while providing an extended analysis on the broader utility of the dataset.

Methodological Foundations for Assessing Continuous Analysis Capabilities To further elaborate on video continuous analysis capabilities introduced In Section 2.2, we formalize the foundational definitions and evaluation protocols for this capacity. Primarily, models are required to possess perception, understanding, and reasoning abilities comparable to state-of-the-art offline video models before addressing real-time streaming contexts. This requirement motivates our two-stage QA design, as elementary offline video analysis capabilities intuitively form the prerequisite for advanced temporal reasoning.

Furthermore, models must demonstrate proficiency in recognizing correct temporal query patterns due to three critical demands inherent to real-time applications: enhanced perception capabilities in highly dynamic scenarios requiring rapid and precise visual processing, deepened understanding of ongoing events under temporal continuity constraints, and effective reasoning about future trajectories based on evolving contextual cues.

Why evaluate across perception, understanding, and reasoning dimensions? To elucidate the necessity of three-dimensional categorization, we subsequently analyze perception, understanding, and reasoning respectively.In egocentric driving environments with high-dynamic scenarios, the video continuous analysis capability fundamentally addresses two aspects of persistent navigation: 1) holistic operational state awareness (e.g., current traffic condition assessment, historical route context) through the integration of offline and online video processing capabilities, and 2) perception speed, comprehension, and analytical capabilities for sudden real-time events.

Regarding perception design, this corresponds to detecting abrupt environmental changes during navigation, such as traffic light transitions, emergent vehicles, and pedestrians - scenarios where conventional video models exhibit critical deficiencies in temporal responsiveness. As visualized in Figure 4, existing architectures struggle to adapt to real-time variations in high-dynamic settings. We systematically decompose this capability into three sub-dimensions: Temporal Perception, Scene Perception and Visual Perception.

Regarding understanding design, it addresses the interpretative capacity for sudden operational changes during driving. For the understanding design, it pertains to scenarios involving the comprehension of abrupt changes during driving, such as interpreting traffic signal indications, road sign semantics, and the rationale behind preceding vehicles' maneuvering strategies. Previous models have been shown to fall short of fundamental requirements in both processing speed and interpre-

Table B.1: Core Evaluation Dimensions: RTV-Bench systematically defines eight essential evaluation dimensions for continuous video understanding systems, accompanied by formal characterizations and discriminative criteria for each dimension.

Category	Dimensions
Model Capabilities	Temporal Perception (TP): Recognizing temporal sequence and duration. Scene Perception (SP): Understanding holistic environment and layout. Visual Perception (VP): Detecting fine-grained visual features. Future Prediction (FP): Anticipating future developments. Phenomenological Understanding (PU): Interpreting surface phenomena. Intent Analysis (IA): Inferring actor motivations. Global Understanding (GU): Grasping video context. Spatiotemporal Reasoning (SR): Logical deduction from observations.

tative depth. For instance, during traffic signal transitions or lane-changing events, current models struggle to detect such changes within reasonable timeframes. In complex real-time traffic scenarios, beyond insufficient processing speed, existing systems also largely fail to meet advanced comprehension requirements for situational awareness. Through systematic categorization and abstraction of diverse scenarios, we decompose this dimension into three sub-components: Intent Analysis, Phenomenological Understanding and Global Understanding.

Regarding reasoning design, it encompasses scenarios requiring continuous temporal analysis, such as traffic signal duration estimation, and anticipating preceding vehicles' strategies, these demand sophisticated continuous analysis capabilities. Based on the dual requirements of historical context integration and prospective forecasting, we architect this dimension into two sub-dimensions: Spatiotemporal Reasoning and Future Prediction.

Why cross-apply the OAE design with eight evaluation dimensions? We observe notable limitations in the dimensional design frameworks of current mainstream benchmarks, suggesting areas that warrant systematic refinement. Current evaluation taxonomies frequently include components such as Object Recognition and Action Reasoning, yet conspicuously omit complementary dimensions like Object Reasoning and Action Recognition. This prevalent pattern reveals significant arbitrariness in dimension partitioning, resulting in evaluation frameworks whose systematic rigor and scientific credibility remain fundamentally compromised. To address these limitations, we propose a novel taxonomy that first independently categorizes analytical subjects and evaluation dimensions, subsequently implementing cross-categorization to establish a structured and systematic grid framework for dimensional organization.

