🐟 TUNA: Comprehensive Fine-grained Temporal Understanding Evaluation on Dense Dynamic Videos

Fanheng Kong[1][*], Jingyuan Zhang[2], Hongzhi Zhang[2], Shi Feng[1][†], Daling Wang[1],
Linhao Yu[2], Xingguang Ji[2], Yu Tian[2], Victoria W., Fuzheng Zhang[2]
[1]Northeastern University   [2]Kuaishou Technology
kongfanheng426@gmail.com, fengshi@cse.neu.edu.cn

Abstract

Videos are unique in their integration of temporal elements, including camera, scene, action, and attribute, along with their dynamic relationships over time. However, existing benchmarks for video understanding often treat these properties separately or narrowly focus on specific aspects, overlooking the holistic nature of video content. To address this, we introduce TUNA, a temporal-oriented benchmark for fine-grained understanding on dense dynamic videos, with two complementary tasks: captioning and QA. Our TUNA features diverse video scenarios and dynamics, assisted by interpretable and robust evaluation criteria. We evaluate several leading models on our benchmark, providing fine-grained performance assessments across various dimensions. This evaluation reveals key challenges in video temporal understanding, such as limited action description, inadequate multi-subject understanding, and insensitivity to camera motion, offering valuable insights for improving video understanding models. The data and code are available at https://friedrichor.github.io/projects/TUNA.

1 Introduction

Vision enables us to perceive the world, and video, as a key form of visual media, offers rich spatial and temporal information (Tang et al., 2023; Madan et al., 2024). With the rapid growth of video content, video understanding has become a crucial area of research, enabling applications that address the increasing volume of video data (Nguyen et al., 2024) and facilitate video generation as general-purpose simulators of the physical world (Brooks et al., 2024). Despite these advancements, the lack of robust evaluation methods remains a pressing challenge for the community. Accurate and comprehensive benchmarks are essential to assess the performance of video understanding models and
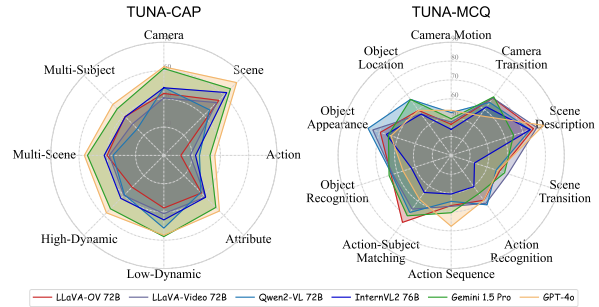

Figure 1: Performance of several advanced models on our TUNA. TUNA offers robust and interpretable evaluations on video captioning and QA tasks, providing clear guidance for advancements in video understanding.

improve their ability to interpret and analyze diverse video data effectively.

Recent works (Fu et al., 2024; Zhou et al., 2024) have evaluated video understanding across various tasks such as temporal perception and reasoning, video captioning, and long-video comprehension, providing metrics to guide the development of video LMMs. However, these evaluations often focus on specific aspects, such as subject actions, while neglecting other crucial video elements like camera states and background scenes along with the relationships between these elements (Chai et al., 2024; Xiong et al., 2024; Polyak et al., 2024). Additionally, the bias toward long-form videos (Fu et al., 2024; Li et al., 2024e; Mangalam et al., 2023) entangles video understanding with long-context modeling, making it difficult to attribute performance to specific capabilities. Furthermore, existing benchmarks lack an analysis of the model's sensitivity towards key factors affecting video understanding, such as diversity of video dynamics and visual characteristics. These limitations hinder comprehensive evaluation and effective error analysis to advance video understanding models.

To address the need for comprehensive video understanding, we introduce TUNA, a challenging multimodal benchmark for **T**emporal **U**nderstanding of dense dy**NA**mic videos. Unlike

Benchmark	#Videos	#Samp.	Anno.	Domain	Temporal Oriented	Scene Trans.	Captioning					VQA	
							Camera	Scene	Key.	Sem.	M.D.	Global	Fine.
VQA Benchmark													
NExT-QA (Xiao et al., 2021)	1,000	8,564	M	daily life	✗	✗	-	-	-	-	-	✗	✓
EgoSchema (Mangalam et al., 2023)	5,063	5,063	M&A	egocentric	✓	✗	-	-	-	-	-	✓	✗
PerceptionTest (Patraucean et al., 2024)	11,620	44,000	M	indoor	✓	✗	-	-	-	-	-	✓	✓
MVBench (Li et al., 2024d)	3,641	4,000	A	open	✓	✓	-	-	-	-	-	✓	✓
Video-MME (Fu et al., 2024)	900	2,700	M	open	✗	✓	-	-	-	-	-	✗	✓
MMBench-Video (Fang et al., 2024)	609	1,998	M	open	✗	✓	-	-	-	-	-	✓	✓
VideoVista (Li et al., 2024e)	894	24,906	A	open	✗	✓	-	-	-	-	-	✗	✓
TOMATO (Shangguan et al., 2024)	1,417	1,484	M	open	✓	✓	-	-	-	-	-	✓	✗
Captioning Benchmark													
DREAM-1K (Wang et al., 2024a)	1,000	1,000	M	open	✓	✓	✗	✗	✗	✓	✗	-	-
VDC (Chai et al., 2024)	1,027	1,027	A	open	✓	✓	✓	✓	✗	✓	✓	-	-
Multi-task Benchmark													
MLVU (Zhou et al., 2024)	1,334	2,593	M	open	✗	✓	✗	✓	✗	✓	✗	✗	✓
TempCompass (Liu et al., 2024f)	410	7,540	M&A	open	✓	✓	✗	✗	✗	✓	✓	✓	✓
E.T.Bench (Liu et al., 2024e)	7,002	7,289	M	open	✓	✓	✗	✗	✗	✓	✗	✓	✓
TemporalBench (Cai et al., 2024)	2,179	2,179	M	open	✓	✗	✗	✗	✗	✗	✗	✓	✓
TUNA	1,000	2,432	M&A	open	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison with various video understanding benchmarks across several aspects: number of videos (**#Videos**); number of samples (**#Samp.**); annotation method (**Anno.**, with M/A denoting manual/automatic); domain (**Domain**); temporal orientation (**Temporal Orientated**); presence of scene transitions (**Scene Trans.**); consideration of camera (**Camera**) and scene (**Scene**); use of keypoints (**Key.**) for controllability and interpretability; Judgement of semantically identical yet diverse representations (**Sem.**); availability of multi-dimensional scores (**M.D.**); if global (**Global**) and fine-grained (**Fine.**) understanding are concerned.

previous evaluations that focus on isolated video elements, TUNA emphasizes holistic video comprehension. We carefully curated 1,000 representative videos from diverse sources, spanning 12 domains such as Film and Driving, categorized across four visual characteristics: High-Dynamic, Low-Dynamic, Multi-Scene, and Multi-Subject. Each video in our dataset, TUNA-1K, is meticulously segmented into fine-grained events and annotated with detailed temporal captions, capturing camera states, background scenes, subject actions, object attributes. Table 1 shows the comparison with various vidoe understanding benchmarks.

Building on TUNA-1K, we propose TUNA, a multi-task benchmark towards temporal dynamics through two complementary tasks: TUNA-CAP for captioning and TUNA-MCQ for VQA. TUNA-CAP features an automated evaluation pipeline that performs event splitting, matching, and relationship classification, closely aligning with human judgment to assess dense captioning capabilities. TUNA-MCQ comprises 1,432 carefully crafted multiple-choice questions that specifically require full video context for accurate answers, ensuring that answers cannot be derived from a single frame or limited frames, providing a rigorous test of temporal understanding. Together, these tasks provide comprehensive evaluation metrics and valuable insights for advancing video understanding research.

We benchmark 21 popular LMMs on TUNA, revealing key challenges in video understanding.

Figure 1 shows the performance of selected models. Dense video captioning remains a difficult task, with GPT-4o (OpenAI, 2024) achieving the best performance but only reaching an F1 score of 58.5%, yet open-source models lag notably behind commercial models. Additionally, LMMs struggle with complex scenarios involving multi-scene, multi-subject, and high-dynamic video content. Interestingly, in the VQA task, open-source models demonstrate competitive performance. However, all models show consistent weaknesses in comprehending camera motion and action sequence. The notable performance disparity between captioning and VQA tasks underscores the current limitations in holistic video understanding capabilities. These findings provide crucial insights for advancing video LMMs, particularly in temporal and visual comprehension capabilities.

In summary, our contributions are:

- We introduce TUNA-1K, a meticulously annotated video-caption dataset that captures fine-grained temporal dynamics across camera, scene, action, and attribute on dense dynamic videos.
- We develop TUNA, a novel benchmark for comprehensive temporal video understanding, measuring the performance across various novel dimensions, such as different visual characteristics, temporal elements and video complexities.
- We conduct a comprehensive evaluation of several popular models, uncovering their strengths and weaknesses across various dimensions.
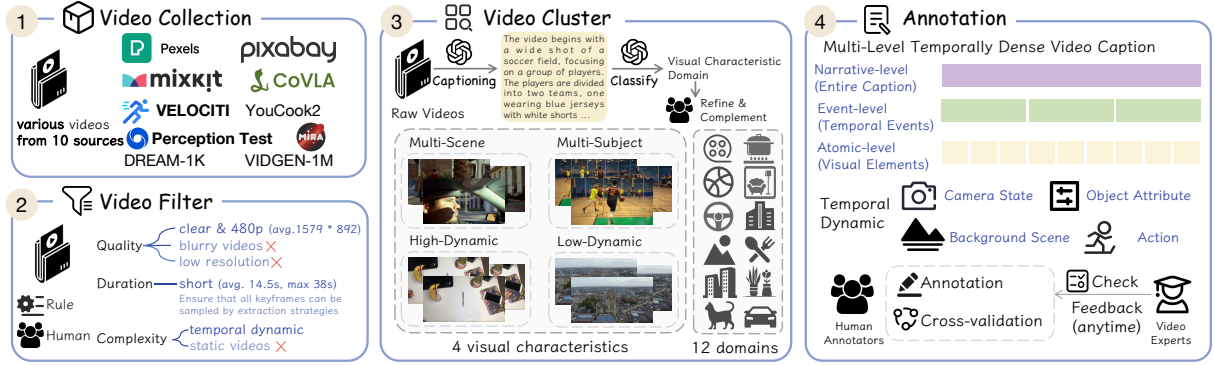
Figure 2: Overview of TUNA-1K construction. We collect and filter high-quality, short videos featuring dynamic temporal content from various sources. Each video is then categorized based on its visual characteristics and domain. Trained annotators provide temporally dense descriptions, followed by cross-validation. Video experts continuously review annotations, guiding annotators to refine their works, thus ensuring quality of the annotations.

Hopefully, this provides solid guidance for advancing video understanding.

2 Related Work

Video Captioning. Recent works (Zhang et al., 2024d,f; Chen et al., 2024a; Liu et al., 2024d) have revealed the importance of detailed captions for video understanding. Compared to image captioning, video captioning presents a greater challenge as it requires advanced techniques to handle the diversity of human and object appearances in various scenes, as well as their evolving relationships over time (de Souza Inácio and Lopes, 2023). While video captioning data is served as training data for video LMMs, it is challenging to robustly and interpretably evaluate video captioning. Traditional n-gram overlaps based metrics (Papineni et al., 2002; Lin, 2004; Vedantam et al., 2015) fail to measure genuine semantic similarity, with weakly consistency with human judgement. LLM-based scoring methods (Chan et al., 2023; Maaz et al., 2023) can deal with captions with the same semantics yet distinct expressions, but directly asking LLM to generate digital scores is not dependable due to their ambiguous meaning of each rating. Recently, Dream-1K (Wang et al., 2024a) evaluates captions from events, providing a robust results. However, these efforts don't centre on temporal dynamics, overlooking the essential features of video and pay minimal attention to changes in camera and scene.

Video QA. Recent works has provided benchmarks for comprehensively evaluating video LMM's ability to understand video, e.g., Video-MME (Fu et al., 2024), MLVU (Zhou et al., 2024). Temporal dynamics are crucial as a unique feature of video. Existing temporal understanding benchmarks (Pa-

traucean et al., 2024; Li et al., 2024d) focus on restricted scenes (e.g., indoor, egocentric), or just on subject's actions and attributes, without attention to changes in camera and scene, which are incomplete for temporal understanding evaluation. Our TUNA aims to comprehensively evaluate temporal perception skills towards open-domain videos.

3 TUNA

In this section, we present TUNA-1K, a temporally dense video-caption dataset, and TUNA, a multi-task temporal understanding benchmark.

3.1 TUNA-1K

The construction workflow of TUNA-1K is shown in Figure 2, consisting of four major phases: video collection, filter, cluster, and annotation.

Video Collection. Temporally dense videos should have diverse contents and include changes in the camera states and scenes besides subject actions and object attributes (Polyak et al., 2024; Xiong et al., 2024). To capture these complexities, we carefully collect 1,000 open-domain videos from 10 sources: (1) *Academic Video Understanding Data*: DREAM-1K (Wang et al., 2024a), Perception Test (Patraucean et al., 2024), VELOCITI (Saravanan et al., 2024), YouCook2 (Zhou et al., 2018); (2) *Academic Video Generation Data*: MiraData (Ju et al., 2024), VIDGEN-1M (Tan et al., 2024)); (3) *Other Academic Video Data*: CoVLA (Arai et al., 2024); and (4) *Web Data*: Pexels (Pexels, 2023), Pixabay (pixabay, 2023), MixKit (mixkit, 2023). Unlike concurrent works (Cai et al., 2024), we maintain the original videos containing multiple scenes or complex actions without segmenting them into clips, as these are essential for our tasks.

Video Filter. We remove blurry, low-resolution and long duration videos to ensure that our videos are
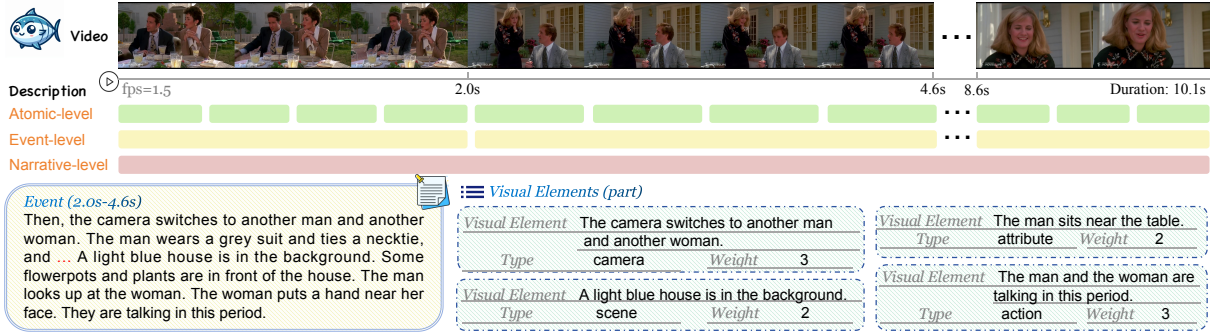
Figure 3: An instance in TUNA-1K consists of three levels of description: (a) an overall caption (Narrative-level), (b) a chronological sequence of events (Event-level), and (c) fine-grained visual elements (Atomic-level) along with their types and weights. A complete sample can be found in Figure 15.

high-quality and short, with an average resolution of $1579*892$, and an average duration of $14.5s$. We select short videos to ensure that sampling frame strategies can extract all keyframes from videos, to purely examine the video understanding ability. To ensure the videos are temporal-dynamic, one criterion is rich in either camera motion, scene transitions, or subject activities. Coarse filtering (e.g., resolution) is achieved by rules, and humans make complex filters (e.g., dynamic degree).

