CROSSTIMEBENCH: BENCHMARKING CROSS-TEMPORAL VIDEO REASONING IN COMPLEX MULTI-EVENT SCENARIOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in video reasoning have increasingly focused on improving the temporal understanding capabilities of Multimodal Large Language Models(MLLMs). However, existing benchmarks mainly focus on short clips and single-segment reasoning. While such benchmarks are useful for assessing basic temporal understanding, they are insufficient for evaluating complex multi-event scenarios where relevant evidence is distributed across different parts of a video. In this paper, we introduce the CrossTimeBench, **a cross-temporal video reasoning benchmark designed to systematically evaluate model reasoning over complex multi-event videos**. Each question in the benchmark requires the integration of information from more than two temporal segments, constructed through both mining and synthesis from existing datasets. The benchmark covers 8 major categories and 35 subcategories, including 2,607 videos and 8,496 QA pairs (with 6,628 unique samples due to multi-label annotations). Among them, 48.4% of open-ended QA pairs and 94.6% of closed-form QA pairs are newly constructed. In addition, we propose a fine-grained evaluation framework based on the benchmark and conduct extensive baseline experiments on a wide range of mainstream open-source and closed-source MLLMs. Our experimental results show that, although current models achieve strong performance on existing benchmarks, they still exhibit substantial limitations in cross-temporal reasoning tasks. Meanwhile, the evaluation results demonstrate noticeable differences across open-ended and closed-form QA settings, as well as among different question types. These observations provide new insights into the limitations of current video understanding models.

1 INTRODUCTION

With the growing scale of video data and the increasing complexity of real-world applications, temporal understanding has become a core challenge in video understanding. Semantic information in videos may be distributed across multiple temporal segments, where events exhibit complex temporal orders, causal relationships, and state transitions, placing greater demands on models' ability to capture long-range dependencies. Models must not only accurately localize key temporal segments but also effectively integrate and reason over information across time.

In recent years, the rapid development of video models has substantially improved temporal understanding and driven the development of benchmarks for systematic evaluation of temporal reasoning. However, existing benchmarks often focus on short-term perception, making them insufficient for evaluating long-range temporal reasoning in complex, multi-event long videos.

Early benchmarks such as MSVD-QA and MSRVTT-QA Xu et al. (2017) primarily assess factual understanding over short clips, with many questions solvable from local visual cues. TVQA Lei et al. (2018) and NExT-QA Xiao et al. (2021) introduce temporal and causal reasoning, while STAR Wu et al. (2021) further considers structured sequence, interaction, and predictive reasoning. Ego4D Grauman et al. (2022) extends evaluation to long-form egocentric videos through tasks such as episodic memory and action anticipation. Nevertheless, these benchmarks generally evaluate

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

localized temporal relations or task-specific capabilities rather than systematically requiring models to integrate evidence across multiple temporally separated events in long-horizon scenarios.

This gap highlights the need for new evaluation benchmarks to assess models' temporal understanding in complex multi-event scenarios. To this end, we propose a cross-temporal video reasoning benchmark for systematically evaluating models' capabilities in complex temporal modeling and reasoning. The benchmark explicitly requires reasoning over temporally disjoint segments, rather than relying solely on continuous temporal modeling. In contrast to conventional approaches that focus on local dependencies, cross-temporal reasoning requires models to establish long-range dependencies across events, providing a more realistic and challenging setting for video understanding.

Our contributions are summarized as follows:

A fine-grained cross-temporal benchmark. We construct CrossTimeBench, a cross-temporal video understanding benchmark in which all questions require integrating information from at least two temporal segments, enabling the evaluation of reasoning ability in complex multi-event scenarios. The benchmark includes 8 major categories and 35 subcategories, covering 2,607 videos and 8,496 QA pairs (with 6,628 unique samples after deduplication due to multi-label annotations). Among them, 48.4% of open-ended QA pairs and 94.6% of closed-form QA pairs are newly constructed.

A structured evaluation framework. We propose an evaluation framework based on both scene types and question types. The scene dimension covers 35 cognitive task categories defined in the benchmark taxonomy. In the evaluation protocol, we separately assess open-ended and closed-form QA tasks, enabling a more complementary and comprehensive analysis of model capabilities.

A systematic evaluation of state-of-the-art models. We conduct extensive evaluations on a variety of state-of-the-art multimodal video models. The results show that although these models perform well on existing video understanding benchmarks, their performance drops significantly on cross-temporal reasoning tasks, revealing key limitations in long-range temporal modeling

2 Related Work

2.1 VideoQA

Video Question Answering (VideoQA) aims to perform visual understanding and reasoning by jointly leveraging video content and natural language questions. Early benchmarks such as TGIF-QA Jang et al. (2017) and MSR-VTT-QA Xu et al. (2016) focused on factual understanding and local temporal relations, while TVQA Lei et al. (2018) and NExT-QA Xiao et al. (2021) introduced more structured temporal, causal, and event-centric reasoning. With the emergence of multimodal large language models (MLLMs), works such as VideoChat Li et al. (2023b) and Video-LLaMA Zhang et al. (2023), building upon vision-language models such as Flamingo Alayrac et al. (2022) and BLIP-2 Li et al. (2023a), have further advanced VideoQA toward open-ended reasoning.

2.2 Temporal Modeling and Long-Video Understanding

The increasing duration and complexity of videos have motivated research on long-range temporal modeling through long-video memory, temporal-aware representations, and temporal localization. MovieChat+ Song et al. (2026) employs question-aware sparse memory to efficiently retain relevant visual information, while TimeChat Ren et al. (2024) incorporates timestamps into frame representations and uses a sliding Video Q-Former for long-video understanding. VTimeLLM Huang et al. (2024) further introduces boundary-aware training for fine-grained temporal understanding. Closely related is temporal grounding, which aims to localize video segments relevant to a language query. Methods such as TALL Gao et al. (2017a) and TURN-TAP Gao et al. (2017b) formulate this as temporal boundary prediction, while Trust Xiao et al. (2024) and GroundVQA Di & Xie (2024) connect temporal grounding with VideoQA. These works emphasize identifying relevant temporal evidence, but do not necessarily evaluate whether models can *jointly integrate multiple temporally separated evidence segments*, which is the focus of our work.

2.3 VIDEOQA BENCHMARKS AND LONG-RANGE VIDEO REASONING

VideoQA benchmarks have evolved from factual understanding toward temporal, causal, and compositional reasoning. MSR-VTT-QA Xu et al. (2016) and ActivityNet-QA Yu et al. (2019) primarily evaluate factual understanding, while TVQA Lei et al. (2018), NExT-QA Xiao et al. (2021), and AGQA Jang et al. (2017) introduce temporal, causal, and compositional spatio-temporal reasoning. Long-video benchmarks such as EgoSchema Mangalam et al. (2023), Video-MME Fu et al. (2025), and MLVU Zhou et al. (2025) further evaluate long-range understanding across diverse capabilities, but cross-temporal evidence integration is not a universal requirement across their questions.

LongVideoBench Wu et al. (2024) and KFS-Bench Li et al. (2026) are closely related to our work. LongVideoBench explores multi-moment relational reasoning through tasks such as event ordering, object tracking, and sequence-of-scenes understanding. However, multi-moment reasoning constitutes only one aspect of its overall evaluation. In contrast, CrossTimeBench makes **cross-temporal evidence integration** a unified task constraint, requiring every question to depend on multiple temporally separated evidence segments. KFS-Bench evaluates evidence coverage in long-video QA from the perspective of key-frame sampling, considering both dispersed and continuous evidence. In contrast, CrossTimeBench imposes the presence of temporally separated evidence at the question construction level and evaluates models' ability to **jointly integrate and reason over** such evidence.

3 BENCHMARK

3.1 TASK DEFINITION

We formalize CrossTimeBench as a reasoning problem grounded in temporally distributed visual evidence within a video. Given an untrimmed video

$$V = \{f_1, f_2, \ldots, f_T\},$$

where f_t denotes the video frame at time step t, the video consists of multiple semantically meaningful events that may occur across different and potentially distant temporal intervals.

Unlike traditional Video Question Answering (VideoQA) settings, which typically assume that the answer can be obtained from a single continuous segment or a short temporal window, CrossTimeBench requires models to perform reasoning over multiple non-contiguous temporal segments. Specifically, for a question Q, the model needs to identify and integrate relevant evidence from a set of temporally separated intervals

$$\{S_1, S_2, \ldots, S_n\},$$

where each segment is defined as

$$S_i = (t_i^{\text{start}}, t_i^{\text{end}}) \subseteq [1, T].$$

The objective of the task is to generate an answer conditioned on both the question and the temporally distributed evidence:

$$A = \mathcal{F}(V, Q),$$

where $\mathcal{F}(\cdot)$ denotes a multimodal reasoning function that requires the following capabilities: (i) temporal localization, (ii) cross-segment alignment, and (iii) compositional reasoning across events.

3.2 LABEL SYSTEM

We further design a fine-grained task and labeling framework, enabling evaluation at multiple levels of granularity, from overall capabilities to specific sub-tasks, all of which can be quantitatively analyzed.