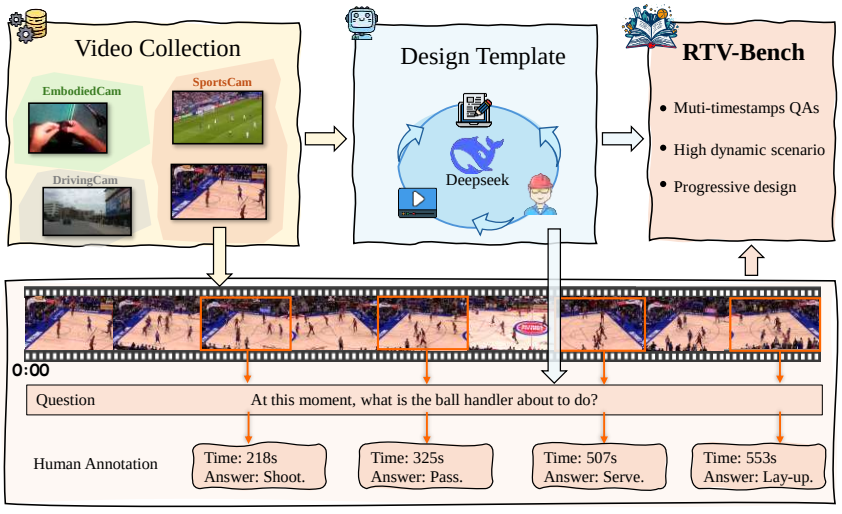


Figure B.1: Our dataset construction pipeline. We develop a dataset generation pipeline consisting of three stages to create RTV-Bench: Video Collection, Templates Design and Human Annotation.

Supplementary Analysis of Dataset Value This work introduces the Real-time Video Understanding Benchmark (RTV-Bench), which holds significant potential to advance video comprehension research and its applications. By defining core capabilities required for real-time video processing through high-dynamic video evaluation, RTV-Bench establishes continuous analysis capabilities as critical developmental objectives. The benchmark demonstrates a novel paradigm for online video capability assessment, potentially accelerating the enhancement of core functionalities in online video understanding models and expediting their deployment in real-time scenarios such as embodied AI, autonomous driving, and live sports commentary.

Experimental analyses using RTV-Bench reveal substantial deficiencies in existing video models' continuous analysis capabilities while illuminating potential improvement pathways. Our results indicate that increasing model scale or video sampling frequency yields diminishing returns. The substantial performance gap between online and offline models in continuous analysis capabilities not only validates RTV-Bench's design efficacy but also suggests architectural innovation as a viable development direction.

The RTV-Bench framework is constructed upon three foundational principles: 1) Balanced Taxonomy Design: Prioritizing structural rationality in evaluation dimensions while implementing novel task formulations. 2) QA Integrity Assurance: Maintaining high-quality question-answer pairs while mitigating validity threats from oversimplified QA designs, particularly accuracy inflation through correct guessing in multiple-choice formats. 3) Capability Anchoring: Conducting specialized evaluations for real-time video processing while systematically monitoring potential degradation in foundational video understanding competencies. Furthermore, the rigorous RTV-Bench generation pipeline B.1 also ensures high-quality QA pair production B.2.

Broader Impact By focusing on real-time video comprehension, RTV-Bench rigorously evaluates models' continuous temporal modeling and dynamic scenario comprehension capabilities. This benchmarking framework directly addresses critical gaps in deploying video understanding systems for latency-sensitive applications, including embodied AI navigation requiring sub-second environmental feedback, autonomous driving systems dependent on frame-by-frame hazard anticipation, and live sports commentary generation demanding real-time event contextualization.