Video Cluster. We use GPT-4o (OpenAI, 2024) to generate description for each video, and cluster videos based on their descriptions, including four visual characteristics and 12 domains. Annotators then correct and complete the classification results.

Annotation. Existing models often miss critical events (Wang et al., 2024a), and lack sensitivity to camera states, struggling to accurately describe camera changes (Chai et al., 2024). Consequently, generating temporally dense video captions through automatic methods is challenging. Instead, our data is manually annotated. Trained human annotators are tasked to provide detailed video descriptions, focusing on camera states, background scenes, subject actions, and object attributes. The target captions features several chronologically evolving events, without summaries and subjective feelings. Additionally, annotators split each event into multiple visual elements, assigning types and weights to these elements. The types include camera, scene, action, and attribute, while weights indicates the element's importance for the video on a scale of 1-3.

Formally, a typical instance in TUNA-1K involves a collection of temporally evolving events $E_{ref} = [r_1, r_2, \ldots, r_T]$ forming an overall caption C_{ref}, where T denotes the count of events in the sequence. Each event r_i further contains various visual elements $V_i = \{v_{i1}, \ldots, v_{i,n_i}\}$, where n_i represents the number of visual elements in event r_i.

Moreover, each visual element v_{ij} is labelled with a type $t \in \{\texttt{camera}, \texttt{scene}, \texttt{action}, \texttt{attribute}\}$ and their weight $w_{ij} \in \{1, 2, 3\}$. An example of TUNA-1K is shown in Figure 3.

Quality Review. All annotated video-caption pairs undergo cross-inspection by annotators. In parallel, video experts (non-authors) review the annotations, providing feedback and prompting annotators to refine results to ensure high-quality annotation.

3.2 TUNA

3.2.1 Task Definition

Temporal dynamics distinguish videos from static images. While several benchmarks consider temporal sequences, they sole focus on actions and attributes, neglecting changes in camera state and scene. Additionally, some evaluation tasks fail to capture the perception ability of relationships and evolution of various elements in the video. For example, many questions are just about a single frame cue in the video. To fill these gaps, we emphasize the in-context understanding throughout the entire video, and measure temporal understanding across 4 key dynamic elements: camera state, background scene, subject action, and object attribute. Specifically, we introduce two complementary tasks: TUNA-CAP for captioning and TUNA-MCQ for VQA.

3.2.2 TUNA-CAP

An effective way to evaluate the temporal understanding ability of LMMs is reflected by their captioning skills (de Souza Inácio and Lopes, 2023; Chen et al., 2024a). However, it remains a challenge to reliably and interpretably assess the correctness and completeness of video captions. Event-level methods (Wang et al., 2024a, 2022) have proven effective but solely focus on subject actions, overlooking camera states and scenes. To this end, we propose a strategy to assess the tem-
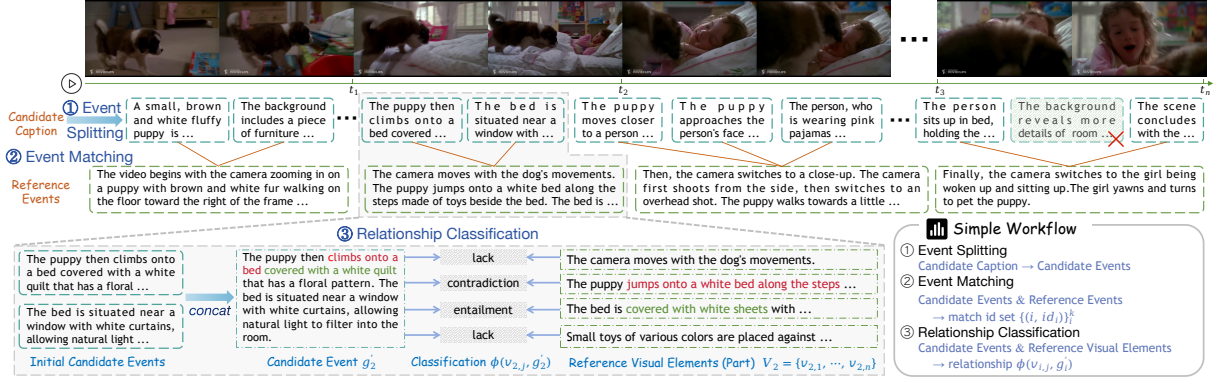
Figure 4: Overview of the evaluation workflow for TUNA-CAP. We first split candidate caption into multiple events and match them to reference events in TUNA-1K. Then we discard the mismatched events (useless content or inconsistent chronology), and connect the matched candidate events with the same reference event, considering the temporal sequence of the captions. Finally, we classify the relationship of visual elements to the candidate event.

porally dense captions that incorporate dynamic elements evolutions over time.

As shown in Figure 4, our evaluation proceeds in three stages: (1) Event Splitting, (2) Event Matching, and (3) Relationship Classification.

Event Splitting & Matching. To examine temporal perception skills through model-generated captions, we consider that an effective solution is to verify whether the models accurately describe several events in the correct temporal sequence. To achieve this, the candidate caption C_{gen} is first split into an event sequence $G = [g_1, g_2, \ldots, g_k]$. Then, each candidate event g_i is matched to a reference event r_j. Formally, the target is to obtain $\{(i, id_i)\}_{i=1}^{k}$ pairs, where $id_i \in \{1, \ldots, T, \text{None}\}$ denotes the index of the reference event r_{id_i} matched with the candidate event g_i and $id_1 \leq id_2 \leq \cdots \leq id_k$. These ensures that events which are effective and described in a correct temporal order are extracted.

Relationship Classification. For the captioning task, the classification-based approach is more interpretable and robust than the direct scoring methods (Wang et al., 2024a). Each reference event r_j corresponds to a set of visual elements V_j. Thus, we can transition from a tuple of concatenated candidate events with reference events (g_i', r_j) to a tuple of candidate events with visual elements (g_i', V_j). Subsequently, the relationship $\phi(v_{ij}, g_i') \in \{\text{entailment}, \text{lack}, \text{contradiction}\}$ between visual element v_{ij} and candidate event g_i' is classified. This element-based approach improves the interpretability of the evaluation. The workflow is implemented by GPT-4o (OpenAI, 2024), an LLM with powerful instruction-following capabilities.

Metrics. We employ precision (P) and recall (R)

to measure the correctness and completeness of the captions, introducing a novel metric calculation:

$$P = \frac{\sum_{i=1}^{T} \sum_{j=1}^{n_i} \mathbb{1}\left(\phi(v_{ij}, g_i') = \text{ent.}\right) \cdot w_{ij}}{\sum_{i=1}^{T} \sum_{j=1}^{n_i} \mathbb{1}\left(\phi(v_{ij}, g_i') \in \{\text{ent.}, \text{con.}\}\right) \cdot w_{ij}} \quad (1)$$

$$R = \frac{\sum_{i=1}^{T} \sum_{j=1}^{n_i} \mathbb{1}(\phi(v_{ij}, g_i') = \text{ent.}) \cdot w_{ij}}{\sum_{i=1}^{T} \sum_{j=1}^{n_i} w_{ij}} \quad (2)$$

$$F1 = \frac{2 \times P \times R}{P + R} \quad (3)$$

where $\mathbb{1}(\cdot)$ denotes the indicator function. Recognizing that each visual element v_{ij} has a distinct importance within the video, each element is weighted by its corresponding factor w_{ij}.

3.2.3 TUNA-MCQ

Based on fine-grained TUNA-1K, we design a pipeline, integrating automatic construction and manual refinement, to create instructions for multi-choice questions. The pipeline involves two main flows: error-prone points extraction and multi-choice QA generation. We consider 10 task types: 1) *camera motion*, e.g, zooming, panning, and rotating. 2) *camera transition*. 3) *scene description*. 4) *scene transition*. 5) *action recognition*. 6) *action sequence*. 7) *action-subject matching*. 8) *object recognition*. 9) *object appearance*, e.g., age, dress, color, shape, number. 10) *object location*. Beyond previous works (Li et al., 2024d; Liu et al., 2024f) that focus on subject actions and object attributes, we additionally emphasize camera states and scene transitions, to provide a more comprehensive assessment of temporal understanding.

Error-prone Points Extraction. To generate challenging questions, we develop an automatic approach to identify error-prone points in videos. The process involves feeding video frames and their ground-truth descriptions to video LMMs, which

Model	Dynamic Element Type				Visual Characteristic				Overall
	Camera	Scene	Action	Attribute	Low-Dynamic	High-Dynamic	Multi-Scene	Multi-Subject	
Open-Source LMMs									
PLLaVA-7B	49.4/22.6/28.9	52.2/30.9/36.6	30.5/12.6/16.5	44.5/19.5/25.3	66.5/23.0/32.7	56.6/17.1/24.7	55.7/15.5/22.8	56.2/15.3/22.5	60.0/19.1/27.4
LongVA-7B	52.3/26.0/32.5	56.5/34.4/40.6	38.9/17.2/22.0	50.6/22.0/28.4	75.9/26.5/37.3	69.4/20.1/29.0	68.3/19.0/27.6	67.3/15.7/23.7	71.6/22.3/31.8
Tarsier-7B	56.9/27.3/34.8	45.3/28.2/33.1	56.7/28.9/36.2	56.4/26.0/33.3	**81.2**/34.3/46.5	68.7/24.5/34.5	71.7/25.3/35.8	67.8/23.2/33.2	73.0/27.9/38.6
Kangaroo	65.2/36.5/44.1	67.8/45.4/51.9	49.3/26.0/31.9	59.8/32.2/39.5	73.2/34.7/45.6	67.6/31.3/41.1	66.2/29.7/39.3	63.5/26.3/35.7	69.5/32.5/42.7
LLaVA-OV-7B	75.2/42.0/51.0	71.8/51.2/57.6	54.1/30.4/36.8	66.2/42.0/49.3	78.6/38.4/50.0	71.0/38.8/48.9	71.7/38.3/48.4	67.1/33.8/43.8	73.6/38.6/49.3
LLaVA-Video-7B	74.0/41.5/50.4	73.6/52.3/58.9	57.0/30.8/37.8	72.1/44.8/53.1	80.7/40.0/52.2	75.1/39.5/50.3	77.1/38.6/50.0	73.5/34.6/45.8	77.0/39.7/51.0
Qwen2-VL-7B	72.3/40.7/49.0	71.9/50.0/56.7	55.9/30.1/37.0	68.2/38.4/46.7	81.2/42.0/53.8	76.0/35.3/46.4	76.8/33.2/44.4	73.6/28.9/39.9	77.8/37.6/48.9
InternVL2-8B	64.8/33.7/41.7	59.4/38.7/44.7	45.2/24.7/30.0	59.8/35.5/42.3	71.6/34.0/44.5	64.9/29.7/38.9	65.6/29.1/38.4	61.5/26.6/35.2	67.2/31.1/40.8
MiniCPM-V-2.6	76.5/47.8/56.0	75.0/54.1/60.6	57.2/31.8/38.8	68.7/42.3/50.2	79.3/41.4/53.0	74.3/40.4/51.0	76.5/40.8/51.7	73.5/38.3/49.0	76.0/40.7/51.7
PLLaVA-34B	60.8/29.6/37.4	56.2/33.7/39.9	38.7/17.3/22.3	55.1/26.1/33.2	74.5/28.1/38.9	64.3/22.6/31.8	63.9/21.3/30.2	60.7/19.2/27.6	67.8/24.5/34.2
Tarsier-34B	63.6/34.3/42.3	59.0/38.4/44.4	65.6/39.9/47.6	63.6/34.3/42.2	79.6/37.2/49.1	75.8/36.5/47.8	77.6/38.1/49.6	74.4/36.0/47.3	77.1/36.7/48.2
LLaVA-OV-72B	73.5/43.7/51.9	71.5/51.1/57.5	51.2/30.2/36.0	65.7/41.4/48.8	75.4/37.3/48.6	71.3/36.7/45.9	71.4/40.1/50.1	72.3/39.1/49.4	72.7/39.2/49.6
LLaVA-Video-72B	72.7/41.7/50.3	71.1/49.9/56.4	55.7/32.7/39.3	68.1/43.2/50.8	77.3/39.2/50.6	71.9/39.8/50.0	73.9/38.6/49.3	70.5/35.1/45.7	73.7/39.6/50.2
Qwen2-VL-72B	73.6/45.9/54.0	67.6/46.3/52.8	59.1/35.7/42.6	66.6/40.7/48.5	79.2/44.6/55.7	72.4/39.3/49.7	73.6/37.2/48.0	69.1/32.8/43.3	74.7/41.1/51.7
InternVL2-76B	75.1/45.4/53.9	73.3/55.8/61.4	55.7/34.9/41.2	64.3/44.5/50.9	72.0/43.1/52.8	70.1/41.9/51.5	71.4/41.1/51.1	68.6/39.7/49.3	70.7/42.3/51.9
Closed-Source LMMs									
Gemini 1.5 Flash	74.6/52.8/59.6	77.2/59.3/65.1	58.7/36.4/42.9	69.0/48.4/55.2	74.0/46.5/56.0	72.0/46.4/55.5	73.4/46.2/55.9	73.4/46.2/55.9	72.7/46.4/55.7
Gemini 1.5 Pro	78.7/53.0/60.7	75.7/57.4/63.3	59.0/40.3/46.3	69.0/49.4/56.0	76.7/48.7/58.7	72.1/47.8/56.7	73.4/47.7/57.0	69.9/44.1/53.3	73.7/48.1/57.4
GPT-4o	80.1/53.3/61.3	79.5/60.2/66.4	64.0/41.1/48.0	73.8/50.1/57.8	79.1/47.3/58.2	77.0/48.6/58.7	78.7/47.2/58.1	76.8/44.4/55.5	77.7/48.2/58.5

Table 2: TUNA-CAP performance of representative video LMMs. We provide detailed scores for selected tested models in various perception skills and visual characteristic categories. Each cell contains "**Precision / Recall / F1 Score**". The best and second-best results are marked with **bold** and underline, respectively.

then identify visual elements that appear inconsistent with the textual descriptions. Leveraging LMMs' inherent limitations in visual interpretation, we utilize their misidentified elements as naturally occurring error-prone points for question generation.