Hierarchical labeling structure. We introduce a hierarchical labeling structure, which contains primary labels and secondary labels. The primary labels define coarse-grained task categories, including *action_activity* (AA), *temporal_reasoning* (TR), *spatial_reasoning* (SR), *object_understanding* (OU), *human_understanding* (HU), *causality_reasoning* (CR), *scene_environment* (SE), and *description_summarization* (DS).

Table 1: Fine-grained secondary labels with question prompts.

Primary_label.Secondary_label	Abbr.	Example(question)
action_activity.simple_action	AA.SA	What does the man do when he reaches the door?
action_activity.complex_activity	AA.CA	What activity is the woman performing?
action_activity.object_interaction	AA.OI	How does the child interact with the toy?
temporal_reasoning.before_after	TR.BA	What does the boy do after dropping the ball?
temporal_reasoning.event_progression	TR.EP	How does the thematic focus shift?
temporal_reasoning.state_transition	TR.ST	How does the door change?
temporal_reasoning.future_prediction	TR.FP	What will the cyclist do next?
spatial_reasoning.absolute_location	SR.AL	Where is the suitcase?
spatial_reasoning.relative_location	SR.RL	Where is the dog relative to the man?
spatial_reasoning.orientation_direction	SR.OD	Which direction is he facing?
spatial_reasoning.path_trajectory	SR.PT	How does the car move?
spatial_reasoning.distance_judgement	SR.DJ	Who is closer to the camera?
object_understanding.object_identification	OU.OI	What is the woman holding?
object_understanding.object_attribute	OU.OA	What color is the car?
object_understanding.object_state	OU.OS	Is the window open?
object_understanding.object_count	OU.OC	How many apples are there?
object_understanding.object_existence	OU.OE	Is there a bicycle?
object_understanding.object_movement	OU.OM	Does the box move?
human_understanding.person_identification	HU.PI	Which person wears a hat?
human_understanding.appearance_attribute	HU.AA	What is she wearing?
human_understanding.emotion	HU.E	How does the child feel?
human_understanding.gaze_attention	HU.GA	What is he looking at?
human_understanding.social_interaction	HU.SI	How are they interacting?
causality_reasoning.cause	CR.C	Why does the glass fall?
causality_reasoning.effect	CR.E	What happens after pressing the button?
causality_reasoning.intention	CR.I	Why does he wear a helmet?
causality_reasoning.multi_step_reasoning	CR.MSR	How does the outage occur?
scene_environment.scene_classification	SE.SC	What is the place?
scene_environment.environment_condition	SE.EC	What is the weather like?
scene_environment.scene_layout	SE.SL	How is the room arranged?
scene_environment.background_activity	SE.BA	What are people in the background doing?
description_summarization.short_description	DS.SD	What happens in the video?
description_summarization.long_summarization	DS.LS	Describe the video.
description_summarization.video_topic	DS.VT	What is the video mainly about?

The secondary labels further characterize fine-grained sub-tasks within each primary category. The detailed configuration of the secondary labels can be found in Table 1.

Multi-label mechanism. Each question is assigned multiple labels to capture its underlying compositional reasoning properties. For example, a single question may simultaneously involve temporal order reasoning and causal inference. This multi-label mechanism helps more accurately reflect task complexity and supports cross-dimensional performance analysis.

3.3 Dataset Construction

Our annotation pipeline constructs QA pairs requiring reasoning across multiple temporal segments through four stages: (1) large-scale mining from existing datasets; (2) data synthesis to complement missing QA types; (3) LLM-based construction of a comprehensive label taxonomy and open-ended and closed-form QA pairs; and (4) dual validation with LLMs and human verification.

3.3.1 Data Mining Pipeline

A comprehensive survey of existing datasets is conducted. Among numerous candidates, we prioritize videos with explicit or implicit temporal structures, such as those containing event transitions or annotated temporal boundaries, including VRBench, Next-GQA, TVQA+, EgoTimeQA, TimeIT, TACoS, and STAR. To enhance data diversity, we also mine datasets without explicit temporal an-

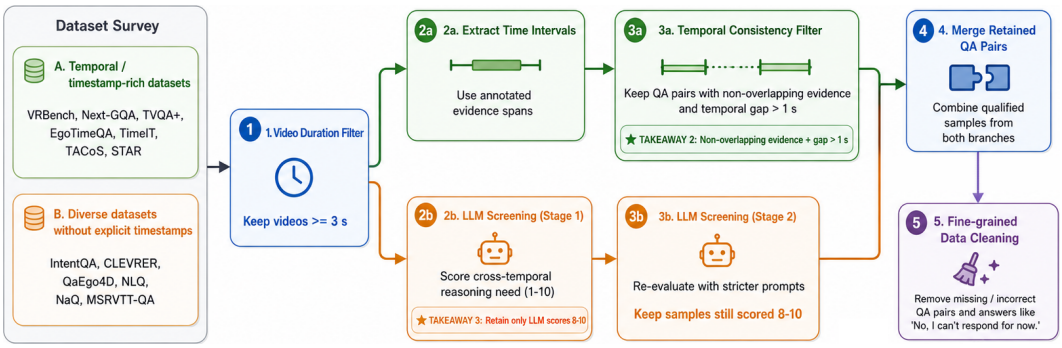


Figure 1: Data mining pipeline

notations, including IntentQA, CLEVRER, QA-Ego4D, NLQ, NExT-QA, and MSRVTT-QA. In addition, we filter out overly short video clips and retain only those with a duration of no less than 3 seconds, in order to support modeling scenarios with meaningful temporal separation between events. This design ensures that the constructed video collection maintains rich temporal information while preserving high annotation reliability, thereby laying a solid foundation for subsequent cross-temporal reasoning tasks.

For datasets that already provide QA annotations, we first perform temporal consistency filtering. When explicit timestamps are available, we extract the corresponding temporal intervals and compute their temporal gaps. We retain only those QA pairs whose supporting evidence is distributed across non-overlapping temporal segments with a gap greater than 1 second. This strategy ensures that the selected samples require reasoning beyond local or short contiguous segments.

For datasets without explicit timestamps, we adopt a two-stage filtering strategy based on LLMs. In the first stage, we prompt the LLM to assess whether a given QA pair requires cross-temporal reasoning and assign a score from 1 to 10, retaining only those with scores above 8. In the second stage, we further analyze the results from the first stage, refine the prompts with stricter and more specific criteria, and re-evaluate the remaining samples, keeping only those that consistently exceed the threshold. This process improves the reliability of the temporal reasoning signal.

Subsequently, we perform more fine-grained data cleaning on the filtered results. For example, we remove low-quality samples with missing or incorrect QA content, as well as responses such as "No, I can't respond for now." Finally, we obtain 34,386 valid QA pairs from an initial pool of 2,095,968 samples across the above datasets.

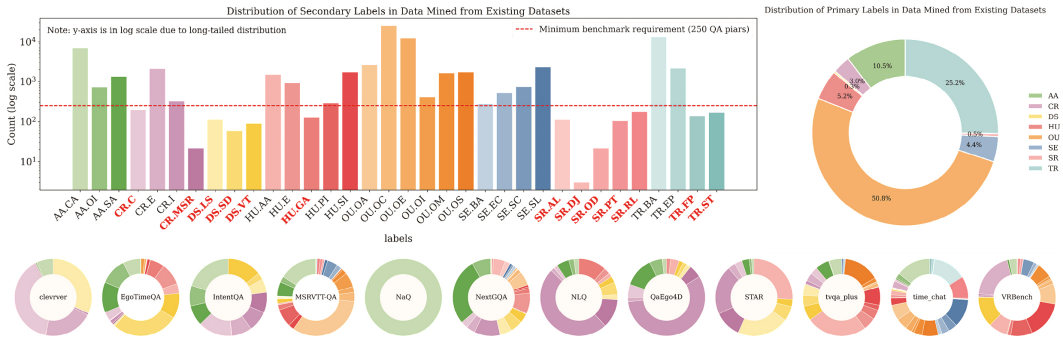


Figure 2: Distribution of Primary Labels in Data Mined from Existing Datasets

Our ultimate goal is to construct a scientific and well-balanced evaluation framework. To this end, it is necessary to understand the distribution of existing QA pair types. Since different datasets adopt inconsistent classification standards for QA types, we establish a unified question type mapping as the label system defined. We then leverage LLMs to classify the mined QA pairs according to this mapping and record the results in the QA database.

3.3.2 Data Synthesis Pipeline

Through an analysis of the label distribution in the mined data, we observe that several label types remain underrepresented, including CR.C, CR.MSR, DS.LS, DS.SD, DS.VT, HU.GA, SR.AL, SR.DJ, SR.OD, SR.PT, SR.RL, TR.FP, and TR.ST (a total of 13 categories). To address this issue, we leverage LLMs for data synthesis based on the ActivityNet Krishna et al. (2017) dataset.