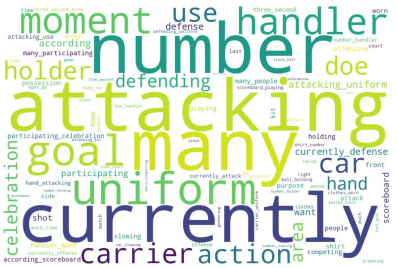
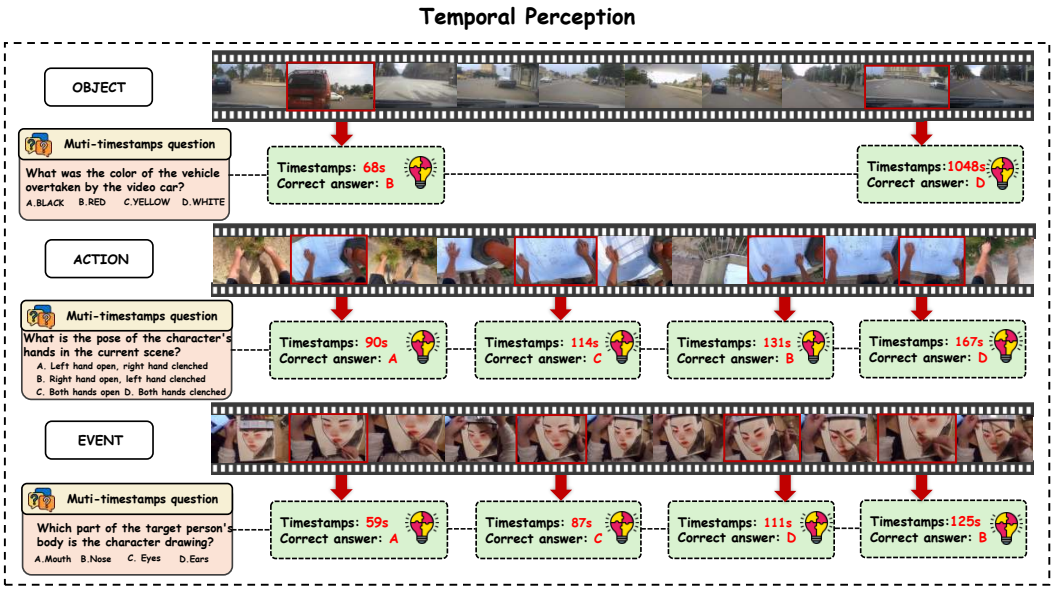


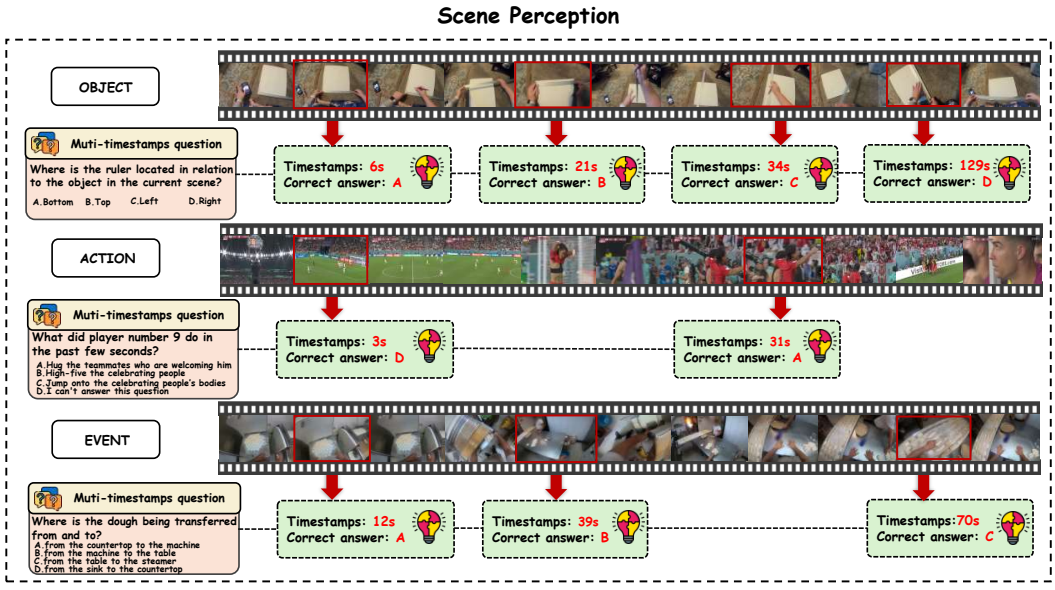
Figure B.2: This figure presents a word cloud visualization derived from QA question corpora, explicitly highlighting the most frequently occurring terms through typographic prominence.