Multi-Choice QA Generation. Based on predefined task types, error-prone points and textual descriptions, LLM generates several multi-choice questions for each video. To ensure these questions effectively capture temporal dynamics, we employ a temporal-indispensability filtering mechanism similar to MMBench-Video (Fang et al., 2024). Specifically, a question is considered temporal-indispensable only if it cannot be correctly answered using a single frame but requires n frames (default $n = 16$) for accurate comprehension. This rigorous filtering process helps maintain a high temporal-indispensability ratio in TUNA-MCQ.

Quality Review. To ensure that data is high-quality and time-sensitive, we employ crowdsourcing to further filter and refine the automatically constructed data. In addition, human annotators perform cross-inspections to ensure annotation quality.

4 Experiments

4.1 Settings

We evaluate 21 closed-source models and open-source models with various sizes, including: Gemini 1.5 Pro (Reid et al., 2024), Gemini 1.5 Flash (Reid et al., 2024), GPT-4o (OpenAI, 2024), PLLaVA (Xu et al., 2024), LongVA (Zhang et al., 2024c), Tarsier (Wang et al., 2024a), InternVL2

(Chen et al., 2024b), Kangaroo (Liu et al., 2024c), LLaVA-OneVision (Li et al., 2024a), MiniCPM-V-2.6 (Yao et al., 2024), LLaVA-Video (Zhang et al., 2024f), and Qwen2-VL (Wang et al., 2024b).

By default, we uniformly sample 32 frames from each video, which is sufficient to capture the entire content of videos in our TUNA. Some models have varying constraints on input length or specific recommended settings. To accommodate these variations, we employ tailored sampling strategies for these models. More details are available in Appendix B.1 and Appendix C.3.

4.2 Video Captioning

We evaluate the temporal understanding skills of the models and their abilities to perceive videos towards different dynamic elements and visual characteristics. Precision reflects the correctness of the content mentioned in the descriptions, while recall reflects the completeness of the descriptions. As shown in Table 2, majority of video LMMs achieve a precision over 70%, but recall is below 50%, indicating that many visual elements in videos are often overlooked or misdescribed. The state-of-the-art model GPT-4o only achieve an F1 score of 58.5%, with a recall of 48.2%, highlighting that LMMs still have a great potential for improvement in the task of temporally dense captioning.

Temporal Dynamic Elements. Recent researches in video understanding and video generation have increasingly emphasized the dynamics of camera states and scenes (Chai et al., 2024; Xiong et al., 2024; Polyak et al., 2024). In this work,
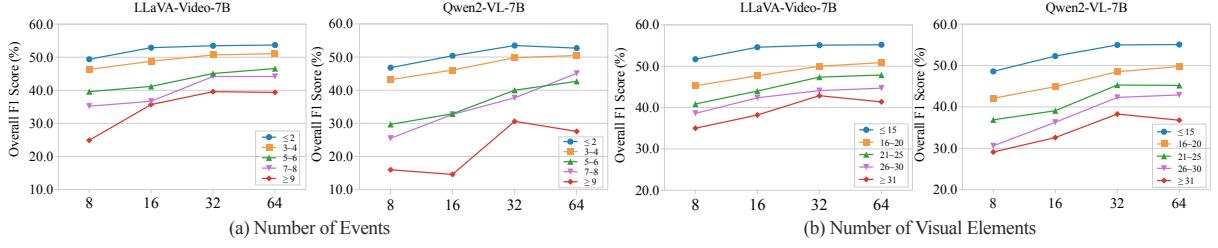
Figure 5: Performance comparison of different input frames with different video complexity for models trained in long contexts (over 8K tokens). The horizontal coordinate is the number of input frames.

we comprehensively thoroughly analyze four key dynamic element types: *camera*, *scene*, *action*, and *attribute*, aiming to explore the challenges that existing models face in captioning dynamic videos. As shown in Table 2, LMMs demonstrate superior performance in scene perception compared to the other dimensions. Existing LMMs often extract multiple frames from videos and treat them as a series of static images, facilitating a better grasp of static visual scenes. However, *camera* and *attribute* elements, which assess overall dynamics and fine-grained perception respectively, remain challenging, with highest scores reaching only 61.3% (56.0% for open-source models) for camera and 57.8% (53.1% for open-source models) for attribute. Notably, action perception shows consistently weaker performance across almost all models compared to the other dimensions, indicating substantial shortcomings in accurately describing the dynamic actions. An interesting exception is Tarsier-34B, which performs exceptionally well in the action dimension, falling only 0.4% behind GPT-4o. This aligns with its strong performance on DREAM-1K (Wang et al., 2024a), a video captioning benchmark focused on action events.

Diverse Visual Characteristics. As shown in Table 2, large performance disparities emerge when models process videos with different visual characteristics. All tested models perform better with low-dynamic content, but they struggle with high-dynamic and multi-scene videos, and show the weakest performance when handling videos containing multiple subjects.
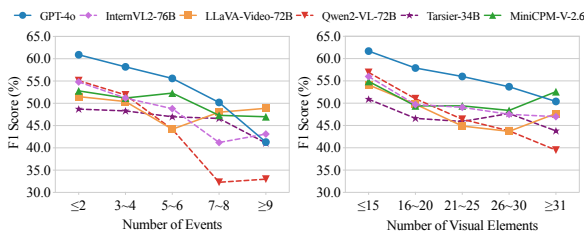

Figure 6: Performance comparison across different video complexities.

Video Complexity. We partition TUNA-1K based on the number of events and visual elements to investigate how increasing video complexity affects model performance. As shown in Figure 6, the F1 scores demonstrate a consistent downward trend as video complexity increases, indicating that the comprehension of complex videos remains a formidable challenge for current models. More details are available in Appendix B.2.1.

Enrichment of Visual Inputs. To explore the challenges posed by complex videos, we further investigate the impact of increasing frame number on videos with varying complexity. As shown in Figure 5, we analyze LLaVA-Video and Qwen2-VL, both trained with longer context lengths. Our findings reveal that F1 scores decrease with increasing video complexity at any given number of input frames. Generally, increasing the number of frames results in greater improvements for more complex samples, suggesting that complex videos require more frames for a complete and precise description. Counterintuitively, an unexpected pattern emerges: for the most complex videos, increasing frames from 32 to 64 actually reduces performance, indicating that highly complex videos remain a prominent challenge for LMMs. Further details be found in Appendix B.2.2.

Measure	Kendall's τ	Spearman's ρ	Pearson r
METEOR (Banerjee and Lavie, 2005)	30.8	44.8	54.7
BERT-Score (Zhang et al., 2019)	27.4	34.8	49.2
CLAIR (Chan et al., 2023)	45.6	56.6	41.0
DREAM-1K (Wang et al., 2024a)	22.2	31.3	24.7
TUNA-CAP	57.2	76.7	69.9

Table 3: Human judgment correlation scores for our automatic evaluation. All p-values < 0.05.

Correlation with Human Judgments. To validate the effectiveness and robustness of our automatic evaluation method, we calculate Kendall's τ, Spearman's ρ, and Pearson r correlation scores between several methods and human evaluation. As shown in Table 3, these results demonstrate strong correlation, confirming that our method provides a

Model	Camera State		Background Scene		Subject Action			Object Attribute			Overall
	Motion	Transition	Description	Transition	Recognition	Sequence	Matching	Recognition	Appearance	Location	
Open-Source LMMs											
PLLaVA-7B	29.7	31.9	48.1	22.4	43.6	34.6	30.4	32.3	38.1	45.2	33.7
LongVA-7B	37.5	41.5	63.0	30.8	44.6	44.7	43.5	41.7	47.6	40.5	42.4
Tarsier-7B	23.0	24.6	40.7	20.6	38.6	26.9	45.7	20.9	25.9	23.8	26.5
Kangaroo	33.2	47.3	53.7	38.3	49.5	38.8	54.3	47.2	43.5	59.5	42.9
LLaVA-OV-7B	42.2	54.6	57.4	48.6	42.6	41.4	60.9	47.9	50.0	59.5	47.4
LLaVA-Video-7B	39.1	50.7	59.3	46.7	52.5	52.4	56.5	53.6	61.9	47.6	50.6
Qwen2-VL-7B	41.0	51.7	66.7	45.8	54.5	52.8	65.2	49.0	60.2	57.1	51.3
InternVL2-8B	41.0	53.1	66.7	40.2	45.5	50.5	50.0	45.8	56.8	45.2	48.4
MiniCPM-V-2.6	39.8	45.9	59.3	34.6	49.5	51.1	52.2	42.2	46.6	50.0	45.7
PLLaVA-34B	42.6	41.5	63.0	43.9	45.5	48.5	56.5	43.2	56.8	57.1	46.9
Tarsier-34B	43.0	48.3	72.2	45.8	51.5	50.2	56.5	49.7	53.7	<u>61.9</u>	50.1
LLaVA-OV-72B	46.5	**67.6**	<u>75.9</u>	<u>57.0</u>	59.4	<u>56.6</u>	**73.9**	**63.5**	69.5	59.5	<u>60.0</u>
LLaVA-Video-72B	<u>47.7</u>	**67.6**	**77.8**	**61.7**	<u>61.4</u>	**57.0**	65.2	62.5	<u>73.7</u>	57.1	**60.7**
Qwen2-VL-72B	**52.7**	<u>64.7</u>	74.1	55.1	**62.4**	54.4	<u>67.4</u>	<u>63.0</u>	**76.3**	**66.7**	**60.7**
InternVL2-76B	43.8	61.8	74.1	43.0	50.5	50.5	54.3	52.1	66.1	57.1	53.1
Closed-Source LMMs											
Gemini 1.5 Flash	40.8	<u>58.3</u>	<u>70.4</u>	52.3	48.0	54.2	<u>63.0</u>	49.0	66.7	<u>64.3</u>	53.3
Gemini 1.5 Pro	<u>49.4</u>	**68.4**	64.8	**59.8**	<u>55.0</u>	<u>60.4</u>	**69.6**	**64.6**	<u>65.0</u>	**66.7**	**60.8**
GPT-4o	**53.9**	56.0	**81.5**	<u>56.1</u>	**59.4**	**67.6**	58.7	<u>56.8</u>	63.6	59.5	<u>60.3</u>

Table 4: TUNA-MCQ performance of representative video LMMs. We provide detailed scores for selected tested models on 10 temporal tasks. The best and second-best results are marked with **bold** and <u>underline</u>, respectively.

robust and accurate solution for captioning evaluation. More details are available in Appendix B.2.4.

4.3 Video QA

TUNA-MCQ specializes in temporal understanding in videos, emphasizing the necessity of the entire video observation rather than single-frame analysis. We assess the temporal understanding skills across 4 dynamic elements and 10 task types.

Overall Performance. Table 4 showcases the performance of selected models on TUNA-MCQ. All tested models demonstrate limited capabilities, with even the best-performing model barely achieving a passing score. However, a promising trend emerges as open-source models illustrate performance on par with commercial counterparts. Specifically, LLaVA-Video-72B and Qwen2-VL-72B achieve an identical score of 60.7%, matching the performance of GPT-4o (60.3%) and Gemini 1.5 Pro (60.8%). This competitive performance of open-source models aligns with findings from recent studies, such as Video-MME (Short) (Fu et al., 2024) and TempCompass (Liu et al., 2024f), suggesting a promising direction for open-source development in video understanding.

Camera State. Recent works (Chai et al., 2024; Tan et al., 2024) emphasize the crucial role of camera state in video understanding and generation. However, open-source video understanding datasets minimally involve this aspect. Our evaluation reveals a considerable weakness in models' camera understanding skill, with average scores notably lower than overall scores. While models show some promise in detecting camera transitions, they struggle particularly with camera motion analysis, achieving a maximum score of only 53.9%.

Subject Action. Action understanding is another challenge, as it requires tracking and interpreting character state evolutions across multiple frames. The action sequence task is notoriously difficult due to its complexity, demanding models to simultaneously recognize individual actions while understanding their temporal order and causal relationships. While GPT-4o leads performance with 67.6% accuracy, all other models struggle to the passing threshold. Additionally, temporal action recognition remains challenging, with even the best-performing model achieving only 62.4%.

Background Scene & Object Attribute. Advanced video LMMs show promising capabilities in scene and attribute understanding. For background scene tasks, models achieve impressive results with GPT-4o reaching 81.5% on scene description, while LLaVA-Video-72B attains 61.7% on scene transition understanding. In object attribute tasks, models also perform well, with top scores of 64.6% in recognition, 76.3% in appearance, and 66.7% in location tasks. These strong performance can be attributed to the transfer of knowledge from well-established image-text understanding techniques, as these tasks share similar characteristics with multi-image analysis scenarios.

These comprehensive results underscore the complex challenges in understanding temporal dynamics in videos, while offering clear directions for future improvements in video LMMs.

4.4 Synthesizing Analysis

Through comprehensive analysis of TUNA-CAP and TUNA-MCQ results, commercial models demonstrate superior performance across both tasks. While open-source models (Qwen2-VL-72B and LLaVA-Video-72B) achieve comparable results on TUNA-MCQ, they notably underperform in TUNA-CAP. This performance gap reveals a critical limitation of open-source LMMs in captioning and even open-ended QA tasks, indicating areas demanding further research efforts.

5 Conclusion

In this paper, we present TUNA-1K, a temporally dense video-caption dataset, and its derivative benchmark TUNA. Our work focuses on temporal dynamics, the distinctive feature between videos and static images, by examining four critical temporal aspects: camera, scene, action, and attribute. TUNA-1K features comprehensive coverage across diverse visual domains with detailed, fine-grained captions. TUNA evaluates LMMs' temporal understanding skills through two complementary tasks: captioning and MCQ. This comprehensive evaluation provides precise insights into models' strengths and weaknesses, offering interpretable metrics for advancing video understanding technology. We envision TUNA serving as a catalyst for future research in video understanding. Moreover, the meticulously annotated TUNA-1K, with its high accuracy and completeness, offers versatile applications beyond our current scope. We anticipate its broad utility in diverse research directions and look forward to seeing its impact on future studies in the field.

Limitations

Our dataset is highly fine-grained, but the data annotation is extremely labor-intensive. making it costly to apply this construction method to other video datasets. For TUNA-CAP, we conduct a comprehensive evaluation of the video captioning capabilities of video LMMs using an interpretable and robust approach. However, our method has certain limitations. Our scoring system focuses on the alignment with annotated visual elements. If the model outputs visual elements that fail to match the annotated events or elements, our method cannot assess their precision. Specifically, when a generated caption includes excessive irrelevant content, even if this content contains substantial hallucinatory information, our method would be unable to provide a valid assessment in such cases.

Ethics Policy

To increase the diversity of our dataset, we collected videos from several sources. These include a number of movies spanning many years and several types. While we made an effort to remove some videos that were poorly observed or NSFW, there may be unintentional data that involve potential social biases and stereotypes, including stereotypical items related to gender, race, ethnicity, age, and socioeconomic status. This requires careful judgment and utilization of the data.