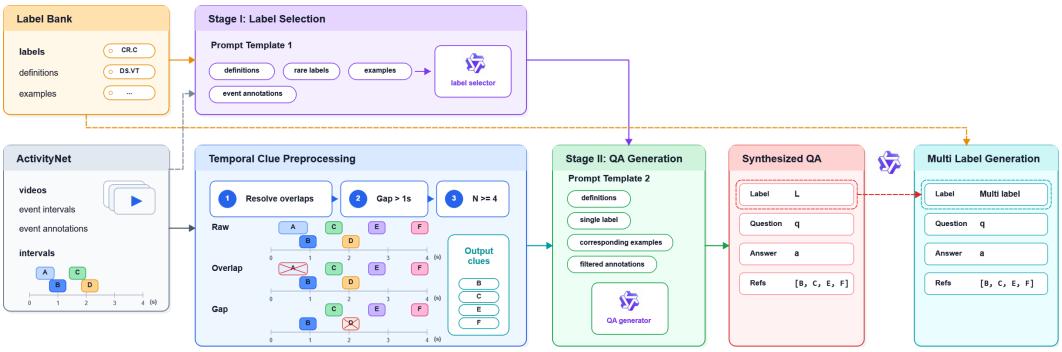


Figure 3: Data synthesis pipeline

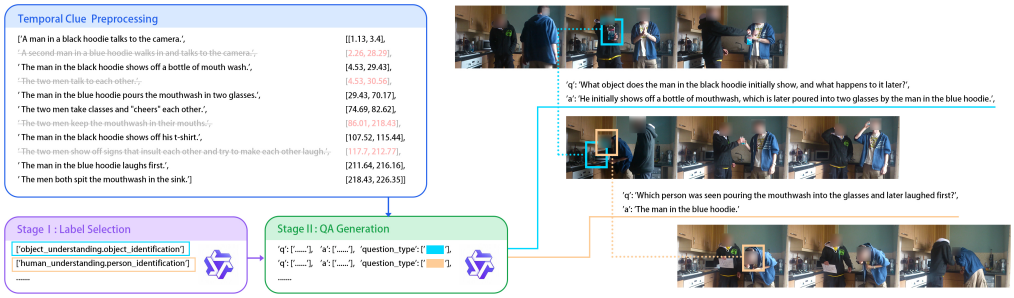


Figure 4: An Example of the Data Synthesis Pipeline. The original video frames are sourced from the ActivityNet video v_MyMYuXh8E0Y.

Our overall pipeline can be summarized in the following steps:

Data preprocessing. We observe that candidate event intervals in the original data may be either continuous or non-contiguous, and may also overlap. To ensure that the synthesized data strictly involves cross-temporal segments, we require the input evidence provided to the LLM to be non-contiguous and non-overlapping. Therefore, we filter out overlapping intervals within each sample, prioritizing the retention of shorter intervals to improve annotation granularity.

Subsequently, we further refine the dataset by (1) enforcing that adjacent intervals are separated by more than 1 second to better satisfy the cross-temporal segmentation criterion, and (2) removing samples containing fewer than four intervals, as we empirically observe that generation quality degrades when the number of intervals is insufficient.

Stage 1: Synthesizing applicable labels generation. Considering that not every video is suitable for QA generation across all rare labels, forcing the model to generate question-answer pairs for a specific label is both unnecessary and likely to degrade generation quality. Therefore, we adopt a two-stage generation strategy when using LLMs. In the first stage, we prompt the LLM to select appropriate labels for each video. The input prompt includes the full set of temporal intervals and their annotations from the original data, as well as the list of target labels and their definitions. In addition, for each label, we incorporate representative QA pairs mined from existing data as few-shot examples to ensure consistency in both label interpretation and QA style.

Stage 2: QA pairs generation. Based on the labels selected in the first stage, we provide the LLM with the processed temporal intervals and event annotations, along with a single target label, its definition, and corresponding few-shot examples, to generate preliminary question-answer pairs together with their supporting temporal intervals. During the benchmark construction stage, we further perform additional verification and quality control on the synthesized labels to ensure data reliability. In total, we synthesize 10,675 QA pairs from 974 videos.

Multi-label generation. Since the scene labels obtained in the previous step are in a one-to-one correspondence with the question-answer pairs, we further leverage the label taxonomy and employ LLMs to perform multi-label annotation for the synthesized QA pairs.

3.3.3 Benchmark Selection

Due to the imbalanced data distribution across labels, we uniformly sample 250 instances per label to construct a balanced benchmark. In total, we collect 8,496 open-ended QA pairs covering 35 labels, including 4,109 synthesized and 4,387 mined samples. For the selected samples, we further perform information alignment and supplement closed-form QA pairs for 8,041 samples. We also analyze the number and maximum span of reference temporal ground-truth intervals in the synthesized data, as shown in Figure 6.

In addition, we release the full set of mined and synthesized data as an expandable data pool in an open-source manner. Researchers can flexibly incorporate the remaining data not included in the benchmark into their evaluation pipelines according to specific needs.

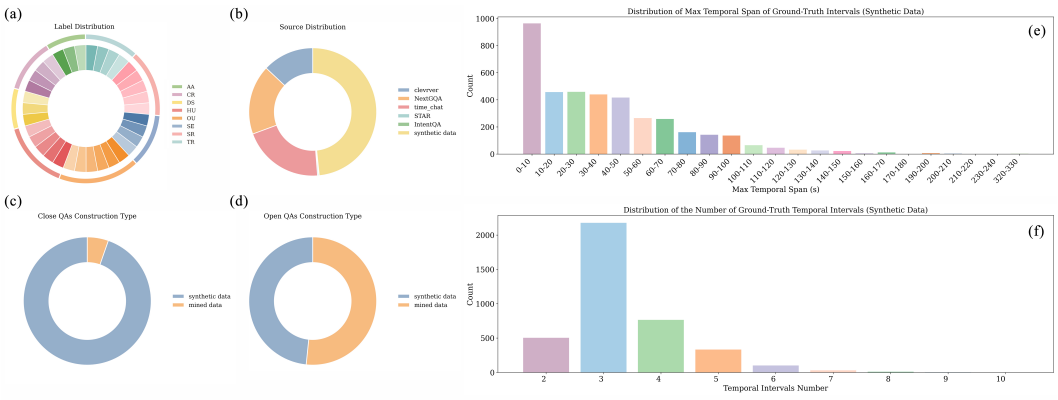


Figure 5: (a) The distribution of labels in the benchmark is balanced. (b) The distribution of QA pair sources in the benchmark. (c) The composition of close QA pairs in the benchmark: 94.6% synthetic and 5.4% mined. (d) The composition of open QA pairs in the benchmark: 48.4% synthetic and 51.6% mined. (e) Distribution of maximum temporal span of ground-truth intervals (synthetic data). (f) Distribution of the number of ground-truth temporal intervals (synthetic data).

3.3.4 Quality Control

To further ensure annotation quality, we perform additional consistency verification from multiple perspectives. Before constructing the benchmark, we re-evaluate the synthesized data using the LLM-based scoring method. In the prompt, we provide the complete temporal intervals and annotations from the original data, and assess each QA pair based on both content accuracy and its cross-temporal characteristics. This process conducts consistency verification through forward and backward reasoning, and only samples that pass the evaluation are retained. In addition, we apply rule-based verification and filtering to the generated temporal intervals to ensure that they satisfy the requirements of cross-temporal segments. All synthesized samples included in the final benchmark pass both verification procedures. Furthermore, after constructing the benchmark, we perform manual spot-checking on the benchmark data, achieving an acceptance rate of approximately 90%.

3.4 Evaluation System

Based on the constructed controlled benchmark, we design a multi-dimensional evaluation framework to systematically assess models' cross-temporal reasoning capabilities from multiple perspectives.

Open-ended QA evaluation. In open-ended QA evaluation, we adopt F1 and CIDEr as the primary automatic metrics. F1 measures the lexical overlap between predicted and ground-truth answers, reflecting the balance of precision and recall. CIDEr evaluates generation quality from the perspective of semantic similarity. These two metrics complement each other by capturing both surface-level matching and semantic consistency, providing a stable and comparable evaluation protocol.

Closed-form QA evaluation. We synthesize closed-form QA pairs for the benchmark data, enabling a closed-loop evaluation pipeline. Each sample is associated with a complete and well-defined answer space, allowing for exact matching and automated evaluation. This reduces the uncertainty inherent in open-ended evaluation and improves the stability and reproducibility of the results.

Label-aware evaluation. The hierarchical labeling system enables evaluation across different task dimensions. We analyze model performance by primary and secondary labels to compare different reasoning types, such as temporal order, causal, and multi-step reasoning. The multi-label annotation further allows each sample to contribute to multiple evaluation dimensions, providing a more comprehensive and interpretable analysis of model capabilities.

4 Experiments

4.1 Experiment Setup

Based on the proposed benchmark, we evaluate a range of open-source and proprietary multimodal large language models. Specifically, we select six open-source models, including Qwen2.5-VL-7B-Instruct Bai et al. (2025), Qwen3-VL-4B-Instruct, Qwen3-VL-8B-Instruct, Qwen3-VL-30B-A3B-Instruct Team (2025), InternVL3-8B Chen et al. (2024), and LLaVA-NeXT-Video-7B-hf Zhang et al. (2024), as well as a proprietary model Claude-Sonnet-4-6 Anthropic (2026). For open-source models, we adopt a video preprocessing configuration of 1 FPS sampling, with a minimum of 8 frames and a maximum of 64 frames, and a spatial resolution of 360P. For proprietary models, due to token length constraints, we limit the maximum number of input frames to 32.