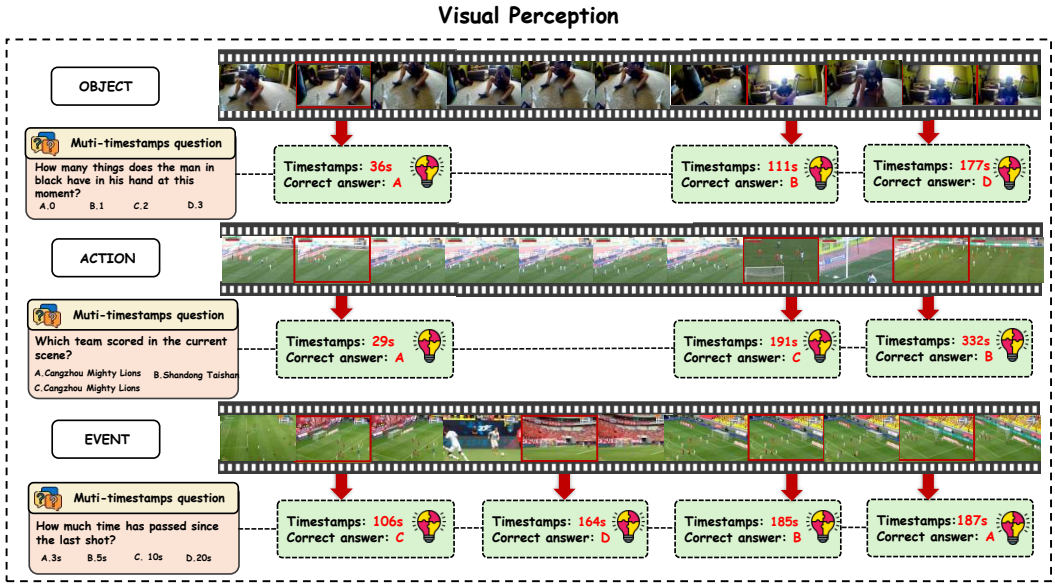
Regarding potential negative societal impact, while RTV-Bench can better facilitate the development and application of real-time video tasks, these technologies may concurrently promote the creation of more sophisticated real-time surveillance systems, which inherently raises privacy concerns. Therefore, the advancement of related technologies must be conducted under the supervision of robust privacy protection measures and ethical standards.