References

Elmira Amirloo, Jean-Philippe Fauconnier, Christoph Roesmann, Christian Kerl, Rinu Boney, Yusu Qian, Zirui Wang, Afshin Dehghan, Yinfei Yang, Zhe Gan, et al. 2024. Understanding alignment in multimodal llms: A comprehensive study. *arXiv preprint arXiv:2407.02477*.

Hidehisa Arai, Keita Miwa, Kento Sasaki, Yu Yamaguchi, Kohei Watanabe, Shunsuke Aoki, and Issei Yamamoto. 2024. Covla: Comprehensive vision-language-action dataset for autonomous driving. *arXiv preprint arXiv:2408.10845*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. pages 65–72, Ann Arbor, Michigan.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. 2024. Video generation models as world simulators.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The revolution of multimodal large language models: A survey. pages 13590–13618, Bangkok, Thailand and virtual meeting.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, et al. 2024. Temporal-bench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*.

Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jeng-Neng Hwang, Saining Xie, and Christopher D Manning. 2024. Auroracap: Efficient, performant video detailed captioning and a new benchmark. *arXiv preprint arXiv:2410.03051*.

David Chan, Suzanne Petryk, Joseph Gonzalez, Trevor Darrell, and John Canny. 2023. CLAIR: Evaluating image captions with large language models. pages 13638–13646, Singapore.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. 2024a. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Andrei de Souza Inácio and Heitor Silvério Lopes. 2023. Evaluation metrics for video captioning: A survey. *Machine Learning with Applications*, 13:100488.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. 2024. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37:89098–89124.

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

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. 2024. Miradata: A large-scale video dataset with long durations and structured captions. *arXiv preprint arXiv:2407.06358*.

Fanheng Kong, Jingyuan Zhang, Yahui Liu, Hongzhi Zhang, Shi Feng, Xiaocui Yang, Daling Wang, Yu Tian, Victoria W, Fuzheng Zhang, and Guorui Zhou. 2025. Modality curation: Building universal embeddings for advanced multimodal information retrieval. *arXiv preprint arXiv:2505.19650*.

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

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, Jianfeng Gao, et al. 2024b. Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision*, 16(1-2):1–214.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024c. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

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

Yunxin Li, Xinyu Chen, Baotian Hu, Longyue Wang, Haoyuan Shi, and Min Zhang. 2024e. Videovista: A versatile benchmark for video understanding and reasoning. *arXiv preprint arXiv:2406.11303*.

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

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. pages 74–81, Barcelona, Spain.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pretraining for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. 2024c. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*.

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. 2025. Llava-plus: Learning to

use tools for creating multimodal agents. In *European Conference on Computer Vision*, pages 126–142. Springer.

Tingkai Liu, Yunzhe Tao, Haogeng Liu, Qihang Fang, Ding Zhou, Huaibo Huang, Ran He, and Hongxia Yang. 2024d. DeVAn: Dense video annotation for video-language models. pages 14305–14321, Bangkok, Thailand.

Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Chang Wen Chen, and Ying Shan. 2024e. E.t. bench: Towards open-ended event-level video-language understanding. In *Neural Information Processing Systems (NeurIPS)*.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024f. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

Neelu Madan, Andreas Møgelmose, Rajat Modi, Yogesh S Rawat, and Thomas B Moeslund. 2024. Foundation models for video understanding: A survey. *arXiv preprint arXiv:2405.03770*.

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

mixkit. 2023. mixkit. https://mixkit.com/videos/.

Thong Nguyen, Yi Bin, Junbin Xiao, Leigang Qu, Yicong Li, Jay Zhangjie Wu, Cong-Duy Nguyen, See-Kiong Ng, and Luu Anh Tuan. 2024. Video-language understanding: A survey from model architecture, model training, and data perspectives. *arXiv preprint arXiv:2406.05615*.

OpenAI. 2024. Hello gpt-4o.

Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. 2023. Kosmos-g: Generating images in context with multimodal large language models. *arXiv preprint arXiv:2310.02992*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318, Philadelphia, Pennsylvania, USA.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. 2024. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36.

Pexels. 2023. Pexels. https://www.pexels.com/videos/.

pixabay. 2023. pixabay. https://pixabay.com/videos/.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. 2024. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Darshana Saravanan, Darshan Singh, Varun Gupta, Zeeshan Khan, Vineet Gandhi, and Makarand Tapaswi. 2024. Velociti: Can video-language models bind semantic concepts through time? *arXiv preprint arXiv:2406.10889*.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. 2024. Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models. *arXiv preprint arXiv:2410.23266*.

Zhiyu Tan, Xiaomeng Yang, Luozheng Qin, and Hao Li. 2024. Vidgen-1m: A large-scale dataset for text-to-video generation. *arXiv preprint arXiv:2408.02629*.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. 2023. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Jiawei Wang, Liping Yuan, and Yuchen Zhang. 2024a. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*.

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

Yuxuan Wang, Difei Gao, Licheng Yu, Weixian Lei, Matt Feiszli, and Mike Zheng Shou. 2022. Geb+: A benchmark for generic event boundary captioning, grounding and retrieval. In *European Conference on Computer Vision*, pages 709–725. Springer.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786.

Tianwei Xiong, Yuqing Wang, Daquan Zhou, Zhijie Lin, Jiashi Feng, and Xihui Liu. 2024. Lvd-2m: A long-take video dataset with temporally dense captions. *arXiv preprint arXiv:2410.10816*.

Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. 2024. Pllava: Parameter-free llava extension from images to videos for video dense captioning. *arXiv preprint arXiv:2404.16994*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. 2024a. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Jingyuan Zhang, Hongzhi Zhang, Zhou Haonan, Chenxi Sun, Jiakang Wang, Fanheng Kong, Yahui Liu, Qi Wang, Fuzheng Zhang, et al. 2025. Data metabolism: An efficient data design schema for vision language model. *arXiv preprint arXiv:2504.12316*.

Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. 2024b. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Hao-ran Tan, Chunyuan Li, and Ziwei Liu. 2024c. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*.

Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, et al. 2024d. Direct preference optimization of video large multi-modal models from language model reward. *arXiv preprint arXiv:2404.01258*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yiqun Zhang, Fanheng Kong, Peidong Wang, Shuang Sun, SWangLing SWangLing, Shi Feng, Daling Wang, Yifei Zhang, and Kaisong Song. 2024e. STICKERCONV: Generating multimodal empathetic responses from scratch. pages 7707–7733, Bangkok, Thailand.

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

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multi-modal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

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

Luowei Zhou, Chenliang Xu, and Jason Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.
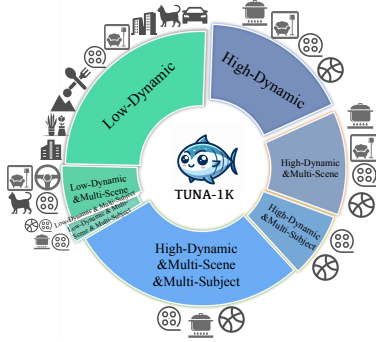
A TUNA-1K

A.1 Statistics

Figure 7: The sample distribution of TUNA-1K, videos covering 4 visual characteristics and 12 domains.

As shown in Table 5, we illustrate the detailed statistics of TUNA-1K. Each video must belong to one of Low-Dynamic or High-Dynamic categories, while Multi-Scene and Multi-Subject are optional.
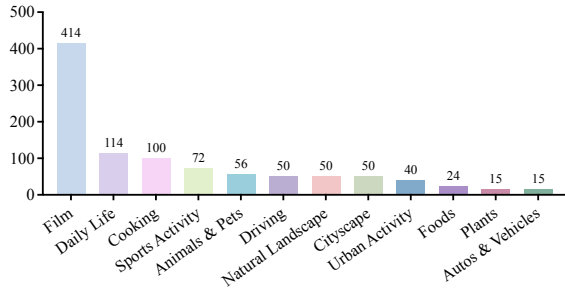

Figure 8: Sample distribution of domains in the TUNA-1K, covering 12 domains.

There are 12 domains contained in TUNA-1K, including: (1) Film, (2) Daily Life, (3) Cooking, (4) Sports Activity, (5) Driving, (6) Animals & Pets, (7) Natural Landscape, (8) Cityscape, (9) Urban Activity, (10) Foods, (11) Plants, and (12) Autos & Vehicles. As shown in Figure 8, we illustrate the domain statistics of the videos in TUNA-1K.

We visualize the sample distribution of video complexity in TUNA-1K in Figure 9, in terms of (a) the number of events, (b) the number of visual elements in each video, and (c) the number of visual elements in each event.

A.2 More Details of TUNA-1K Construction

A.2.1 Video Collection

Table 6 shows the video sources that make up the TUNA-1K, along with their descriptions.

A.2.2 Annotators

We employ crowdsourcing for data annotation. All annotators have TEM-4 or TEM-8 English proficiency, and have experience in video captioning annotation (e.g., several annotators have previously annotated video-caption pairs for Kling[1] project). Prior to formal annotation, they undergo our specialized training to guarantee the quality of their annotation.

A.2.3 Annotator Training

We prepare a detailed note document for instructing human annotators on annotation. The documentation encompasses 5 key components: (1) Visual Characteristic Classification, (2) Video Element Guidelines, (3) Video Captioning Protocol, (4) Event Splitting and Element Extraction Criterion, and (5) Annotation Examples.

Visual Characteristic Classification. Detailed criteria for categorizing videos based on their visual characteristics.

- **Low/High-Dynamic:** Based on the number and frequency of dynamic elements in the video.
- **Multi-Scene:** Presence of at least one camera transition or scene transition. Excludes those that just have camera zooming, panning, or rotating.
- **Multi-Subject:** Presence of at least two subjects. Non-major objects are not counted.

Video Element Guidelines. Comprehensive definitions and key considerations for essential video elements:

- **Camera:** Camera states, including panning, rotating, zooming, following, shaking, transition, etc. It is necessary to indicate a specific direction.
- **Scene:** Describe the background scene, including environment, weather, time, etc.
- **Action:** Recognize actions and their temporal evolving sequences.
- **Attribute:** Identify objects and describe their appearance (e.g., characters' gender, age, and dress, objects' color, shape, and number) and spatial orientation (location and relative positional relationships).

Video Captioning Protocol. We emphasizing:

- Strict chronological ordering of events.
- Objective descriptions without summarization and subjective feelings.
- If multiple similar characters/objects appear, distinguish them in expression by unique attributes (e.g., age, dress, etc.).

---

[1] https://kling.kuaishou.com

	Low-Dynamic	High-Dynamic	Multi-Scene	Multi-Subject	Total
#Videos	340	660	493	385	1,000
Duration	18.0s	12.8s	12.5s	9.5s	14.53s
#Events	2.8	3.4	3.8	3.8	3.2
#Elements (Narrative-level)	15.8	18.3	19.9	20.2	17.5
#Elements (Event-level)	5.7	5.4	5.3	5.3	5.48
#Tokens	198.8	247.1	255.9	267.6	230.7

Table 5: Detailed statistics for TUNA-1K, including: number of videos (**#Videos**), video duration (**Duration**), number of events (**#Events**), number of visual elements in captions (**#Elements (Narrative-level)**), number of visual elements in events (**#Elements (Narrative-level)**), number of tokens of caption (**#Tokens**).
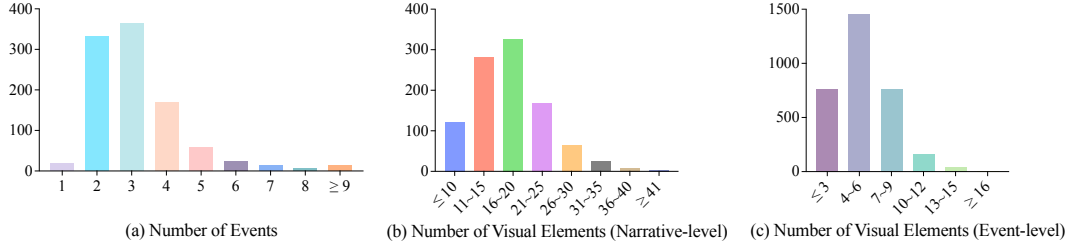

Figure 9: Visual statistics of the number of events and the number of visual elements in TUNA-1K.

Event Splitting and Element Extraction Criterion. To ensure systematic and standardized annotation, we establish the following comprehensive guidelines:

- Divide captions into chronologically ordered events, where each event represents distinct temporal activities. Further decompose each event into its constituent visual elements.
- Ensure explicit subject identification in all visual elements. Replace missing subjects and pronouns with their corresponding specific noun references to maintain clarity and precision.
- Element weighting criteria for scoring: (1) Weight 3: primary and conspicuous contents in the video. (2) Weight 2: primary and inconspicuous contents, or secondary but conspicuous contents. (3) Weight 1: secondary and inconspicuous contents.

Annotation Examples. Some complete annotation examples provide human annotators with a further guidance for annotation.

To ensure annotation quality and consistency, we implemented a rigorous annotator selection and training process. Initially, all potential annotators underwent a trial annotation phase using a shared subset of videos. This phase served both as a training exercise and a qualification assessment. Through careful evaluation of their trial annotations, we selected only those annotators who demonstrated high consistency, accuracy, and thorough understanding of the annotation guidelines. These qualified annotators then proceeded to partic-

ipate in the main annotation task. This systematic approach helped maintain annotation quality while minimizing potential inconsistencies across different annotators.

A.2.4 Annotation

Video Filter. We first filter out undesired videos based on specific rules, e.g., videos with low resolution (not satisfying 480p) and long duration (>40s). Then, human annotators filter out near-static or NSFW videos to ensure the high quality and temporal dynamics of the selected videos.

Video Cluster. Firstly, we assign a caption for each video. If the original source provides a caption, it is utilized; otherwise, a caption is generated using `gpt-4o-2024-05-13`. Then, we utilize GPT-4o to classify visual characteristic category and domain for each video. Thus far, we have obtained raw videos with initial model-generated visual characteristics and domains. Initially, we obtain raw videos with model-generated visual characteristics and domains. Annotators then observe the videos, correcting and supplementing the visual characteristic categories and domains as needed. The prompt instruction used in this step is shown in Figure 23.