4.2 Baseline

We adopt F1 and CIDEr as the primary metrics for open-ended QA evaluation, and further assess model performance under constrained answer spaces through closed-form QA tasks.

For closed-form multiple-choice question answering tasks, Qwen3-VL-30B-A3B-Instruct achieves the best performance in 6 out of 8 scenarios. Notably, despite the substantial difference in parameter size between Qwen3-VL-8B-Instruct and Qwen3-VL-30B-A3B-Instruct, their average performance differs by only 0.08. In addition, the relatively poorer performance of closed-source models in our experiments may be attributed to the reduced number of input frames; further ablation studies on frame count are needed to verify this effect. For open-ended question answering tasks, the trends observed in F1 and CIDEr metrics are similar. Qwen3-VL-30B-A3B-Instruct still achieves the best performance, followed by Qwen2.5-VL-7B-Instruct. The baselines of each model can be found in Table 2, Table 3, and appendix D.

We analyze the correlations between answer accuracy and four temporal metrics on the synthetic dataset for the top three models in terms of closed-form QA performance: the number of temporal intervals, average interval span, total temporal coverage, and maximum temporal gap. We find that the number of temporal intervals does not exhibit a simple monotonic relationship with accuracy. In contrast, both average interval span and total temporal coverage are generally positively correlated with accuracy, suggesting that questions with broader temporal coverage are not necessarily more difficult. Meanwhile, the result of max gap shows a negative trend for some models, indicating that larger temporal gaps may increase the difficulty of integrating cross-temporal evidence. These results suggest that the difficulty of CrossTimeBench may not simply arise from temporal span

Table 2: Performance comparison across primary label on the closed-form QA task.

Model	AA	TR	SR	OU	HU	CR	SE	DS	Avg
Qwen2.5-VL-7B-Instruct	0.42	0.48	0.46	0.49	0.51	0.51	0.48	0.54	0.49
Qwen3-VL-4B-Instruct	0.54	0.35	0.23	0.52	0.49	0.41	0.58	0.45	0.45
Qwen3-VL-8B-Instruct	0.52	0.67	0.62	**0.66**	0.73	0.75	0.70	0.70	0.67
Qwen3-VL-30B-A3B-Instruct	**0.56**	**0.76**	**0.69**	0.65	**0.78**	**0.86**	**0.81**	**0.91**	**0.75**
InternVL3-8B	0.40	0.47	0.42	0.44	0.46	0.49	0.48	0.49	0.46
LLaVA-NeXT-Video-7B-hf	0.31	0.41	0.42	0.35	0.40	0.42	0.42	0.42	0.39
Claude-Sonnet-4-6	0.06	0.09	0.13	0.06	0.09	0.07	0.11	0.17	0.09

Table 3: Performance comparison across primary label using F1 and CIDEr metrics.

	F1									CIDEr								
Model	AA	TR	SR	OU	HU	CR	SE	DS	Avg	AA	TR	SR	OU	HU	CR	SE	DS	Avg
Qwen2.5-VL-7B-Instruct	**0.23**	0.36	**0.47**	0.42	0.25	**0.26**	0.26	0.34	**0.32**	**0.55**	0.59	**1.39**	0.96	0.40	**0.32**	0.18	0.23	0.58
Qwen3-VL-4B-Instruct	0.13	0.30	0.39	0.30	0.26	0.25	0.26	0.32	0.28	0.16	0.41	0.97	0.53	0.40	0.26	0.18	0.21	0.39
Qwen3-VL-8B-Instruct	0.19	0.32	0.30	0.36	**0.28**	**0.26**	0.25	0.30	0.28	0.30	0.53	0.65	0.45	0.41	0.21	0.18	0.15	0.36
Qwen3-VL-30B-A3B-Instruct	0.21	**0.38**	0.39	**0.45**	**0.28**	**0.26**	0.27	**0.36**	**0.32**	0.46	**0.71**	0.94	**1.23**	**0.52**	0.30	**0.27**	**0.25**	**0.59**
InternVL3-8B	0.17	0.32	0.35	0.37	0.23	0.22	**0.28**	0.32	0.28	0.28	0.49	0.75	0.78	0.29	0.17	0.23	0.15	0.39
LLaVA-NeXT-Video-7B-hf	0.09	0.19	0.36	0.17	0.22	0.24	0.22	0.35	0.23	0.07	0.19	0.85	0.27	0.26	0.19	0.10	0.20	0.27
Claude-Sonnet-4-6	0.12	0.24	0.43	0.21	0.20	0.21	0.21	0.31	0.24	0.04	0.11	0.63	0.18	0.12	0.11	0.05	0.12	0.17

or the number of evidence segments, but rather from the need to bridge temporal gaps, establish relationships across separated segments, and integrate the corresponding evidence.

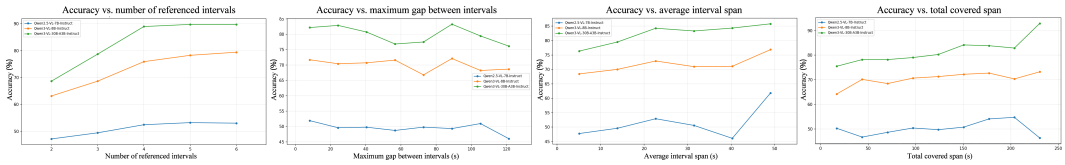


Figure 6: Relationship Between QA Accuracy and Temporal Factors: Number of Temporal Intervals, Maximum Temporal Gap, Average Interval Span, and Total Temporal Coverage.

Furthermore, we visualize the attention of Qwen2.5-VL-7B on both correctly and incorrectly answered samples, using the ground-truth temporal evidence intervals as references. The results show that the model's attention is not completely misaligned with the ground-truth evidence intervals, but a noticeable gap remains. Detailed case analyses are provided in Appendix D, with examples shown in Fig. 7 and Fig. 8.

5 Conclusion

We introduce CrossTimeBench, a video understanding benchmark designed to evaluate complex temporal modeling and multi-event reasoning across multiple temporally disjoint segments. With a fine-grained label taxonomy and structured evaluation protocols, CrossTimeBench enables systematic analysis of cross-temporal reasoning.

Experiments on state-of-the-art multimodal models show that their performance drops when cross-temporal reasoning is required. Further analysis reveals that accuracy tends to decrease with larger temporal gaps, while models show limited attention coverage over key temporal evidence, highlighting challenges in temporal evidence integration. We hope CrossTimeBench can serve as a standardized testbed for cross-temporal video understanding and facilitate the development of more effective long-range temporal reasoning methods.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing of the paper, retrieve and discover related work, draft portions of the manuscript, and generate certain question-answer pairs and annotations in the dataset. We did not use generative AI tools for research ideation or execution, or for proving mathematical claims.

We have thoroughly reviewed, verified, and, where necessary, revised all AI-assisted content. The generated data were examined and validated following our established data construction and quality control procedures. The authors take responsibility for the research design, methodology, experimental results, data analysis, and final presentation of the paper. We take full responsibility for all final content of this work, including any text, claims, or other research artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work focuses on the cross-temporal reasoning capabilities of multimodal large language models (MLLMs) in complex videos. Since CrossTimeBench is constructed based on multiple existing video datasets and uses large language models for data filtering, question generation, annotation, and quality control, we pay particular attention to data provenance and release, as well as privacy protection.

Data Sources and Release. The data sources of CrossTimeBench are derived from existing publicly available or licensed video datasets. For each data source, we follow the corresponding terms of use and licensing requirements. We will not redistribute video content that is restricted by the original licenses. For annotations that are permitted to be publicly shared, we will release only the corresponding annotations and necessary video identifiers, allowing users to obtain restricted video content in accordance with the licenses of the original datasets. For the extended annotations constructed by us, we will clearly specify their provenance, construction process, and applicable licensing terms upon release.

Privacy and Personal Information. Since some source datasets contain real-world videos, they may include identifiable individuals, bystanders, and other personal information. We do not additionally collect or release personal information that is unrelated to the original datasets. For video content subject to the access and privacy policies of the original datasets, we will provide only the necessary identifiers and will not redistribute the videos or video frames. The example images and other visual materials presented in the paper will also avoid unnecessary disclosure of identifiable individuals, and the data documentation will clearly describe the relevant privacy and access restrictions of the source datasets.

Potential Misuse and Broader Impact. To mitigate potential risks, we will clearly document the intended use, known limitations, and high-risk use cases for which the benchmark is not intended. We also encourage researchers to comply with applicable legal, privacy, and ethical requirements when applying related models to real-world scenarios.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of CrossTimeBench, we provide comprehensive descriptions of the data construction pipeline, task definitions, taxonomy, and experimental settings throughout the paper, appendix, and supplementary materials. The main paper and appendix describe the data sources, data filtering and processing procedures, question construction and quality control methods, as well as the definitions and statistics of the different task categories. The supplementary materials further provide detailed descriptions of the data processing and construction pipeline, along with the prompts used for data generation and quality control. For the experiments, we report the models and evaluation metrics used to facilitate the reproduction of our experimental results. We will also release publicly shareable annotations, the extended annotations constructed by us, and the necessary video identifiers. For video content restricted by the licenses of the original datasets, we will not redistribute the videos or video frames, but instead provide the necessary video identifiers so that researchers can reproduce the experiments by obtaining the corresponding videos in accordance with the access requirements of the original datasets.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022.