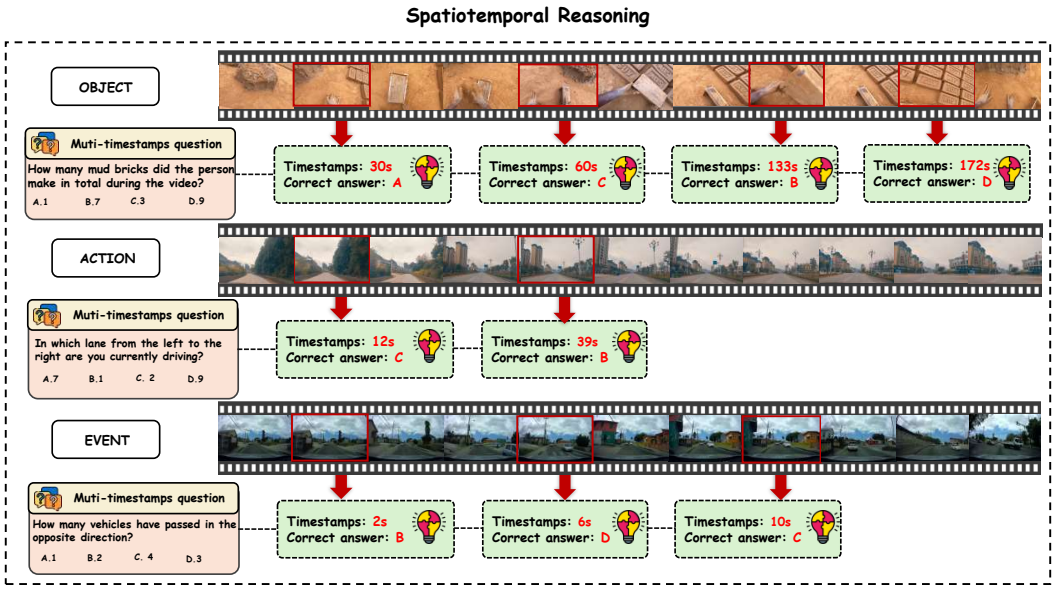
C Case

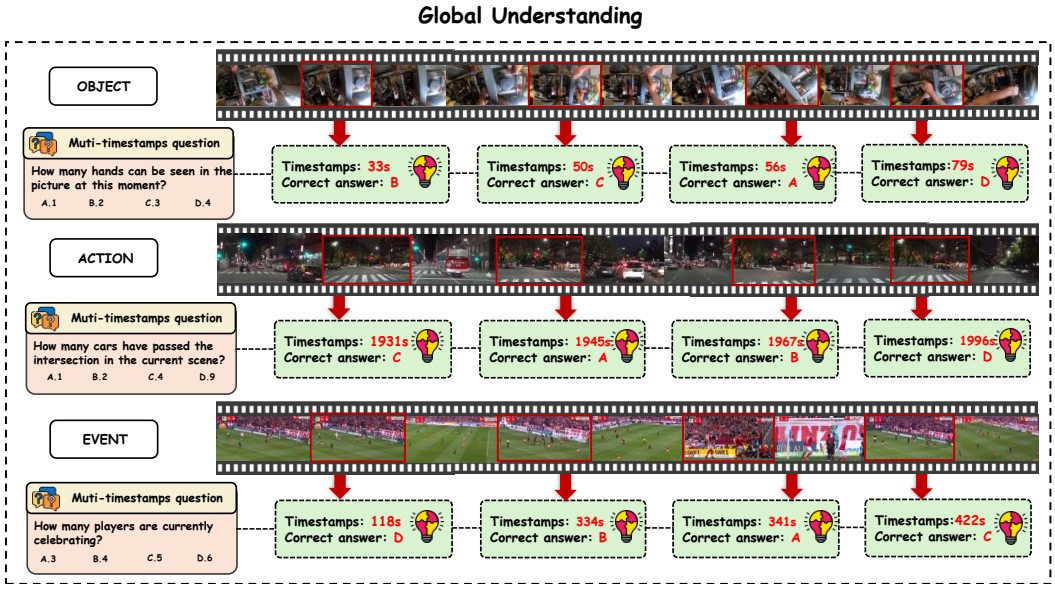
We present additional examples from RTV-Bench, covering eight evaluation dimensions and three analytical perspectives, to demonstrate the holistic and systematic framework for assessing continuous analysis capabilities.