Temporally Dense Caption Annotation. Annotators are tasked with providing a detailed chronological description of each video. They must divide the caption into multiple events based on criteria such as camera transitions, scene transitions, or story advancements. Each event is further split into multiple atomic visual elements, categorized
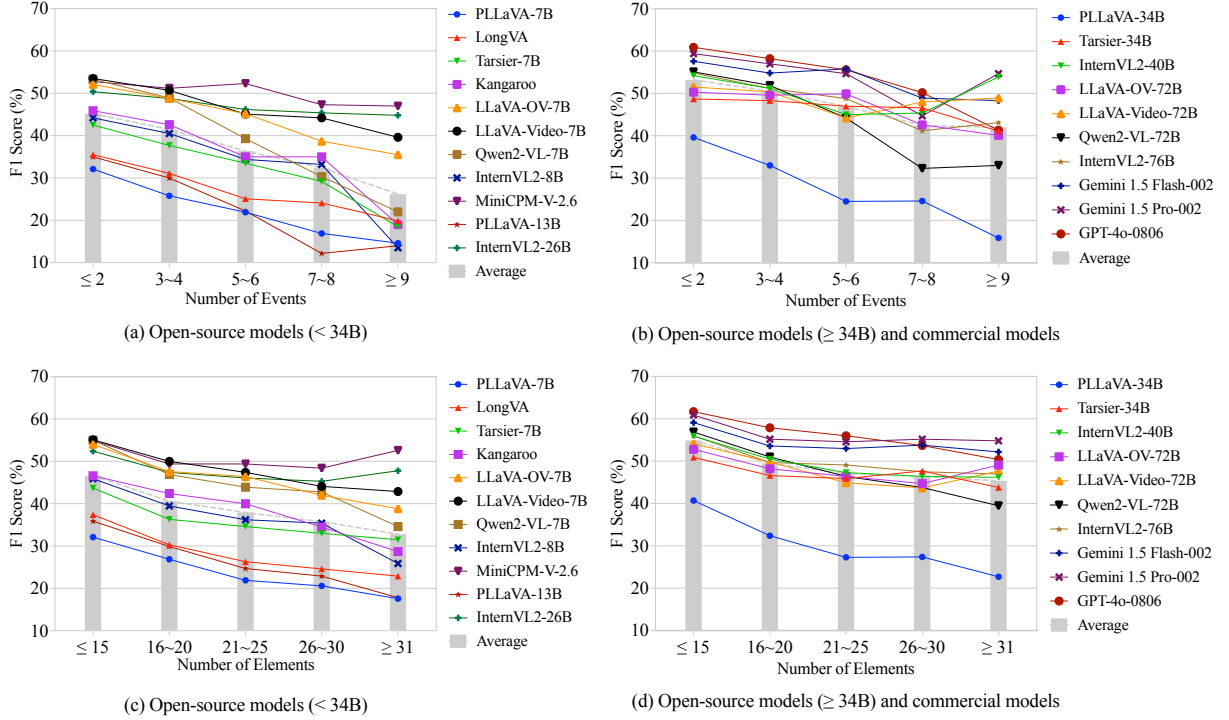
Figure 10: The whole performance comparison on TUNA-CAP across different video complexities.

by type and weighted by importance on a scale of 1-3. The types include *camera*, *scene*, *action*, and *attribute*.

Quality Review. For quality assurance, cross-inspections are performed between annotators. Furthermore, trained video experts (non-authors) continuously review the annotations, offering feedback and prompting annotators to refine their work to ensure the high-quality annotations. During cross-inspections and expert reviews, the checking covers all annotation results including video caption, event splitting and visual element extraction as well as the type and weight of the elements.

A.2.5 Visualized Examples

A detailed example in TUNA-1K is shown in Figure 15.

B TUNA-CAP

B.1 Experimental Settings

The configuration and experimental settings for all test models are shown in Table 7.

The specific version of the closed-source models we tested are `gemini-1.5-flash-002`, `gemini-1.5-pro-002`, and `gpt-4o-2024-08-06`. Incidentally, a few samples (less than 5) in our TUNA-CAP and TUNA-MCQ do not receive any responses from the Gemini (Reid et al., 2024) series, possibly due to security mechanisms. Therefore,

we calculated the scores using only the samples with responses, rather than assigning a score of 0 to those without responses.

Input Frames. By default, we uniformly sample 32 frames from each video, which is sufficient to capture the entire content of the video in our TUNA. For Qwen2-VL (Wang et al., 2024b) and PLLaVA (Xu et al., 2024), the official strategy is followed to sample frames at 2 FPS and uniformly sample 16 frames, respectively. For closed-source models, we sampled frames dynamically with 1/2 FPS, meaning that when the video event is less than 16s, it is sampled at 2 FPS, otherwise at 1 FPS.

Detailed Prompts. The default prompt template for captioning is shown in Figure 19. Figures 20, 21, and 22 illustrate the prompt templates used to evaluate TUNA-CAP.

B.2 More Experimental Analysis

TUNA-CAP results of all tested models are shown in Table 8 and Table 9, as a complement result to Table 2.

B.2.1 Video Complexity

We partition the video complexity according to the number of events and the number of visual elements in the video, to observe the impact of the model on increasing video complexity. The visualization results of selected models are shown in Figure 6. The detailed results of all tested models

are in Table 10, and its visualization results are shown in Figure 10.

As demonstrated in Table 10 and Figure 10, model performance consistently declines with increasing video complexity. Larger models ($\geq$34B parameters) exhibit better robustness to complex videos, showing smaller performance drops (2.8% for event count, 2.5% for element count) compared to their smaller counterparts (<34B parameters), which experience steeper declines (4.7% and 3.5% respectively). Moreover, the performance gap between large and small models becomes more pronounced in highly complex videos. When event count exceeds 9 (from 7~8 events), small models suffer a substantial 6.2% performance drop, while large models remain stable with only a 0.7% variation. Similarly, for videos with more than 31 elements (increased from 26~30), small models show a 3.0% fluctuation compared to just 0.7% for large models. This evidence strongly suggests that larger models possess superior adaptability to complex video content.

B.2.2 Enrichment of Visual Inputs

The number of input frames is crucial for video understanding, as it directly impacts whether the model receives sufficient visual content. This is particularly important in long-video scenarios, where the model's ability to answer a question depends on whether the sampled frames contain the necessary visual information. Limited by the number of input frames in existing LMMs, our TUNA-1K ensures that 32 frames are sufficient to cover the content of each video, considering that TUNA-1K has an average duration of $15s$ and a maximum duration of $38s$. To explore the effect of frame number on performance, we compare the TUNA-CAP performance with different input frame numbers across several classical models.
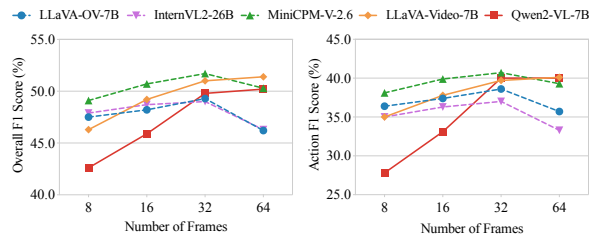

Figure 11: Performance comparison across different number of input frames.

As shown in Figure 11, increasing the number of frames generally improves the F1 score, with an average increase of 1.86% from 8 to 16 frames and

1.62% from 16 to 32 frames. This underscores the importance of providing sufficient visual information, especially when the frame count is low. Similar pattern is shown in action perception, with average improvements of 3.48% from 8 to 16 frames and 2.48% from 16 to 32 frames, indicating that dynamic actions are more sensitive to frame numbers. However, we observe a performance drop in some earlier models, including LLaVA-OV-7B (Li et al., 2024a), InternVL2-26B (Chen et al., 2024b), MiniCPM-V-2.6 (Yao et al., 2024), when the frame number is increased to 64. We attribute this decline to the fact that these earlier models rarely involve 64 frames of input (context length over 8K) during training, leading to poorer performance at 64 frames. In contrast, LLaVA-Video-7B (Zhang et al., 2024f) and Qwen2-VL-7B (Wang et al., 2024b), which are trained on longer contexts, achieve better results when the number of frames reaches 64. This indicates that providing more frames can indeed enhance performance when the context length is not constrained. More frames can improve the ability to capture intricate temporal dynamics and rich contextual information in videos. Consequently, exploring how to efficiently utilize more frames for training will emerge as a pivotal topic in the field of multimodal video understanding.

To further explore the effect of frame numbers on video understanding, we select LLaVA-Video and Qwen2-VL, which are trained with longer contexts, to illustrate the performance disparity across different video complexities with varying input frame numbers. Figure 5 presents the visualized results, while Table 12 provides the corresponding specific scores. These results demonstrate that increasing the number of frames is more beneficial for understanding more complex videos. However, excessive complexity can lead to performance anomalies, indicating that understanding highly complex videos remains a prominent challenge.

B.2.3 Scaling Law

As shown in Table 2, there is a general law that the performance of models increases as the model scale increases. Therefore, the scaling law is equally valid for video captioning task. Larger models typically have more parameters, enabling them to capture more complex patterns and nuances in the data, leading to improved performance. However, we notice that the LLaVA-Video series shows inconsistent performance scaling with model size. This anomaly may be attributed to the Slow-Fast

approach used in LLaVA-Video-72B, which results in 2/3 of the visual tokens being compressed to 1/4 of the others. This compression leads to a extensive loss of fine-grained information, which is crucial for detailed video understanding and accurate captioning. This observation suggests that the efficient usage of visual information is essential and may even outweigh the impact yielded by the language model scale. The quality and richness of the visual tokens play a critical role in the overall performance of video captioning models.

Discussion. This observation has sparked an intriguing discussion in the field: video LMMs demonstrate superior performance when processing a higher number of input frames. While increased frame coverage provides a more comprehensive representation of video content, capturing nuanced details and temporal dynamics, this advantage is constrained by context length limitations. Specifically, accommodating more frames typically involves the compression of visual tokens, a process that remains a key technical challenge. Future research should focus on the development of more efficient visual token compression techniques and the innovation in architectural designs that can handle extended context lengths, to unlock the full potential of large-scale models in video understanding tasks.

B.2.4 Correlation with Human Judgments

Given a video-caption pair, this task is to check whether the metric is consistent with human scoring. Specifically, we randomly sample 40 videos containing 687 visual elements. We provided human scorers with reference meta-information and model-generated captions. The scorers were asked to sequentially determine whether each reference visual element appeared accurately and completely in the candidate captions in the correct temporal order, ultimately resulting in human-assigned scores. Finally, we calculate Kendall's τ, Spearman's ρ, and Pearson r to test the consistency of TUNA-CAP's automatic evaluation method with human scoring. The calculated Kendall's τ, Spearman's ρ, and Pearson r are 57.2%, 76.7%, and 69.9%, respectively, with all p-values < 0.05, demonstrating the validity of our automatic evaluation method. CLAIR (Chan et al., 2023), an evaluation method for image captioning, is an LLM-based strategy for scoring based on reference captions. We migrate this approach seamlessly to assess video captioning as a comparative object. DREAM-1K (Wang et al.,

2024a) is a recently proposed method for video captioning evaluation with interpretability. However, it only focuses on subject actions, leading to its weak performance on our comprehensive video captioning data that focuses on camera, scene, action, and attributes.

C TUNA-MCQ

C.1 Statistics
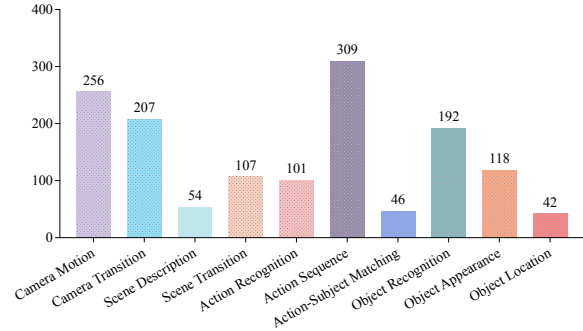
Figure 12: Sample distribution of task types in the TUNA-MCQ, covering 10 task types.

Figure 12 illustrates the sample distribution of the TUNA-MCQ, across 10 tasks: (1) camera motion, e.g, zooming, panning, and rotating. (2) camera transition. (3) scene description. (4) scene transition. (5) action recognition. (6) action sequence. (7) action-subject matching. (8) object recognition. (9) object appearance, e.g., gender, age, dress, color, shape, number. and (10) object location.
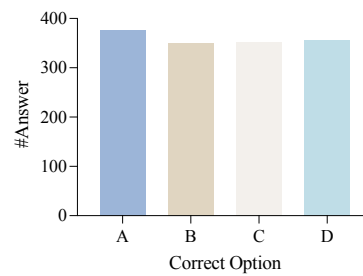
Figure 13: Sample distribution of correct option in the TUNA-MCQ.

To eliminate the bias and varied sensitivity of the models towards order and token, we ensure that the distribution of correct options is uniform, as shown in Figure 13.

C.2 More Details of TUNA-MCQ Construction

Error-prone Points Extraction. To obtain challenging questions, we obtain some error-prone points through an automated approach. Specifically, we provide the video LMM with 8 frames

from the video and its ground-truth textual description, and ask it to generate what it thinks it sees that the video is inconsistent with the textual description. The prompt instruction used in this step is shown in Figure 25.

Multi-Choice QA Generation. Based on a predefined set of task types, error-prone points and textual descriptions, LLM generates several multi-choice QAs for each video. The prompt instruction used in this step is shown in Figure 26.

Quality Review. To ensure that data is high-quality and time-sensitive, we employ crowdsourcing to optimize the automatically generated data. In addition, human annotators perform cross-inspections to ensure quality. To guarantee that the questions are relevant to capture temporal dynamics, we employ LLaVA-Video-7B to filter them. A question is deemed temporal-indispensable if it can be accurately answered using both a single frame and multiple frames. Specifically, we deem the question to be temporal-indispensable if it can be answered correctly by both 1-frame and 16-frame inputs.

C.2.1 Visualized Examples

Several examples in TUNA-MCQ are shown in Figure 16, 17, and 18.

C.3 Experimental Settings

The number of input frames in TUNA-MCQ is consistent with TUNA-CAP, which is shown in Table 7. The default prompt template for multi-choice QA is shown in Figure 24.

Incidentally, a few samples (less than 10) in our TUNA-MCQ do not receive any responses from the Gemini series, possibly due to security mechanisms. Therefore, we calculated the scores using only the samples with responses, rather than assigning a score of 0 to those without responses.

C.4 More Experimental Analysis

TUNA-MCQ results of all tested models are shown in Table 13, as a complement result to Table 4.

C.4.1 Scaling Law

On TUNA-MCQ, while most models demonstrate predictable scaling patterns, InternVL2 (Chen et al., 2024b) exhibits an unexpected trend where its 76B variant underperforms the 40B version and its 26B variant underperforms the 8B version. This anomaly is consistently observed across multiple video comprehension benchmarks: Video-MME (76B: 64.7% vs. 40B: 66.1%), MVBench (76B:

69.6% vs. 40B: 72.0%), MMBench-Video (76B: 1.71% vs. 40B: 1.78%), MLVU (76B: 69.9% vs. 40B: 71.0%). Notably, this counter-intuitive scaling behavior can be attributed to architectural differences: each InternVL2 variant employs distinct LLM backbone families and vision encoders, making direct performance comparisons less meaningful for establishing scaling laws.