Anthropic. Claude sonnet 4.6 system card, 2026. URL `https://www.anthropic.com/claude-sonnet-4-6-system-card`. Accessed: 2026-05-07.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

Shangzhe Di and Weidi Xie. Grounded question-answering in long egocentric videos. In *CVPR*, 2024.

Chaoyou Fu, Yi Dai, Yuhui Luo, Shuhuai Li, Jun Liao, Renrui Zheng, Ren Zhang, Yifei Wang, Yawei Wang, Huaxia Zhou, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23411–23422, 2025.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *ICCV*, 2017a.

Jiyang Gao, Zhenheng Yang, Kan Chen, Chen Sun, and Ram Nevatia. Turn tap: Temporal unit regression network for temporal action proposals. In *ICCV*, 2017b.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *CVPR*, 2024.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2758–2766, 2017.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*, 2017.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. TVQA: Localized, compositional video question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1369–1379, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1167. URL https://aclanthology.org/D18-1167/.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *PMLR*, 2023a.

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*, 2023b.

Zongyao Li, Kengo Ishida, Satoshi Yamazaki, Xiaotong Ji, and Jianquan Liu. Kfs-bench: Comprehensive evaluation of key frame sampling in long video understanding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5643–5652, 2026.

Karttikeya Mangalam, Ruslan Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *CVPR*, 2024.

Enxin Song, Wenhao Chai, Tian Ye, Jenq-Neng Hwang, Xi Li, and Gaoang Wang. Moviechat+: Question-aware sparse memory for long video question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(1):374–389, 2026. doi: 10.1109/TPAMI.2025.3604614.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. STAR: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding, 2024. URL https://arxiv.org/abs/2407.15754.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9777–9786, June 2021.

Junbin Xiao, Angela Yao, Yicong Li, and Tat Seng Chua. Can i trust your answer? visually grounded video question answering. In *CVPR*, 2024.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*, 2017.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016.

Zhou Yu, Jun Xu, Jianlong Yu, Qingming Zhao, Yu Wang, and Zhen Zhang. Activitynet-qa: A new benchmark for understanding complex video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 9127–9134, 2019.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 543–553, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.49. URL https://aclanthology.org/2023.emnlp-demo.49/.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

Junjie Zhou, Ziqi Zhang, Dong Zhu, et al. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

A Compute resources

We used 138.236M input tokens and 485k output tokens for the Claude model.

For open-source VLM evaluation, experiments with Qwen2.5-VL-7B, Qwen3-VL-8B, Qwen3-VL-4B, LLaVA-NeXT-Video-7B, and InternVL3-8B each required approximately 30 GPU hours. Experiments with Qwen3-VL-30B required roughly twice the compute ($\sim$60 GPU hours), as inference was performed using two GPUs in parallel.

InternVL3-8B and Qwen3-VL-30B were evaluated on NVIDIA H200 GPUs, while all other models were evaluated on NVIDIA A100 GPUs.

B Temporal Attention Analysis of Qwen2.5-VL-7B

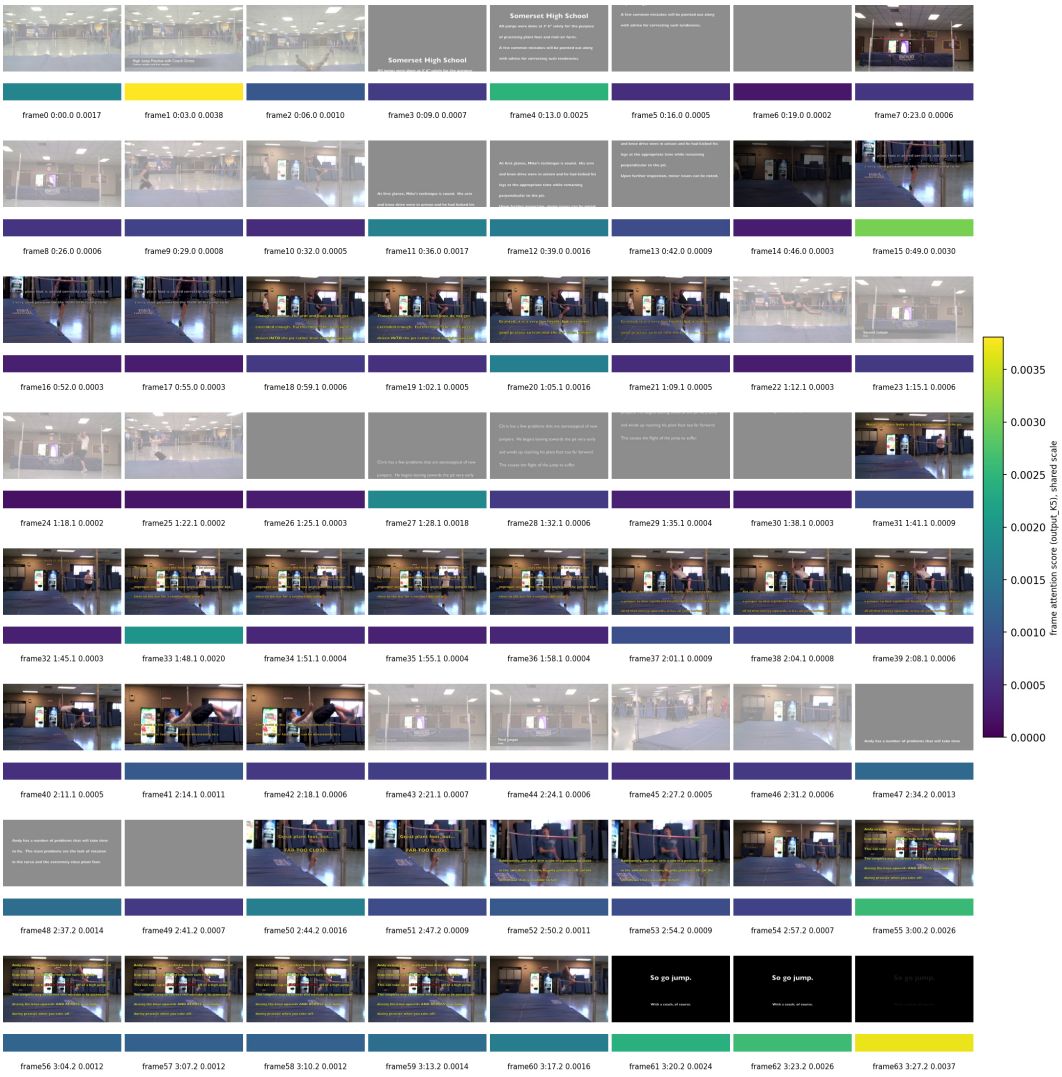
Taking Qwen2.5-VL-7B as an example, we analyze several correctly and incorrectly answered QA pairs along with their corresponding temporal evidence. We find that the frames attended to by the model are not always well aligned with the temporal evidence, as illustrated by the two cases in Fig. 7 and Fig. 8. Specifically, approximately 60% and 50% of the model's Top-10 attended frames fall within the corresponding evidence intervals in the two cases, respectively. For Question 1, although the question involves multiple temporal segments, it only requires summarizing the overall intent of the events. The model can therefore effectively integrate the relevant information and provide a correct answer. In contrast, Question 2 requires the model not only to identify temporally separated events, but also to summarize each event and establish their relationships to form a unified answer. This places greater demands on cross-temporal evidence integration, and the model fails to answer the question correctly.

C 35-category Label Baseline Experiment

Due to space constraints in the main text, we are unable to include the evaluation results of all 35-category label baselines; therefore, we provide these results in the appendix. See Table 5 and Table 6.

D Prompt Template

Prompt templates used for (a) scoring cross-temporal segment annotations; (b) determining question types; (c) identifying which types of questions can be generated from a given content; (d) synthesizing open-ended QA pairs; (e) evaluating the quality of synthesized QA pairs; (f) synthesizing closed-form QA pairs; (g) evaluating open-ended QA; (h) evaluating closed-form QA. See Figure 9, Figure 10, Figure 11, Figure 12, and Figure 13.