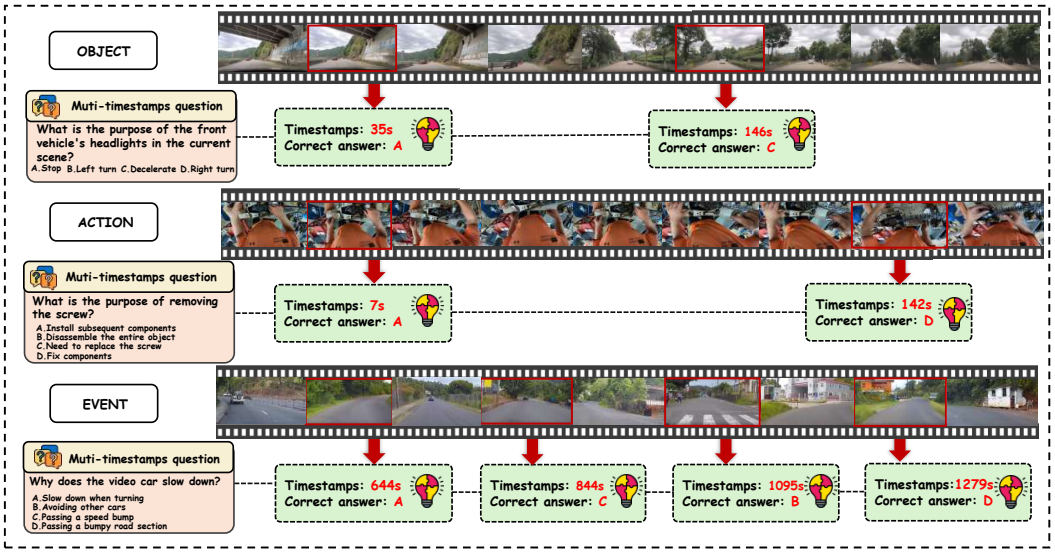

Temporal Perception
OBJECT
Muti-timestamps question
What was the color of the vehicle overtaken by the video car?
A.BLACK B.RED C.YELLOW D.WHITE
Timestamps: 68s
Correct answer: B
Timestamps:1048s
Correct answer: D
ACTION
Muti-timestamps question
What is the pose of the character's hands in the current scene?
A. Left hand open, right hand clenched
B. Right hand open, left hand clenched
C. Both hands open D. Both hands clenched
Timestamps: 90s
Correct answer: A
Timestamps: 114s
Correct answer: C
Timestamps: 131s
Correct answer: B
Timestamps: 167s
Correct answer: D
EVENT
Muti-timestamps question
Which part of the target person's body is the character drawing?
A.Mouth B.Nose C. Eyes D.Ears
Timestamps: 59s
Correct answer: A
Timestamps: 87s
Correct answer: C
Timestamps: 111s
Correct answer: D
Timestamps:125s
Correct answer: B


Scene Perception
OBJECT
Muti-timestamps question
Where is the ruler located in relation to the object in the current scene?
A.Bottom B.Top C.Left D.Right
Timestamps: 6s
Correct answer: A
Timestamps: 21s
Correct answer: B
Timestamps: 34s
Correct answer: C
Timestamps: 129s
Correct answer: D
ACTION
Muti-timestamps question
What did player number 9 do in the past few seconds?
A.Hug the teammates who are welcoming him
B.High-five the celebrating people
C.Jump onto the celebrating people's bodies
D.I can't answer this question
Timestamps: 3s
Correct answer: D
Timestamps: 31s
Correct answer: A
EVENT
Muti-timestamps question
Where is the dough being transferred from and to?
A.from the countertop to the machine
B.from the machine to the table
C.from the table to the steamer
D.from the sink to the countertop
Timestamps: 12s
Correct answer: A
Timestamps: 39s
Correct answer: B
Timestamps:70s
Correct answer: C