D Future Work

Considering that different models have diverse capabilities in following complex instructions, we deliberately adopted simple prompting templates to ensure fair comparison and clear assessment. While this approach helps isolate models' inherent temporal understanding abilities, advanced prompting strategies like Multimodal-CoT (Zhang et al., 2023) reasoning show promising potential for performance enhancement. Although such sophisticated prompting techniques may improve performance on TUNA-MCQ, their applicability to captioning tasks like TUNA-CAP remains challenging. We encourage future research to explore advanced prompting strategies that can effectively enhance temporal understanding across different tasks while maintaining a balance between performance optimization and the assessment of fundamental temporal comprehension abilities.

E More Related Work

Video LMMs. Large Mulitmodal Models (LMMs) have mushroomed, showcasing impressive visual understanding capabilities (Li et al., 2024b; Zhang et al., 2024a; Caffagni et al., 2024; Amirloo et al., 2024; Zhang et al., 2025). These advances have catalyzed the development of diverse and innovative applications across multiple domains. (Pan et al., 2023; Zhang et al., 2024e; Liu et al., 2025; Kong et al., 2025). Existing works bridge visual encoders and Large Language Models (LLMs) using a small intermediate architecture, as seen in models like LLaVA (Liu et al., 2024b,a), BLIP-2 (Li et al., 2023), and MiniGPT-4 (Zhu et al., 2023), which facilitate the evolution of visual-language LMMs. On this basis, recent researches (Li et al., 2024c; Zhang et al., 2024b; Lin et al., 2024; Cheng et al., 2024; Lin et al., 2023; Maaz et al., 2023) have extended these techniques from static images to dynamic videos, demonstrating promising results in video understanding by processing videos as multiple image frames.

Type	Source	Domain	Visual Characteristic	Description
Web Data	Pexels (Pexels, 2023)	Animals & Pets / Autos & Vehicles / Cityscape / Foods / Natural Landscape / Urban Activity	**Low-Dynamic**	A website offer stock videos and motion graphics free from copyright issues, which are usually exceptionally high-quality videos uploaded by skilled photographers. We sample 46 videos, as a source of Low-Dynamic scenarios, covering diverse domains.
	Pixabay (pixabay, 2023)	Animals & Pets / Cityscape / Foods / Natural Landscape / Urban Activity	**Low-Dynamic**	A website offer stock videos and motion graphics free from copyright issues, which are usually exceptionally high-quality videos uploaded by skilled photographers. We sample 13 videos, as a source of Low-Dynamic scenarios, covering diverse domains.
	MixKit (mixkit, 2023)	Natural Landscape	**Low-Dynamic**	A website offer stock videos and motion graphics free from copyright issues, which are usually exceptionally high-quality videos uploaded by skilled photographers. We sample 7 videos, as a source of Low-Dynamic scenarios.
Academic Video Understanding Data	DREAM-1K (Wang et al., 2024a)	Film	Low-Dynamic / **High-Dynamic** / **Multi-Scene** / **Multi-Subject**	DREAM-1K consists of 1,000 video clips from five categories: live-action movies, animated movies, stock videos, YouTube videos, and TikTok-style short videos. These videos typically feature multiple events and subjects across various shots. We sample 148 videos from live-action movies that meet our selection principles, mostly as a source of High-Dynamic, Multi-Scene, and Multi-Subject scenarios.
	VELOCITI (Saravanan et al., 2024)	Film	Low-Dynamic / **High-Dynamic** / **Multi-Scene** / **Multi-Subject**	A benchmark using complex movie clips and dense semantic role label annotations to test perception and binding in video LMMs. The videos feature challenging scenarios with frequent shot changes, fast action sequences, multi-event situations, role switching, and entity co-referencing over time. We sample 266 videos, mostly as a source of High-Dynamic, Multi-Scene, and Multi-Subject scenarios.
	PerceptionTest (Patraucean et al., 2024)	Daily Life (Indoor)	Low-Dynamic / **High-Dynamic** / Multi-Scene	A dataset evaluates performance across skill areas (memory, abstraction, physics, semantics) and reasoning types (descriptive, explanatory, predictive, counterfactual). We sample 114 videos, mostly as a source of High-Dynamic scenarios.
	YouCook2 (Zhou et al., 2018)	Cooking	**High-Dynamic** / Multi-Scene / Multi-Subject	A dataset of YouTube videos covering 89 recipes from four major cuisines (Africa, Americas, Asia, Europe), featuring diverse cooking styles and challenges like fast camera motion, camera zooms, video defocus, and scene-type changes. We sample 100 videos, mostly as a source of High-Dynamic scenarios.
Academic Video Generation Data	VIDGEN-1M (Tan et al., 2024)	Animals & Pets / Autos & Vehicles / Cityscape / Foods / Natural Landscape / Plants / Urban Activity / Sports Activity	**Low-Dynamic** / **High-Dynamic** / **Multi-Scene**	Open-domain Text-to-Video dataset with high video quality, high temporal consistency, and balanced categories. We sample 154 videos, as a source of High-Dynamic, Multi-Scene (Sports Activity) scenarios, and Low-Dynamic (other domains) scenarios.
	MiraData (Ju et al., 2024)	Animals & Pets / Autos & Vehicles / Cityscape / Foods / Natural Landscape / Plants / Urban Activity	**Low-Dynamic** / Multi-Scene	A large-scale, high-quality video dataset designed to meet the key expectations of video generation tasks: diverse content, high visual quality, long duration, and significant motion strength. Unlike existing text-to-video datasets that primarily source videos from YouTube, MiraData includes videos from YouTube, Videvo, Pixabay, and Pexels, ensuring a more comprehensive and suitable data source. We sample 102 videos, mostly as a source of Low-Dynamic scenarios, covering diverse domains.
Others	CoVLA (Arai et al., 2024)	Driving	Low-Dynamic / **Multi-Scene** / Multi-Subject	The CoVLA (Comprehensive Vision-Language-Action) dataset is a novel large-scale resource designed to advance autonomous driving research. The dataset includes synchronized multi-modal data streams from front-facing cameras, in-vehicle signals, and other sensors, providing a comprehensive view of diverse driving scenarios. We choose it due to its complex scene variations. We sample 50 videos as a source of Multi-Scene scenarios.

Table 6: Rich video sources within TUNA-1K. **Domain** denote the domains represented in the sampled data. **Visual Characteristic** indicates the visual characteristics present in the sampled data, with **bold** representing major features and grey representing minor features.. We also provide a brief description of each dataset, along with our our selection criteria and counts.

Figure 14: Several video understanding benchmark examples and analysis.

Model	LLM	Vision Model	#Frames
Open-Source LMMs			
Qwen2-VL-72B	Qwen2-72B	ViT-600M	2FPS
Qwen2-VL-7B	Qwen2-7B	ViT-600M	2FPS
LLaVA-Video-72B	Qwen2-72B	SigLIP-400M	32
LLaVA-Video-7B	Qwen2-7B	SigLIP-400M	32
LLaVA-OneVision-72B	Qwen2-72B	SigLIP-400M	32
LLaVA-OneVision-7B	Qwen2-7B	SigLIP-400M	32
InternVL2-76B	Llama-3-70B-Instruct	InternViT-6B	32
InternVL2-40B	Nous-Hermes-2-Yi-34B	InternViT-6B	32
InternVL2-26B	InternLM2-20B	InternViT-6B	32
InternVL2-8B	InternLM2.5-7B	InternViT-300M	32
Tarsier-34B	Nous-Hermes-2-Yi-34B	CLIP ViT-L/14	32
Tarsier-7B	Vicuna-v1.5-7B	CLIP ViT-L/14	32
PLLaVA-34B	Nous-Hermes-2-Yi-34B	CLIP ViT-L/14	16
PLLaVA-13B	Vicuna-v1.5-13B	CLIP ViT-L/14	16
PLLaVA-7B	Vicuna-v1.5-7B	CLIP ViT-L/14	16
MiniCPM-V-2.6	Qwen2-7B	SigLIP-400M	32
Kangaroo	Llama3-8B-Instruct	EVA-CLIP-L	32
LongVA-7B	Qwen2-7B-Instruct-224K	CLIP ViT-L/14	32
Closed-Source LMMs			
GPT-4o	Unknown	Unknown	1/2 FPS*
Gemini 1.5 Pro	Unknown	Unknown	1/2 FPS*
Gemini 1.5 Flash	Unknown	Unknown	1/2 FPS*

Table 7: The number of frames used in the TUNA evaluation in Section 4.2, 4.3. By default, 32 frames are sampled uniformly, which is enough to cover the content of each video in TUNA-CAP. Some models take a different number of frames because they are limited by the input length or according to their sampling recommendations. * indicates that 2 FPS is employed when the video duration $< 16s$, otherwise 1 FPS is employed. The versions of the closed-source models are `gpt-4o-2024-08-06`, `gemini-1.5-pro-002`, `gemini-1.5-flash-002`.

Model	Camera			Scene			Action			Attribute			Overall		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Open-Source LMMs															
PLLaVA-7B	49.4	22.6	28.9	52.2	30.9	36.6	30.5	12.6	16.5	44.5	19.5	25.3	60.0	19.1	27.4
LongVA-7B	52.3	26.0	32.5	56.5	34.4	40.6	38.9	17.2	22.0	50.6	22.0	28.4	71.6	22.3	31.8
Tarsier-7B	56.9	27.3	34.8	45.3	28.2	33.1	56.7	28.9	36.2	56.4	26.0	33.3	73.0	27.9	38.6
Kangaroo	65.2	36.5	44.1	67.8	45.4	51.9	49.3	26.0	31.9	59.8	32.2	39.5	69.5	32.5	42.7
LLaVA-OV-7B	75.2	42.0	51.0	71.8	51.2	57.6	54.1	30.4	36.8	66.2	42.0	49.3	73.6	38.6	49.3
LLaVA-Video-7B	74.0	41.5	50.4	73.6	52.3	58.9	57.0	30.8	37.8	72.1	44.8	53.1	77.0	39.7	51.0
Qwen2-VL-7B	72.3	40.7	49.0	71.9	50.0	56.7	55.9	30.1	37.0	68.2	38.4	46.7	77.8	37.6	48.9
InternVL2-8B	64.8	33.7	41.7	59.4	38.7	44.7	45.2	24.7	30.0	59.8	35.5	42.3	67.2	31.1	40.8
MiniCPM-V-2.6	76.5	47.8	56.0	75.0	54.1	60.6	57.2	31.8	38.8	68.7	42.3	50.2	76.0	40.7	51.7
PLLaVA-13B	57.0	25.8	33.0	57.3	34.0	40.3	36.2	13.8	18.5	50.0	23.3	29.8	65.0	21.4	30.6
InternVL2-26B	73.2	43.2	51.6	72.5	52.6	58.7	51.7	30.9	37.0	63.9	42.3	49.1	70.0	39.2	49.0
PLLaVA-34B	60.8	29.6	37.4	56.2	33.7	39.9	38.7	17.3	22.3	55.1	26.1	33.2	67.8	24.5	34.2
Tarsier-34B	63.6	34.3	42.3	59.0	38.4	44.4	65.6	39.9	47.6	63.6	34.3	42.2	77.1	36.7	48.2
InternVL2-40B	77.8	46.3	55.1	71.9	53.1	59.0	53.4	33.1	39.3	65.9	45.7	52.3	71.3	42.1	51.7
LLaVA-OV-72B	73.5	43.7	51.9	71.5	51.1	57.5	51.2	30.2	36.0	65.7	41.4	48.8	72.7	39.2	49.6
LLaVA-Video-72B	72.7	41.7	50.3	71.1	49.9	56.4	55.7	32.7	39.3	68.1	43.2	50.8	73.7	39.6	50.2
Qwen2-VL-72B	73.6	45.9	54.0	67.6	46.3	52.8	59.1	35.7	42.6	66.6	40.7	48.5	74.7	41.1	51.7
InternVL2-76B	75.1	45.4	53.9	73.3	55.8	61.4	55.7	34.9	41.2	64.3	44.5	50.9	70.7	42.3	51.9
Closed-Source LMMs															
Gemini 1.5 Flash	74.6	52.8	59.6	77.2	59.3	65.1	58.7	36.4	42.9	69.0	48.4	55.2	72.7	46.4	55.7
Gemini 1.5 Pro	78.7	53.0	60.7	75.7	57.4	63.3	59.0	40.3	46.3	69.0	49.4	56.0	73.7	48.1	57.4
GPT-4o	80.1	53.3	61.3	79.5	60.2	66.4	64.0	41.1	48.0	73.8	50.1	57.8	77.7	48.2	58.5

Table 8: Evaluation results in terms of dynamic element categoryies on TUNA-CAP. The best and second-best results are marked with orange and blue, respectively.

1829

Model	Low-Dynamic			High-Dynamic			Multi-Scene			Multi-Subject			Overall		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Open-Source LMMs															
PLLaVA-7B	66.5	23.0	32.7	56.6	17.1	24.7	55.7	15.5	22.8	56.2	15.3	22.5	60.0	19.1	27.4
LongVA-7B	75.9	26.5	37.3	69.4	20.1	29.0	68.3	19.0	27.6	67.3	15.7	23.7	71.6	22.3	31.8
Tarsier-7B	81.2	34.3	46.5	68.7	24.5	34.5	71.7	25.3	35.8	67.8	23.2	33.2	73.0	27.9	38.6
Kangaroo	73.2	34.7	45.6	67.6	31.3	41.1	66.2	29.7	39.3	63.5	26.3	35.7	69.5	32.5	42.7
LLaVA-OV-7B	78.6	38.4	50.0	71.0	38.8	48.9	71.7	38.3	48.4	67.1	33.8	43.8	73.6	38.6	49.3
LLaVA-Video-7B	80.7	40.0	52.2	75.1	39.5	50.3	77.1	38.6	50.0	73.5	34.6	45.8	77.0	39.7	51.0
Qwen2-VL-7B	81.2	42.0	53.8	76.0	35.3	46.4	76.8	33.2	44.4	73.6	28.9	39.9	77.8	37.6	48.9
InternVL2-8B	71.6	34.0	44.5	64.9	29.7	38.9	65.6	29.1	38.4	61.5	26.6	35.2	67.2	31.1	40.8
MiniCPM-V-2.6	79.3	41.4	53.0	74.3	40.4	51.0	76.5	40.8	51.7	73.5	38.3	49.0	76.0	40.7	51.7
PLLaVA-13B	69.8	25.7	36.0	62.5	19.1	27.8	62.3	17.6	26.0	60.3	16.3	24.3	65.0	21.4	30.6
InternVL2-26B	71.9	39.1	49.4	69.0	39.2	48.9	70.3	38.6	48.4	67.2	36.3	45.8	70.0	39.2	49.0
PLLaVA-34B	74.5	28.1	38.9	64.3	22.6	31.8	63.9	21.3	30.2	60.7	19.2	27.6	67.8	24.5	34.2
Tarsier-34B	79.6	37.2	49.1	75.8	36.5	47.8	77.6	38.1	49.6	74.4	36.0	47.3	77.1	36.7	48.2
InternVL2-40B	75.0	43.8	53.9	69.5	41.2	50.5	70.7	40.8	50.5	67.9	38.7	48.0	71.3	42.1	51.7
LLaVA-OV-72B	75.4	37.3	48.6	71.3	36.7	45.9	71.4	40.1	50.1	72.3	39.1	49.4	72.7	39.2	49.6
LLaVA-Video-72B	77.3	39.2	50.6	71.9	39.8	50.0	73.9	38.6	49.3	70.5	35.1	45.7	73.7	39.6	50.2
Qwen2-VL-72B	79.2	44.6	55.7	72.4	39.3	49.7	73.6	37.2	48.0	69.1	32.8	43.3	74.7	41.1	51.7
InternVL2-76B	72.0	43.1	52.8	70.1	41.9	51.5	71.4	41.1	51.1	68.6	39.7	49.3	70.7	42.3	51.9
Closed-Source LMMs															
Gemini 1.5 Flash	74.0	46.5	56.0	72.0	46.4	55.5	73.4	46.2	55.9	73.4	46.2	55.9	72.7	46.4	55.7
Gemini 1.5 Pro	76.7	48.7	58.7	72.1	47.8	56.7	73.4	47.7	57.0	69.9	44.1	53.3	73.7	48.1	57.4
GPT-4o	79.1	47.3	58.2	77.0	48.6	58.7	78.7	47.2	58.1	76.8	44.4	55.5	77.7	48.2	58.5

Table 9: Evaluation results in terms of visual characteristic categoryies on TUNA-CAP. The best and second-best results are marked with orange and blue , respectively.