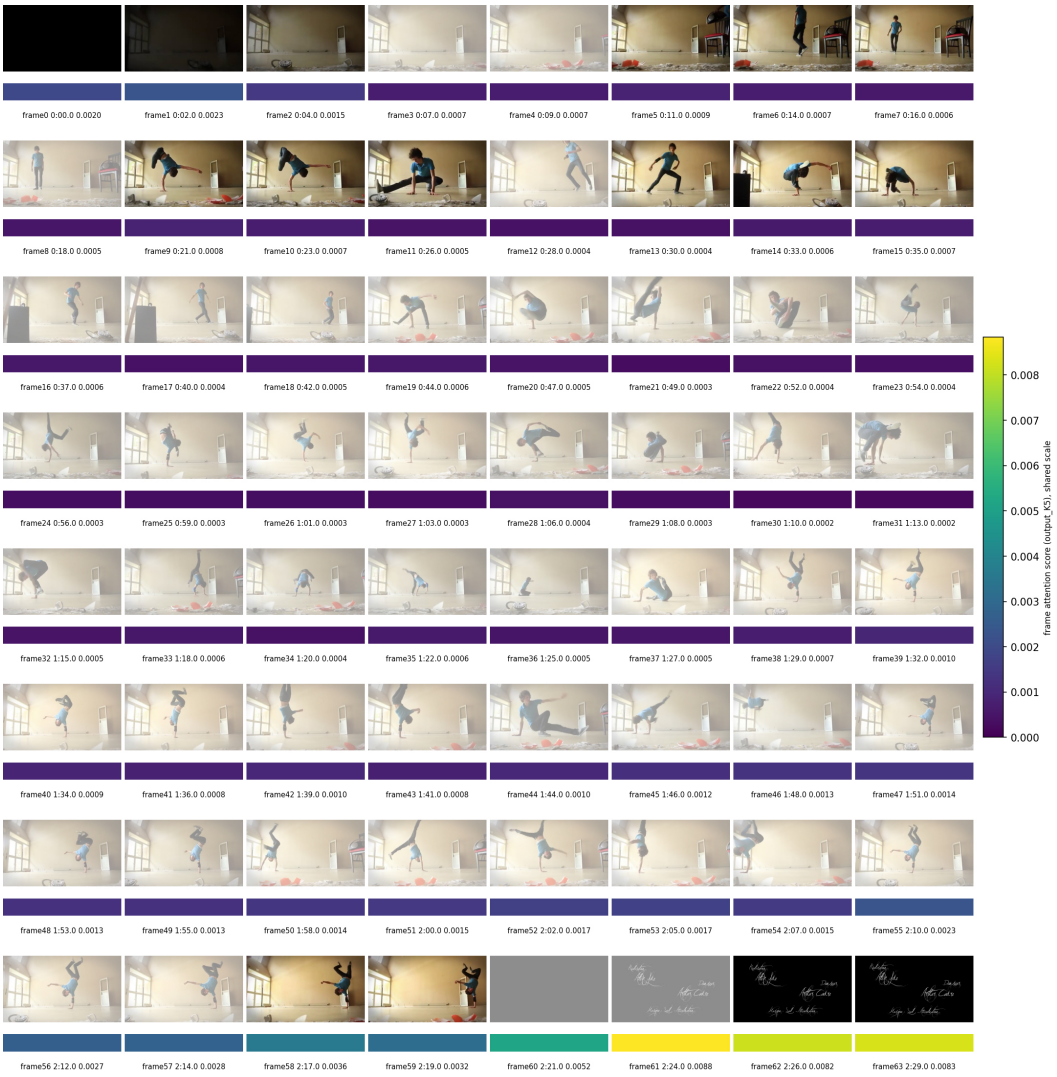


Q: What change in the high jump performance is highlighted through repeated replays involving different boys?

A1: The replays are used to highlight the boys' emotional reactions and audience reactions during their jumps.
A2: The replays emphasize the boys' successful landings and perfect form to celebrate their achievements. <-correct
A3: The repeated replays show technical errors made by each boy during their high jumps, with specific focus on form and execution flaws.
A4: The replays focus on the boys' equipment and mat conditions to analyze safety and infrastructure.

Answers from Qwen2.5-VL-7B-Instruct:
close: A2
open: The video highlights the importance of proper technique, including plant foot positioning, arm drive, and knee drive, which can significantly impact the height achieved in a high jump.

Figure 7: Correctly Answered Case. The QA pair is constructed by us, while the original video frames are sourced from the ActivityNet video v_0-F3q8Aj9Zk.



Q: How does the man's activity evolve from the beginning to near the end of the video?

A1: He walks in, warms up with spins, and ends by setting his stereo down again before the credits roll.
A2: He starts with a handstand during his warm-up, then transitions directly into credits without performing any breakdance moves.
A3: He begins by setting down his stereo immediately, then performs a full breakdance routine without any warm-up.
A4: He walks into the room, sets down his equipment, warms up with a handstand and leg stretches, performs a full breakdance routine, and finally the video ends with credits. <-correct

Answers from Qwen2.5-VL-7B-Instruct:
close: A3
open: The man's activity evolves from standing in a room with a chair and window, to performing various acrobatic moves such as handstands and cartwheels, and finally to standing on his hands with his legs in the air.

Figure 8: Incorrectly Answered Case. The QA pair is constructed by us, while the original video frames are sourced from the ActivityNet video v_Jifw8dC5yTM.

Table 4: Performance comparison across secondary label on the closed-form QA task.

Label	Qwen3-VL-30B-A3B	Qwen3-VL-4B	Qwen2.5-VL-7B	Qwen3-VL-8B	InternVL3-8B	LLaVA-NeXT-Video-7B	claude-4.6
AA.CA	0.24	0.32	0.30	0.22	0.33	0.27	0.14
AA.OI	0.84	0.68	0.54	0.70	0.47	0.34	0.05
AA.SA	0.70	0.62	0.44	0.65	0.40	0.32	0.07
TR.BA	0.49	0.40	0.52	0.55	0.41	0.27	0.02
TR.EP	0.80	0.59	0.36	0.57	0.43	0.30	0.10
TR.FP	0.90	0.19	0.53	0.75	0.54	0.63	0.04
TR.ST	0.90	0.20	0.54	0.81	0.47	0.44	0.14
SR.AL	0.65	0.27	0.42	0.56	0.39	0.36	0.11
SR.DJ	0.62	0.22	0.46	0.65	0.48	0.53	0.14
SR.OD	0.56	0.21	0.40	0.52	0.34	0.30	0.22
SR.PT	0.89	0.27	0.58	0.75	0.47	0.51	0.11
SR.RL	0.75	0.21	0.46	0.61	0.42	0.41	0.08
OU.OA	0.81	0.71	0.47	0.69	0.52	0.41	0.06
OU.OC	0.45	0.40	0.47	0.54	0.43	0.28	0.05
OU.OE	0.60	0.48	0.58	0.69	0.39	0.23	0.14
OU.OI	0.82	0.34	0.50	0.68	0.46	0.50	0.07
OU.OM	0.72	0.62	0.49	0.67	0.43	0.35	0.06
OU.OS	0.85	0.66	0.39	0.69	0.44	0.33	0.05
HU.AA	0.82	0.55	0.47	0.76	0.47	0.43	0.07
HU.E	0.85	0.56	0.52	0.79	0.47	0.39	0.11
HU.GA	0.69	0.31	0.48	0.63	0.42	0.39	0.09
HU.PI	0.80	0.39	0.51	0.76	0.44	0.47	0.08
HU.SI	0.79	0.68	0.54	0.74	0.50	0.31	0.09
CR.C	0.87	0.33	0.52	0.74	0.45	0.41	0.10
CR.E	0.74	0.63	0.43	0.70	0.50	0.38	0.09
CR.I	0.84	0.54	0.51	0.81	0.47	0.40	0.03
CR.MSR	0.90	0.19	0.56	0.78	0.54	0.47	0.08
SE.BA	0.80	0.37	0.55	0.75	0.49	0.45	0.16
SE.EC	0.96	0.58	0.50	0.77	0.49	0.44	0.07
SE.SC	1.00	0.79	0.44	0.68	0.52	0.46	0.07
SE.SL	0.57	0.62	0.42	0.64	0.45	0.31	0.11
DS.LS	0.88	0.50	0.56	0.78	0.50	0.32	0.33
DS.SD	0.91	0.41	0.47	0.79	0.50	0.41	0.17
DS.VT	0.94	0.44	0.60	0.74	0.48	0.53	0.02

Table 5: Performance comparison across primary label using F1 metric.