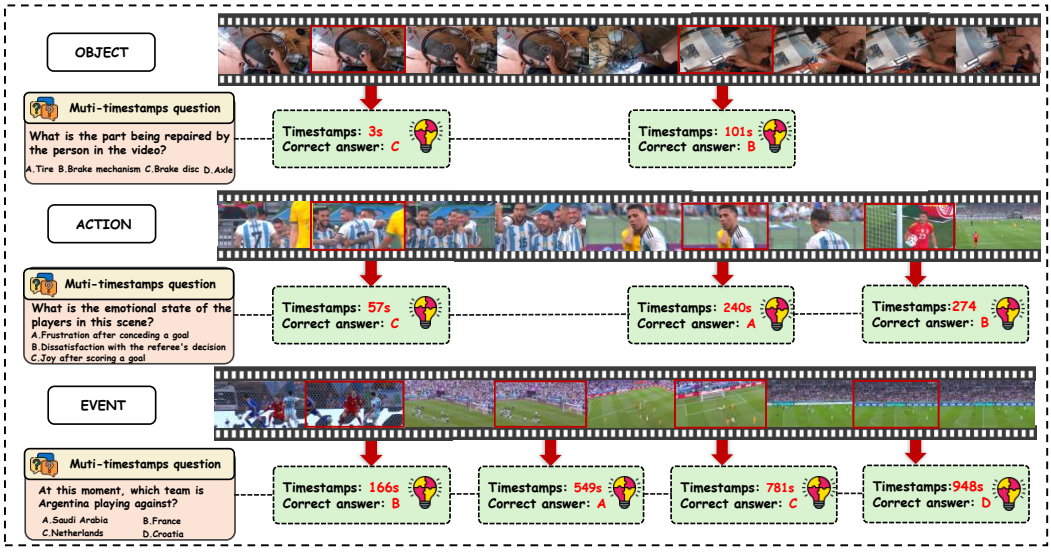

Visual Perception
OBJECT
Muti-timestamps question
How many things does the man in black have in his hand at this moment?
A.0 B.1 C.2 D.3
Timestamps: 36s
Correct answer: A
Timestamps: 111s
Correct answer: B
Timestamps: 177s
Correct answer: D
ACTION
Muti-timestamps question
Which team scored in the current scene?
A.Cangzhou Mighty Lions B.Shandong Taishan
C.Cangzhou Mighty Lions
Timestamps: 29s
Correct answer: A
Timestamps: 191s
Correct answer: C
Timestamps: 332s
Correct answer: B
EVENT
Muti-timestamps question
How much time has passed since the last shot?
A.3s B.5s C. 10s D.20s
Timestamps: 106s
Correct answer: C
Timestamps: 164s
Correct answer: D
Timestamps: 185s
Correct answer: B
Timestamps:187s
Correct answer: A


Spatiotemporal Reasoning
OBJECT
Muti-timestamps question
How many mud bricks did the person make in total during the video?
A.1 B.7 C.3 D.9
Timestamps: 30s
Correct answer: A
Timestamps: 60s
Correct answer: C
Timestamps: 133s
Correct answer: B
Timestamps: 172s
Correct answer: D
ACTION
Muti-timestamps question
In which lane from the left to the right are you currently driving?
A.7 B.1 C. 2 D.9
Timestamps: 12s
Correct answer: C
Timestamps: 39s
Correct answer: B
EVENT
Muti-timestamps question
How many vehicles have passed in the opposite direction?
A.1 B.2 C. 4 D.3
Timestamps: 2s
Correct answer: B
Timestamps: 6s
Correct answer: D
Timestamps: 10s
Correct answer: C