Model	#Events					#Elements					Overall
	≤ 2	3~4	5~6	7~8	≥ 9	≤ 15	16~20	21~25	26~30	≥ 31	
Open-Source LMMs											
PLLaVA-7B	32.1	25.8	21.9	16.9	14.6	32.1	26.9	21.9	20.6	17.6	27.4
LongVA-7B	35.5	31.1	25.1	24.1	19.9	37.4	30.3	26.3	24.6	22.9	31.8
Tarsier-7B	42.5	37.7	33.5	29.2	18.5	43.7	36.3	34.6	33.0	31.5	38.6
Kangaroo	45.9	42.6	35.0	35.0	19.0	46.6	42.4	40.0	34.5	28.7	42.7
LLaVA-OV-7B	52.1	48.8	45.2	38.7	35.5	54.0	47.6	46.4	42.0	38.8	49.3
LLaVA-Video-7B	53.5	50.7	45.1	44.2	39.6	55.1	50.0	47.4	44.1	42.9	51.0
Qwen2-VL-7B	53.3	48.9	39.3	30.3	22.0	55.0	46.9	43.9	42.8	34.6	48.9
InternVL2-8B	44.2	40.5	34.4	33.2	13.5	45.9	39.5	36.2	35.4	25.9	40.8
MiniCPM-V-2.6	52.8	51.2	52.3	47.3	47.0	54.9	49.4	49.4	48.4	52.6	51.7
PLLaVA-13B	35.0	30.0	22.2	12.2	14.0	35.9	29.9	24.7	22.9	17.8	30.6
InternVL2-26B	50.4	48.8	46.2	45.4	44.8	52.4	47.4	46.1	45.3	47.8	49.0
Avg (<34B)	45.2	41.5 (-3.7)	36.4 (-5.1)	32.4 (-4.0)	26.2 (-6.2)	46.6	40.6 (-6.0)	37.9 (-2.7)	35.8 (-2.1)	32.8 (-3.0)	42.0
PLLaVA-34B	39.6	33.0	24.5	24.6	15.9	40.7	32.4	27.3	27.4	22.7	34.2
Tarsier-34B	48.7	48.3	47.0	46.6	41.1	50.9	46.6	45.9	47.7	43.8	48.2
InternVL2-40B	54.2	51.2	45.0	45.4	53.9	55.9	50.5	47.3	46.4	46.2	51.7
LLaVA-OV-72B	50.3	49.6	49.9	42.6	40.1	52.8	48.2	46.4	44.7	49.1	49.6
LLaVA-Video-72B	51.5	50.4	44.2	48.0	48.9	54.1	49.8	44.9	43.7	47.7	50.2
Qwen2-VL-72B	55.1	51.9	44.2	32.3	33.0	56.9	51.0	46.4	43.8	39.5	51.7
InternVL2-76B	54.8	51.2	48.8	41.2	43.1	56.0	49.7	49.1	47.5	47.0	51.9
Avg (≥34B)	50.6	47.9 (-2.7)	43.4 (-4.6)	40.1 (-3.3)	39.4 (-0.7)	52.5	46.9 (-5.6)	43.9 (-3.0)	43.0 (-0.9)	42.3 (-0.7)	48.2
Closed-Source LMMs											
Gemini 1.5 Flash	57.6	54.8	55.8	48.9	48.3	59.1	53.6	53.0	53.8	52.2	55.7
Gemini 1.5 Pro	59.4	57.0	54.7	44.7	54.7	60.9	55.2	54.6	55.2	54.8	57.4
GPT-4o	60.9	58.2	55.6	50.2	41.3	61.7	57.9	56.0	53.7	50.4	58.5
Avg (close-source)	59.3	56.7 (-2.6)	55.4 (-1.3)	47.9 (-7.4)	48.1 (+0.2)	60.6	55.6 (-5.0)	54.5 (-1.0)	54.2 (-0.3)	52.5 (-1.8)	57.2
Avg (Total)	49.0	45.8 (-3.2)	41.4 (-4.4)	37.2 (-4.2)	33.7 (-3.4)	50.6	44.8 (-5.7)	42.3 (-2.6)	40.8 (-1.4)	38.8 (-2.0)	46.2

Table 10: Detailed performance comparison with varying video complexities. Video complexity is measured by the number of events and the number of visual elements in the video. The inference setup is consistent with Table 7.

Model	Frames	Camera	Scene	Action	Attribute	Low-Dynamic	High-Dynamic	Multi-Scene	Multi-Subject	Overall
LLaVA-OV-7B	8	50.2	56.6	33.0	47.8	50.4	46.1	46.2	42.6	47.5
	16	49.3 (-0.9)	57.2 (+0.6)	35.6 (+2.6)	48.9 (+1.1)	50.2 (-0.2)	47.1 (+1.0)	47.0 (+0.8)	42.5 (+0.1)	48.2 (+0.7)
	32	51.0 (+1.7)	57.6 (+0.4)	36.8 (+1.2)	49.3 (+0.4)	50.0 (-0.2)	48.9 (+1.8)	48.4 (+1.4)	43.8 (+1.3)	49.3 (+1.1)
	64	47.4 (-3.6)	54.6 (-3)	33.5 (-3.3)	45.9 (-3.4)	48.8 (-1.2)	44.8 (-4.1)	44.6 (-3.8)	39.9 (-3.9)	46.2 (-3.1)
MiniCPM-V-2.6	8	56.3	59.8	33.0	47.3	52.9	47.1	48.3	44.8	49.1
	16	55.5 (-0.8)	60.5 (+0.7)	36.7 (+3.7)	47.9 (+0.6)	52.6 (-0.3)	49.7 (+2.6)	50.8 (+2.5)	48.1 (+3.3)	50.7 (+1.6)
	32	56.0 (+0.5)	60.6 (+0.1)	38.8 (+2.1)	50.2 (+2.3)	53.0 (+0.4)	51.0 (+1.3)	51.7 (+0.9)	49.0 (+0.9)	51.7 (+1.0)
	64	52.6 (-3.4)	58.2 (-2.4)	39.1 (+0.3)	48.6 (-1.6)	50.5 (-2.5)	50.3 (-0.7)	50.0 (-1.7)	46.9 (-2.1)	50.3 (-1.4)
InternVL2-26B	8	50.1	58.3	35.0	48.8	49.6	47.1	47.0	43.9	47.9
	16	50.0 (-0.1)	59.1 (+0.8)	36.3 (+1.3)	49.9 (+1.1)	49.4 (-0.2)	48.4 (+1.3)	48.3 (+1.3)	45.4 (+1.5)	48.7 (+0.8)
	32	51.6 (+1.6)	58.7 (-0.4)	37.0 (+0.7)	49.1 (-0.8)	49.4 (-)	48.9 (+0.5)	48.4 (+0.1)	45.8 (+0.4)	49.0 (+0.3)
	64	49.6 (-2)	55.1 (-3.6)	33.3 (-3.7)	46.4 (-2.7)	47.5 (-1.9)	45.7 (-3.2)	44.3 (-4.1)	42.2 (-3.6)	46.3 (-2.7)
LLaVA-Video-7B	8	49.3	55.1	31.8	46.8	49.8	44.6	44.3	41.0	46.3
	16	50.7 (+1.4)	57.0 (+1.9)	36.3 (+4.5)	49.0 (+2.2)	51.7 (+1.9)	47.9 (+3.3)	47.0 (+2.7)	43.0 (+2.0)	49.2 (+2.9)
	32	50.4 (+0.3)	58.9 (+1.9)	37.8 (+1.5)	53.1 (+4.1)	52.2 (+0.5)	50.3 (+2.4)	50.0 (+3.0)	45.8 (+2.8)	51.0 (+1.8)
	64	51.0 (+0.6)	58.7 (-0.2)	39.0 (+1.2)	52.4 (-0.7)	51.3 (-0.9)	51.4 (+1.1)	50.1 (+0.1)	46.9 (+1.1)	51.4 (+0.4)
Qwen2-VL-7B	8	44.2	55.5	27.8	41.7	49.6	39.0	37.1	33.2	42.6
	16	47.7 (+3.5)	55.9 (+0.4)	33.1 (+5.3)	43.6 (+1.9)	51.5 (+1.9)	43.0 (+4.0)	42.0 (+4.9)	36.7 (+3.5)	45.9 (+3.3)
	32	48.8 (+1.1)	57.0 (+1.1)	40.0 (+6.9)	47.1 (+3.5)	52.6 (+1.1)	48.4 (+5.4)	46.5 (+4.5)	43.0 (+6.3)	49.8 (+3.9)
	64	50.1 (+1.3)	53.1 (-3.9)	40.0 (-)	49.4 (+2.3)	53.2 (+0.6)	48.7 (+0.3)	47.1 (+0.6)	43.4 (+0.4)	50.2 (+0.4)

Table 11: Detailed performance comparison with different number of input frames. Consistent visual results in Figure 11.

Model	Frames	#Events ≤ 2	3~4	5~6	7~8	≥ 9	#Elements ≤ 15	16~20	21~25	26~30	≥ 31	Overall
LLaVA-Video-7B	8	49.4	46.3	39.6	35.2	24.9	51.7	45.3	40.9	38.6	35.0	46.3
	16	52.9 (+3.5)	48.8 (+2.5)	41.2 (+1.6)	36.7 (+1.5)	35.7 (+10.8)	54.6 (+2.9)	47.7 (+2.4)	44.0 (+3.1)	42.4 (+3.8)	38.2 (+3.2)	49.2 (+2.9)
	32	53.5 (+0.6)	50.7 (+1.9)	45.1 (+3.9)	44.2 (+7.5)	39.6 (+3.9)	55.1 (+0.5)	50.0 (+2.3)	47.4 (+3.4)	44.1 (+1.7)	42.9 (+4.7)	51.0 (+1.8)
	64	53.7 (+0.2)	51.1 (+0.4)	46.6 (+1.5)	44.2 (-)	39.4 (-0.2)	55.2 (+0.1)	50.9 (+0.9)	47.9 (+0.5)	44.7 (+0.6)	41.4 (-1.5)	51.4 (+0.4)
Qwen2-VL-7B	8	46.8	43.2	29.7	25.5	16.0	48.6	42.1	36.9	30.6	29.1	42.6
	16	50.4 (+3.6)	46.1 (+2.9)	32.9 (+3.2)	32.7 (+7.2)	14.6 (-1.4)	52.3 (+3.7)	44.9 (+2.8)	39.1 (+2.2)	36.3 (+5.7)	32.6 (+3.5)	45.9 (+3.3)
	32	53.5 (+3.1)	49.8 (+3.7)	40.0 (+7.1)	37.7 (+5)	30.6 (+16)	55.0 (+2.7)	48.5 (+3.6)	45.3 (+6.2)	42.3 (+6)	38.3 (+5.7)	49.8 (+3.9)
	64	52.7 (-0.8)	50.5 (+0.7)	42.7 (+2.7)	45.1 (+7.4)	27.6 (-3)	55.1 (+0.1)	49.8 (+1.3)	45.2 (-0.1)	42.9 (+0.6)	36.8 (-1.5)	50.2 (+0.4)

Table 12: Performance comparison across different video complexities with varying input frame numbers. Consistent visualization results in Figure 5.

Model	Camera State Motion	Transition	Background Scene Description	Transition	Subject Action Recognition	Sequence	Matching	Object Attribute Recognition	Appearance	Location	Overall
Open-Source LMMs											
PLLaVA-7B	29.7	31.9	48.1	22.4	43.6	34.6	30.4	32.3	38.1	45.2	33.7
LongVA-7B	37.5	41.5	63.0	30.8	44.6	44.7	43.5	41.7	47.6	40.5	42.4
Tarsier-7B	23.0	24.6	40.7	20.6	38.6	26.9	45.7	20.9	25.9	23.8	26.5
Kangaroo	33.2	47.3	53.7	38.3	49.5	38.8	54.3	47.2	43.5	59.5	42.9
LLaVA-OV-7B	42.2	54.6	57.4	48.6	42.6	41.4	60.9	47.9	50.0	59.5	47.4
LLaVA-Video-7B	39.1	50.7	59.3	46.7	52.5	52.4	56.5	53.6	61.9	47.6	50.6
Qwen2-VL-7B	41.0	51.7	66.7	45.8	54.5	52.8	65.2	49.0	60.2	57.1	51.3
InternVL2-8B	41.0	53.1	66.7	40.2	45.5	50.5	50.0	45.8	56.8	45.2	48.4
MiniCPM-V-2.6	39.8	45.9	59.3	34.6	49.5	51.1	52.2	42.2	46.6	50.0	45.7
PLLaVA-13B	31.2	31.9	46.3	23.4	48.5	41.1	45.7	37.0	41.5	45.2	37.2
InternVL2-26B	38.7	45.4	63.0	42.1	48.5	46.0	58.7	42.7	55.1	50.0	45.9
PLLaVA-34B	42.6	41.5	63.0	43.9	45.5	48.5	56.5	43.2	56.8	57.1	46.9
Tarsier-34B	43.0	48.3	72.2	45.8	51.5	50.2	56.5	49.7	53.7	61.9	50.1
InternVL2-40B	40.2	58.0	74.1	51.4	56.4	53.4	63.0	57.3	66.9	61.9	54.7
LLaVA-OV-72B	46.5	67.6	75.9	57.0	59.4	56.6	73.9	63.5	69.5	59.5	60.0
LLaVA-Video-72B	47.7	67.6	77.8	61.7	61.4	57.0	65.2	62.5	73.7	57.1	60.7
Qwen2-VL-72B	52.7	64.7	74.1	55.1	62.4	54.4	67.4	63.0	76.3	66.7	60.7
InternVL2-76B	43.8	61.8	74.1	43.0	50.5	50.5	54.3	52.1	66.1	57.1	53.1
Closed-Source LMMs											
Gemini 1.5 Flash	40.8	58.3	70.4	52.3	48.0	54.2	63.0	49.0	66.7	64.3	53.3
Gemini 1.5 Pro	49.4	68.4	64.8	59.8	55.0	60.4	69.6	64.6	65.0	66.7	60.8
GPT-4o	53.9	56.0	81.5	56.1	59.4	67.6	58.7	56.8	63.6	59.5	60.3

Table 13: TUNA-MCQ performance of all tested video LMMs. We provide detailed scores on 10 temporal-dynamic tasks. The best and second-best results are marked with orange and blue, respectively.