Label	Qwen3-VL-30B-A3B	Qwen3-VL-4B	Qwen2.5-VL-7B	Qwen3-VL-8B	InternVL3-8B	LLaVA-NeXT-Video-7B	claude-4.6
AA.CA	0.18	0.02	0.23	0.17	0.17	0.00	0.01
AA.OI	0.25	0.23	0.22	0.24	0.20	0.18	0.16
AA.SA	0.21	0.15	0.26	0.18	0.15	0.09	0.09
TR.BA	0.48	0.30	0.62	0.40	0.49	0.01	0.01
TR.EP	0.16	0.21	0.19	0.22	0.22	0.16	0.14
TR.FP	0.35	0.37	0.32	0.35	0.29	0.30	0.28
TR.ST	0.32	0.33	0.31	0.31	0.29	0.30	0.30
SR.AL	0.48	0.47	0.46	0.41	0.40	0.42	0.39
SR.DJ	0.17	0.12	0.53	0.04	0.06	0.12	0.54
SR.OD	0.47	0.56	0.52	0.39	0.50	0.47	0.46
SR.PT	0.32	0.34	0.31	0.32	0.28	0.29	0.31
SR.RL	0.49	0.48	0.53	0.36	0.49	0.50	0.45
OU.OA	0.43	0.22	0.27	0.27	0.28	0.21	0.15
OU.OC	0.41	0.26	0.51	0.40	0.44	0.01	0.01
OU.OE	0.61	0.46	0.80	0.61	0.63	0.01	0.02
OU.OI	0.47	0.37	0.41	0.34	0.39	0.39	0.40
OU.OM	0.29	0.22	0.26	0.26	0.24	0.19	0.14
OU.OS	0.21	0.24	0.22	0.26	0.24	0.18	0.17
HU.AA	0.20	0.22	0.21	0.29	0.19	0.17	0.16
HU.E	0.19	0.20	0.22	0.25	0.17	0.16	0.13
HU.GA	0.39	0.39	0.37	0.38	0.34	0.32	0.28
HU.PI	0.31	0.26	0.26	0.24	0.26	0.32	0.29
HU.SI	0.22	0.21	0.22	0.20	0.17	0.14	0.12
CR.C	0.26	0.26	0.27	0.24	0.24	0.27	0.24
CR.E	0.12	0.18	0.18	0.23	0.19	0.18	0.13
CR.I	0.24	0.24	0.25	0.29	0.19	0.18	0.16
CR.MSR	0.31	0.31	0.31	0.28	0.27	0.31	0.30
SE.BA	0.25	0.25	0.26	0.21	0.22	0.27	0.25
SE.EC	0.29	0.30	0.27	0.31	0.31	0.24	0.23
SE.SC	0.14	0.27	0.23	0.27	0.29	0.20	0.18
SE.SL	0.45	0.24	0.28	0.24	0.31	0.16	0.16
DS.LS	0.38	0.36	0.36	0.36	0.34	0.36	0.33
DS.SD	0.36	0.34	0.35	0.32	0.31	0.34	0.32
DS.VT	0.34	0.27	0.31	0.22	0.30	0.35	0.28

Table 6: Performance comparison across primary label using CIDEr metric.

Label	Qwen3-VL-30B-A3B	Qwen3-VL-4B	Qwen2.5-VL-7B	Qwen3-VL-8B	InternVL3-8B	LLaVA-NeXT-Video-7B	claude-4.6
AA.CA	0.46	0.05	0.57	0.44	0.43	0.00	0.00
AA.OI	0.50	0.22	0.42	0.43	0.22	0.17	0.07
AA.SA	0.42	0.21	0.65	0.02	0.20	0.04	0.01
TR.BA	1.19	0.58	1.55	1.04	1.23	0.03	0.00
TR.EP	0.12	0.09	0.07	0.24	0.12	0.04	0.00
TR.FP	0.64	0.62	0.41	0.66	0.40	0.45	0.16
TR.ST	0.28	0.27	0.23	0.19	0.17	0.21	0.16
SR.AL	1.18	1.16	1.01	0.65	0.74	0.79	0.35
SR.DJ	0.31	0.18	2.38	0.01	0.01	0.26	1.31
SR.OD	1.15	1.67	1.35	0.71	1.25	1.15	0.50
SR.PT	0.31	0.33	0.28	0.22	0.16	0.23	0.18
SR.RL	1.74	1.44	1.91	0.71	1.60	1.80	0.85
OU.OA	0.86	0.20	0.34	0.26	0.31	0.11	0.01
OU.OC	1.02	0.52	1.28	0.80	1.10	0.03	0.00
OU.OE	1.52	0.83	1.99	1.32	1.58	0.03	0.00
OU.OI	1.71	0.92	1.34	0.77	1.04	1.16	0.63
OU.OM	0.45	0.30	0.45	0.46	0.42	0.17	0.04
OU.OS	0.44	0.24	0.18	0.28	0.24	0.10	0.01
HU.AA	0.25	0.24	0.26	0.34	0.14	0.11	0.06
HU.E	0.21	0.26	0.35	0.28	0.12	0.13	0.05
HU.GA	0.64	0.62	0.44	0.58	0.43	0.43	0.16
HU.PI	0.67	0.48	0.39	0.38	0.45	0.46	0.29
HU.SI	0.38	0.20	0.41	0.29	0.14	0.07	0.01
CR.C	0.34	0.35	0.42	0.21	0.24	0.34	0.12
CR.E	0.17	0.14	0.13	0.27	0.16	0.07	0.01
CR.I	0.39	0.35	0.49	0.31	0.14	0.13	0.04
CR.MSR	0.24	0.21	0.21	0.16	0.15	0.23	0.25
SE.BA	0.15	0.16	0.16	0.12	0.10	0.18	0.13
SE.EC	0.26	0.23	0.13	0.23	0.19	0.08	0.04
SE.SC	0.41	0.20	0.09	0.21	0.25	0.07	0.00
SE.SL	1.13	0.14	0.30	0.18	0.39	0.05	0.00
DS.LS	0.20	0.15	0.17	0.13	0.12	0.14	0.10
DS.SD	0.23	0.22	0.23	0.17	0.15	0.17	0.13
DS.VT	0.32	0.29	0.29	0.16	0.19	0.29	0.15

```
"""
You are an evaluator for a video QA dataset. Your task is to determine if a question **requires reasoning across multiple distinct time segments** of a video, or if it can be answered from a single continuous segment.

## Scoring Guidelines

**Score 0** - Single segment sufficient
- Questions where all information is concentrated in one segment
- Uses "when" or "while" to describe simultaneous events
- No clear temporal sequence required
- *Example: "What is the color of the man's jacket?" → Color visible in single frame*

**Scores 1-3** - Minimal temporal separation
- Uses "before/after" but actions likely occur in rapid succession
- Events probably happen within the same continuous scene
- *Example: "Who walks in when Sheldon and Leonard are talking?" → Simultaneous events in same scene*

**Scores 4-7** - Clear temporal separation
- Uses "before/after" with actions clearly separated in time
- Events unlikely to be simultaneous
- *Example: "Did the woman pick up the cup before or after waving?" → Temporal order but may be in single segment*

**Scores 8-10** - Definite multi-segment requirement
- "How many times" questions counting separated events
- Requires explicit temporal counting or comparison
- *Example: "How many times did I pour fertilizer?" → Counting repeated actions needs multiple segments*

## Evaluation Principles

**Be Conservative**
- When uncertain, assign LOWER scores
- Penalize ambiguous questions that might be answerable from one segment
- Require clear evidence that multiple segments are NECESSARY

**Key Definitions**
- Time Segment: A continuous temporal interval (e.g., seconds 12-18)
- Multi-segment: Answer cannot be derived from any single short interval

## Output Format
```json
{
  "score": 0-10,
  "min_segments_estimate": 1/2/3...,
  "brief_reason": "One sentence explanation under 30 words"
}
Now evaluate this question:

<<QUESTION>>
"""
```

(a)

Figure 9: (a) The prompt template used for scoring cross-temporal segment annotations.
```

```
"""
You are an expert annotator for Video Question Answering datasets.
Your task is to assign one or more semantic types to each question–answer pair,
based on the fine-grained video understanding taxonomy below.

Each type has a primary category and a secondary subcategory.
Use the format:

primary_subcategory

Multiple labels should be separated by commas.
Do NOT invent new labels.

--------------------------------------------------------------------------------
TAXONOMY (Primary → Secondary)
--------------------------------------------------------------------------------

1. action_activity
   - simple_action
   - complex_activity
   - human_motion_detail
   - object_interaction

2. temporal_reasoning
   - before_after
   - duration
   - event_progression
   - state_transition
   - future_prediction

3. spatial_reasoning
   - absolute_location
   - relative_location
   - orientation_direction
   - path_trajectory
   - distance_judgement

4. object_understanding
   - object_identification
   - object_attribute
   - object_state
   - object_count
   - object_existence
   - object_movement

5. human_understanding
   - person_identification
   - appearance_attribute
   - emotion
   - gaze_attention
   - social_interaction
   - speech_content

6. causality_reasoning
   - cause
   - effect
   - intention
   - multi_step_reasoning

7. scene_environment
   - scene_classification
   - environment_condition
   - scene_layout
   - background_activity

8. description_summarization
   - short_description
   - long_summarization
   - video_topic

--------------------------------------------------------------------------------
TASK
--------------------------------------------------------------------------------
Given a question and an answer, output ALL relevant types from the taxonomy.

Only output the labels in the exact format:
labels: <label1>, <label2>, ...

Do NOT repeat the question or the answer.

--------------------------------------------------------------------------------
FEW-SHOT EXAMPLES

Example 1:
Question: What is the woman doing after she opens the door?
Answer: She walks inside the room.
Output:
labels: temporal_reasoning.before_after, action_activity.simple_action, spatial_reasoning.path_trajectory

Example 2:
Question: Why does the man drop the box?
Answer: Because it is too heavy for him to carry.
Output:
labels: causality_reasoning.cause, human_understanding.appearance_attribute, object_understanding.object_attribute

Example 3:
Question: Where is the boy sitting while eating?
Answer: He is sitting on the sofa.
Output:
labels: spatial_reasoning.absolute_location, action_activity.simple_action, human_understanding.social_interaction

--------------------------------------------------------------------------------
INPUT
Question: <<QUESTION>>
Answer: <<ANSWER>>
--------------------------------------------------------------------------------
OUTPUT
labels:
"""
```