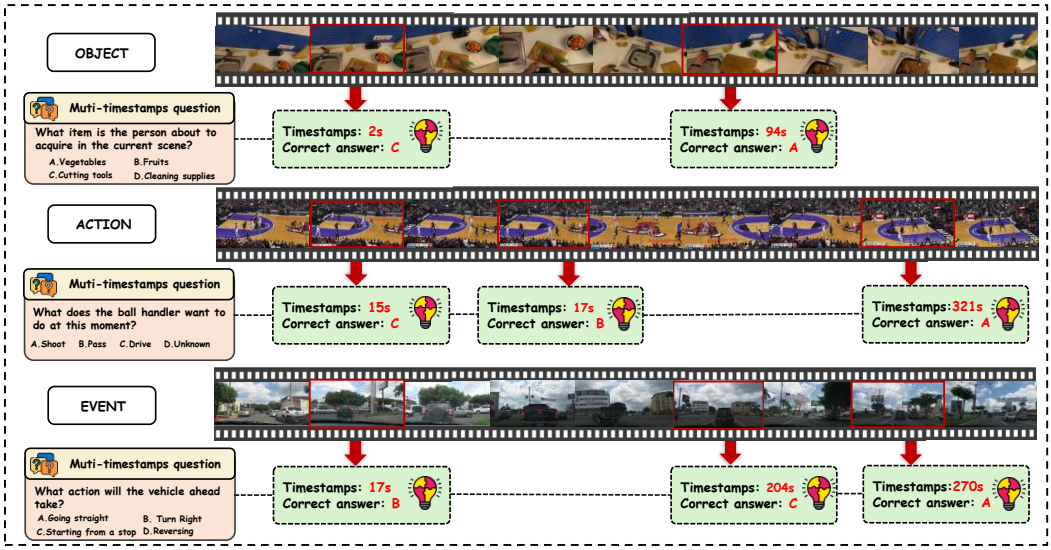

Global Understanding
OBJECT
Muti-timestamps question
How many hands can be seen in the picture at this moment?
A.1 B.2 C.3 D.4
Timestamps: 33s
Correct answer: B
Timestamps: 50s
Correct answer: C
Timestamps: 56s
Correct answer: A
Timestamps:79s
Correct answer: D
ACTION
Muti-timestamps question
How many cars have passed the intersection in the current scene?
A.1 B.2 C.4 D.9
Timestamps: 1931s
Correct answer: C
Timestamps: 1945s
Correct answer: A
Timestamps: 1967s
Correct answer: B
Timestamps: 1996s
Correct answer: D
EVENT
Muti-timestamps question
How many players are currently celebrating?
A.3 B.4 C.5 D.6
Timestamps: 118s
Correct answer: D
Timestamps: 334s
Correct answer: B
Timestamps: 341s
Correct answer: A
Timestamps:422s
Correct answer: C


Intent Analysis
OBJECT
Muti-timestamps question
What is the purpose of the front vehicle's headlights in the current scene?
A.Stop B.Left turn C.Decelerate D.Right turn
Timestamps: 35s
Correct answer: A
Timestamps: 146s
Correct answer: C
ACTION
Muti-timestamps question
What is the purpose of removing the screw?
A.Install subsequent components
B.Disassemble the entire object
C.Need to replace the screw
D.Fix components
Timestamps: 7s
Correct answer: A
Timestamps: 142s
Correct answer: D
EVENT
Muti-timestamps question
Why does the video car slow down?
A.Slow down when turning
B.Avoiding other cars
C.Passing a speed bump
D.Passing a bumpy road section
Timestamps: 644s
Correct answer: A
Timestamps: 844s
Correct answer: C
Timestamps: 1095s
Correct answer: B
Timestamps:1279s
Correct answer: D

Phenomenological Understanding



Future Prediction



NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly reflect the paper's core contributions and research scope. Specifically, they explicitly present the research motivation for RTV-Bench, identify the limitations of existing paradigms, and introduce continuous analysis capabilities along with the associated evaluation protocol.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The limitations of the work are detailed in Section 5

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: As a dataset work, this paper does not contain theoretical derivations.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental section of this paper includes dataset evaluation benchmarks, with the dataset open-sourced on public hosting platforms and accompanied by a Croissant metadata file submission.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: RTV-Bench has been open-sourced on public hosting platforms and submitted with a Croissant metadata file.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental configurations for evaluation are comprehensively detailed in Section 3.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not report error bars or statistical signiffcance tests in this paper due to the extensive computational cost associated with our experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The experimental configurations for evaluation are comprehensively detailed in Section 3.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in this paper fully conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The broader impacts of this work are detailed in Section B

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This work does not contain high-risk content.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All referenced papers are fully cited with complete bibliographic entries in the References section.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [No]

Justification: The documentation has not yet been finalized and made publicly available in public repositories as of the current stage of development.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve research with human subjects or crowdsourcing.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [No]

    Justification: This work is a dataset project where LLMs are solely used as evaluation targets and utilized to inform the dataset construction pipeline.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.