(b)

Figure 10: (b) The prompt template used for determining question types.

```
"""
You are given a dictionary of candidate label types in string form:
<<label_type_dict>>
(key = label index number, value = label type name)

You are also given subtitles in dictionary string form:
<<subtitles_dict>>
(key = 1-based subtitle index, value = subtitle text)

You are also given a dictionary of label explanations:
<<label_explanation_dict>>
(key = label type name, value = explanation describing the label's semantics)

---

### Your task:
Select all label types whose semantics are **capable of producing exactly ONE (1) valid cross-temporal reasoning QA
pair**, satisfying all conditions below. Use the label explanations to help determine if a label can support cross-temporal
reasoning.

A label type is considered **valid (selectable)** only if it could support generating a QA question that:

#### Question format & quality:
- Is a **single, natural, standalone question** (not multiple questions stitched together)
- Is **directly shaped by the label's semantics**
- Requires a clear **reasoning goal aligned with the label**

#### Temporal & reasoning requirements:
- Requires reasoning across **multiple subtitle indices**
- Requires reasoning across **non-adjacent time segments**
- **Cannot be answered from a single subtitle index**
- Uses information from **at least THREE distinct subtitle indices whenever possible**
- Relies on **temporal inference, tracking, evolution, state change, comparison, or cross-moment correlation**
  (examples include identity inference, action continuity, object persistence, cause-effect chains, body-region reasoning
across steps, event progression, etc.)

#### Grounding constraints:
- Must be grounded **only in the provided subtitles**
- Must be based on actual subtitle content, **not external knowledge**

#### Output constraints for QA generation feasibility:
- The reference range would include **3 or more subtitle indices**
- The range would represent the **temporal span needed to infer the answer**
- Subtitle indices must be interpreted as **1-based**

---

### Final output format:
Return **ONLY a JSON list of valid label indices** (the keys from the label type dictionary), e.g.:

[1, 4, 6]

- Do NOT return label names
- Do NOT return explanations
- Do NOT wrap the output in code fences
- The output must be a **pure JSON array of integers**
"""
```

(c)

```
---
### Input:
You are given subtitles describing a video clip.
Each subtitle is indexed and loosely corresponds to a temporal segment of the video. The use of non-consecutive indices (e.g., [2, 4, 5]) serves as evidence for
cross-temporal reasoning.

Your task is to generate **EXACTLY ONE (1)** high-quality, cross-temporal QA pair
for the following label:

LABEL: <<single_label>>
LABEL_DESCRIPTION: <<single_label_description>>

---

### Core Generation Principles (STRICT PRIORITY ORDER):

1. **Cross-temporal reasoning is the highest priority.**
  - The QA must require reasoning across **multiple, temporally separated and not all consecutive moments**.
    For example, [2,3,5] is acceptable, but [2,3,4] is not.
  - Strongly prefer **3 or more subtitle indices** that reflect **different phases or states**.
  - Using only **2 indices** is acceptable *only if* they are clearly non-adjacent, such as [1,3],
    and still support meaningful temporal reasoning.
  - If cross-temporal reasoning cannot be satisfied without degrading quality, output `None`.

2. **Distinct temporal phases requirement.**
  - The selected subtitles must correspond to **distinct moments, stages, behaviors,
    or interaction patterns**, not a single continuous action.
  - QA pairs that rely on a single uninterrupted event
    (even if multiple indices are referenced) are invalid.

3. **Single, self-contained question constraint.**
  - The question must be **one natural, standalone sentence**.
  - Do NOT combine multiple questions or reasoning targets.
  - Do NOT ask "what and why", "what and how", or similar compound questions.
  - The question must have a clear reasoning objective aligned with the label.
  - If a valid benchmark-quality QA cannot be generated, output `None`.

4. **Visual grounding and modality independence constraint.**
  - Answerability must rely **primarily on visual content**, not audio or text.
  - Subtitles are **only proxies for visual evidence**, not sources of factual information.
  - The QA must be **comprehensible with muted audio** (visual-only understanding).
  - Avoid dependence on dialogue, narration, textual clues, or external knowledge.
---

### QA Construction Requirements:

#### Question:
- One complete, independent interrogative sentence
- Directly shaped by the **label's semantic intent**
- Focused on **visually observable actions, interactions, or temporal changes**
- NOT answerable from any single subtitle alone

#### Answer:
- A concise but complete response that directly answers the question
- Explicitly grounded in the visual events implied by the referenced subtitles
- No speculation, no external knowledge, no dialogue dependence

---

### Grounding Range Constraints:

- Use **only** the provided subtitles as evidence anchors
- Do NOT introduce entities, events, or motivations not implied by the subtitles
- Subtitles serve as proxies for visual evidence, not as textual facts

---

### Output Format (STRICT — NO DEVIATION):

If a valid QA pair can be generated, output **exactly one JSON object**:

{
  "q": "...",
  "a": "...",
  "range": [index_a, index_b, ...],
}

Where:
- `"range"` is an explicit set of **all subtitle indices directly relied upon**
- Indices must be **question&answer-based**
- **Absolute avoidance of fully consecutive indices**:
  Using a range consisting entirely of consecutive indices (e.g., [3, 4, 5]) fundamentally violates the cross-temporal reasoning principle and is strictly prohibito

If **any** of the following are true, MUST output `None`:
- The QA can be answered from a single moment
- The reasoning collapses to a single continuous event
- The indices are consecutive
- The result would not meet SOTA benchmark standards

---

Now generate the QA pair for the specified LABEL
based ONLY on the subtitles below.

<<subtitles_clip>>
"""
```

(d)

Figure 11: (c) The prompt template used for identifying which types of questions can be generated from a given content. (d) The prompt template used for synthesizing open-ended QA pairs.

```
"""
You are an extremely strict QA consistency evaluator.

Your decision must be binary and conservative.

A QA pair passes ONLY IF:
- The answer is explicitly and unambiguously supported by the subtitles.
- There is clear semantic evidence in the subtitles.
- No inference beyond the subtitles is required.

Fail the QA if:
- The answer adds extra details.
- The answer makes assumptions.
- The answer generalizes beyond the subtitles.
- The subtitles do not clearly confirm the answer.
- There is any contradiction.

Important:
When uncertain, choose fail.

Output exactly one word:
pass
or
fail

Question:
<<question>>

Answer:
<<answer>>

Subtitles:
<<subtitles>>
"""
```

(e)

```
"""
You are given a question and its correct answer.

Your task is to generate three plausible but incorrect answer options (distractors) for a multiple-choice
question.

Requirements:
1. Each option must be semantically related to the question and consistent with the context implied by the
correct answer.
2. All options must be clearly incorrect, but still realistic and plausible.
3. The distractors should be close to the correct answer in meaning, but contain subtle differences,
inaccuracies, or misleading details.
4. Avoid trivial negations (e.g., simply adding "not") or obviously wrong or irrelevant answers.
5. Ensure diversity among the three options — they should differ from each other in content and reasoning.
6. Match the style, tone, and length of the correct answer.
7. Do not copy phrases from the correct answer verbatim.

Input:
Question:
<<question>>

Correct Answer:
<<answer>>

Output:
Provide exactly three incorrect answer options, formatted as:
A. ...
B. ...
C. ...
"""
```

(f)

Figure 12: (e) The prompt template used for evaluating the quality of synthesized QA pairs. (f) The prompt template used for synthesizing closed-form QA pairs.

```
"""Answer the question concisely and directly.

- Provide a short answer (preferably one sentence).
- Do not include unnecessary details or explanations.
- Ensure the answer is complete and self-contained.

"""
```

(g)

```
"""
You are given a question and multiple-choice options.

You must produce TWO independent outputs:

IMPORTANT:
- The OPEN answer must be generated WITHOUT using the options.
- The MCQ answer must be selected based on the options.
- These two answers must be generated independently.

---------------------------------------------------------------------
STEP 1: OPEN-ENDED ANSWER (NO OPTIONS)
---------------------------------------------------------------------
First, answer the question as if NO options are provided.
Do NOT look at or use the options.
Provide a concise, correct natural language answer.

---------------------------------------------------------------------
STEP 2: MULTIPLE-CHOICE ANSWER
---------------------------------------------------------------------
Then, using the given options, select the best answer.
1. Carefully compare question with all options.
2. Choose the most correct option.
3. Do NOT explain reasoning.

---------------------------------------------------------------------
INPUT:
Question:
<<question>>

Options:
<<options>>
(Each option is labeled 1, 2, 3, 4, ...)

---------------------------------------------------------------------
OUTPUT FORMAT (STRICT):
OPEN_ANSWER:
<your open-ended answer>

MCQ_ANSWER:
<option index only: 1 / 2 / 3 / 4 / ...>
"""
```

(h)

Figure 13: (g) The prompt template used for evaluating open-ended QA. (h) The prompt template used for evaluating closed-form QA.