Video Caption

The video begins with the camera focused on a wooden decorative piece, behind which a man is watching through it.

Then, the camera cuts to an outdoor scene with blurred edges and a clear center. A white news van with the logo "KXBD 6 News at 6" is visible by the roadside. Next to the van is a set-up camera, and a military green vehicle passes in front of the lens. In the background, greenery and a pedestrian path are visible. A woman with a bag on her right shoulder and a bag in her left hand walks along the sidewalk. The camera moves to the right, where a person is standing by the front passenger door of the news van, making a phone call.

Next, the camera cuts back indoors, where a man in a black suit suddenly turns to look inside. Behind him is an ornately decorated wall. The man in the suit turns again to look outside, then steps back while closing the door in front of him. He then turns and walks further into the room.

Events & Visual Elements

Event	The video begins with the camera focused on a wooden decorative piece, behind which a man is watching through it.		
Visual Elements	The video begins with the camera focused on a wooden decorative piece.	camera	3
	Behind the wooden decoration, there is a man.	attribute	3
	The man looks outside through the wooden decoration.	action	3

Event	Then, the camera cuts to an outdoor scene with blurred edges and a clear center. A white news van with the logo "KXBD 6 News at 6" is visible by the roadside. Next to the van is a set-up camera, and a military green vehicle passes in front of the lens. In the background, greenery and a pedestrian path are visible. A woman with a bag on her right shoulder and a bag in her left hand walks along the sidewalk. The camera moves to the right, where a person is standing by the front passenger door of the news van, making a phone call.		
Visual Elements	Then, the camera cuts to an outdoor scene.	camera	3
	The edges of the frame are blurred, with a clear center.	attribute	3
	A white news van is visible by the roadside.	attribute	3
	The van has the logo "KXBD 6 News at 6" on its side.	attribute	2
	A set-up camera stands beside the van.	scene	2
	A military green vehicle passes in front of the lens.	action	3
	Greenery and a pedestrian walkway are visible in the background.	scene	2
	A woman with a bag on her right shoulder and another in her left hand walks along the sidewalk.	action	2
	The camera moves to the right.	camera	3
	A person is standing outside the front passenger door of the news van.	attribute	2
	The person is making a phone call.	action	2

Event	Next, the camera cuts back indoors, where a man in a black suit suddenly turns to look inside. Behind him is an ornately decorated wall. The man in the suit turns again to look outside, then steps back while closing the door in front of him. He then turns and walks further into the room.		
Visual Elements	Next, the camera cuts back to the interior scene.	camera	3
	A man in a black suit suddenly turns around, looking inside.	attribute	3
	Behind the man in the suit is an ornately decorated wall.	attribute	2
	The man in the suit turns again to look outside.	action	3
	The man then steps back while closing the door in front of him.	action	3
	The man in the suit turns and walks further into the room.	action	3

Figure 15: A detailed example in TUNA-1K.

Task Type: Camera Motion

What is the order of camera movements throughout the video?
A. stationary, follows the woman to the left   B. follows the woman to the right, stationary
C. follows the woman to the left, stationary   D. stationary, follows the woman to the right
Answer: D

Task Type: Camera Transition

What is the order of camera transitions in the video?
A. man -> woman and girl -> woman -> man   B. man -> woman -> woman and girl -> man
C. man -> woman -> man -> woman and girl   D. woman -> man -> woman and girl -> man
Answer: C

Task Type: Scene Description

What is the order of focus changes in the video?
A. blurry green background, leaves of the plant
B. blurry green background, flowers of the plant
C. leaves of the plant, flowers of the plant
D. leaves of the plant, blurry green background
Answer: A

Task Type: Scene Description

What is the sequence of changes about the signboards at the beginning of the video?
A. (1) blue signboard with a downwards arrow. (2) yellow square signboard. (3) double round white signboards.
B. (1) yellow square signboard. (2) double round white signboards. (3) blue signboard with a downwards arrow.
C. (1) double round white signboards. (2) blue signboard with a downwards arrow. (3) yellow square signboard.
D. (1) blue signboard with a downwards arrow. (2) double round white signboards. (3) yellow square signboard.
Answer: A

Figure 16: Several examples in TUNA-MCQ, involving *Camera Motion*, *Camera Transition*, *Scene Description* and *Scene Transition* tasks.
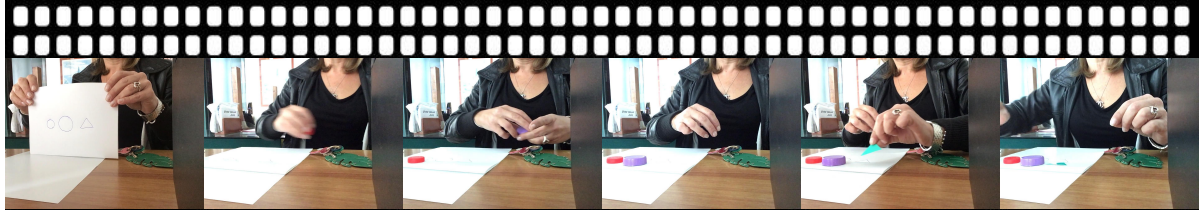
Task Type: Action Recognition

What happens after the man in the gray sweater waves his arms widely?
A. The man in the white shirt turns and enters the kitchen.
B. The man in the white shirt walks out of the kitchen.
C. The man in the gray sweater lowers his hands and glances back.
D. The man in the gray sweater raises both hands and waves them.
Answer: A

Task Type: Action Sequence

What is the order of the woman's actions involving the paper?
A. (1) places bottle caps. (2) holds the paper up. (3) draws shapes. (4) picks up a pen.
B. (1) picks up a pen. (2) draws shapes. (3) holds the paper up. (4) places bottle caps.
C. (1) draws shapes. (2) picks up a pen. (3) places bottle caps. (4) holds the paper up.
D. (1) holds the paper up. (2) picks up a pen. (3) draws shapes. (4) places bottle caps.
Answer: B
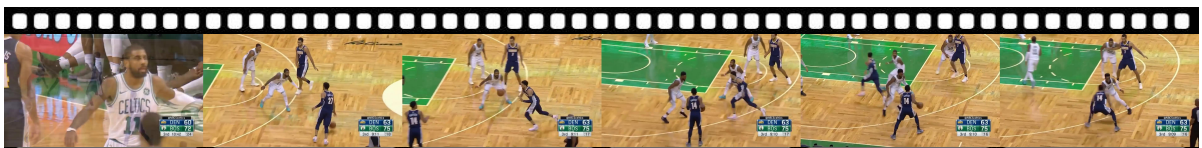
Task Type: Action-Subject Matching

What is the order of the woman's actions involving the paper?
A. Player in white jersey No. 7   B. Player in blue jersey No. 14
C. Player in blue jersey No. 7   D. Player in white jersey No. 11
Answer: B

Figure 17: Several examples in TUNA-MCQ, involving *Action Recognition*, *Action Sequence*, and *Action-Subject Matching* tasks.

Task Type: Object Recognition

What is the sequence of movements of the vehicles in the video?
A. (1) Black car. (2) White truck. (3) Blue truck. B. (1) White truck. (2) Black car. (3) Blue truck.
C. (1) Blue truck. (2) White truck. (3) Black car. D. (1) Blue truck. (2) Black car. (3) White truck.
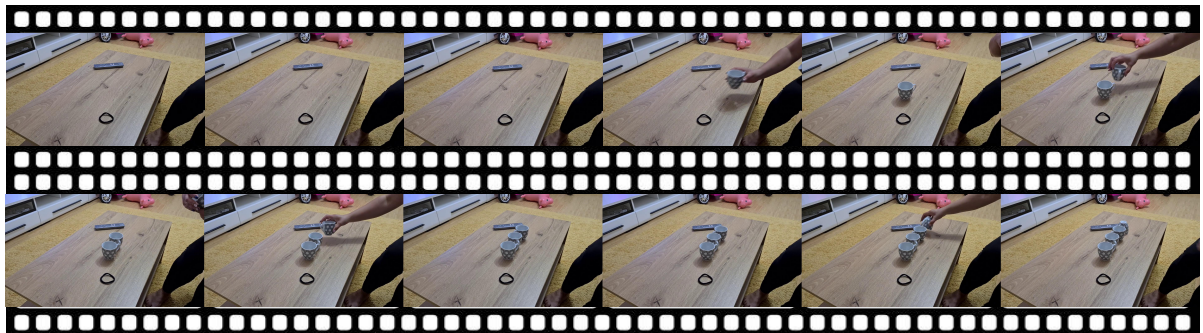Answer: B

Task Type: Object Appearance

What color is the first railing the man jumps over?
A. red railing   B. metal railing   C. blue curved railing   D. blue railing
Answer: B

Task Type: Object Location

Which direction are the cups being placed on the coffee table?
A. from the right to the left   B. from top to bottom
C. in disorder   D. from the left to the right
Answer: D

Figure 18: Several examples in TUNA-MCQ, involving *Object Recognition*, *Object Appearance*, and *Object Location* tasks.

Default Prompt for Video Captioning

Please provide a "chronological" detailed description of the video, focusing on the camera states, background scenes, and subjects' actions and attributes.
The description should consist of several events that evolve chronologically.
Don't have any summary, such as "throughout the video". Don't have any surmises, and subjective feelings.
Only output the video description. DON'T ANSWER IN POINTS.

Figure 19: The default prompt used for the TUNA-CAP experiments in Section 4.2.

Prompt for Splitting Events

Given a chronological video caption, split it into multiple chronologically evolving events.
All events spliced together should be equal to the original caption.

Video Caption:
{model_generated_caption}

Output a List event_list formed as:
[event1, event2, ...]
where ```video_caption = ''.join(event_list)```
Note: If there are lots of repeats, please delete the repeats.

DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only output the List. Output:

Figure 20: The prompt used to split events for the TUNA-CAP experiments in Section 3.2.2.

Prompt for Matching Events

Given a chronological list of candidate events and a chronological list of reference events.
Finds a matching reference event for each candidate event and returns a tuple of ids (candidate_id, reference_id) for both. If there is no match then reference_id is None.
Note that event matching should also be done in chronological order.
Each reference event can be matched by multiple candidate events.

Candidate Events:
{candidate_events}
Reference Events:
{reference_events}

Output a List formed as:
[(1, reference_id_1), (2, reference_id_2), ...]
where, reference_id_1 <= reference_id_2 <= ... <= reference_id_n if reference_id_i is not None.

DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only output the List. Output:

Figure 21: The prompt used to match events for the TUNA-CAP experiments in Section 3.2.2.

Prompt for Classifying Relationships

Given a series of candidate events and corresponding ground-truth visual elements contained in a video, you need to judge whether the candidate event accurately and completely describes each visual elements.
For each event, classify the relationship between the candidate event and the ground-truth visual elements into three classes: entailment, lack, contradiction.
- "entailment": the candidate event entails the visual element.
- "lack": the candidate event lacks the visual element.
- "contradiction": some detail in the candidate event contradicts with the visual element. Pay attention to the correspondence between the character and the action.

Candidate Events and Ground-truth Visual Elements:
{match_data}

Output a JSON formed as:
[
  {
    "candidate_event": "copy the candidate_event here",
    "visual_elements": [{"content": "copy the visual_element here", "relationship": "put class name here"}, ...]
  },
  ...
]

DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only output the JSON. Output:

Figure 22: The prompt used to classify relationships for the TUNA-CAP experiments in Section 3.2.2.

Prompt for Video Classification

Bellow is a description of a video clip:
{caption}

Please categorise the video in two ways, based on the video description.
- Visual Characteristic Category
  - Low Dynamic: The subjects in the video have a minimal amount of action.
  - High Dynamic: The subjects in the video have a lot of action.
  - Multi-Scene: There is at least one camera transition or scene transition.
  - Multi-Subject: At least 2 subjects appear in the video, where the subject must be the main object in the video, non-main object is not considered as a subject.
- Domain: Natural Landscape, Plants, Animals & Pets, Foods, Cityscape, Urban Activity, Autos & Vehicles, Sports Activity, Kitchen Cooking, Film, Driving, Daily Life.
  "Others" if the video does not belong to all of the above categories.

A video can belong to multiple Visual Characteristic Category, and must belong to one of Low Dynamic and High Dynamic, with Multi-Scene and Multi-Subject being optional.
A video belongs to only one Video Content Category.

Output a JSON formed as:
{"visual_characteristic": "select a visual characteristic category", "domain": "select a domain"}

DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only output the JSON. Output:

Figure 23: The prompt used to classify videos for the TUNA-CAP construction in Section 3.2.2.

Default Prompt for Multi-Choice QA (uniform sampling)

You will be provided with {num_frames} separate frames uniformly sampled from a video, the frames are provided in chronological order of the video. Analyze these frames and provide the answer to the question about the video content. Answer the multiple-choice question about the video content.
You must use these frames to answer the question; do not rely on any external knowledge or commonsense.
Question: {question}
Answer with the option's letter from the given choices directly.

Default Prompt for Multi-Choice QA (fps sampling)

You will be provided with separate frames sampled at {fps} fps from a video, the frames are provided in chronological order of the video. Analyze these frames and provide the answer to the question about the video content. Answer the multiple-choice question about the video content.
You must use these frames to answer the question; do not rely on any external knowledge or commonsense.
Question: {question}
Answer with the option's letter from the given choices directly.

Figure 24: The default prompt used for the TUNA-MCQ experiments in Section 4.3.

Prompt for Error-prone Points Generation

You are an AI visual assistant specialized in analyzing videos.
Given 8 frames uniformly sampled from a video clip and a human-annotated video description, your task is to extract the elements of the video frames that differ from the textual description. Only the inconsistent elements are output.

Output a JSON formed as:
[
  {"content_frame": "", "content_description": ""}}
  ...
]

Video Description
{video_caption}

Output
DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only output the List. Output:

Figure 25: The prompt used to generate error-prone points for the TUNA-MCQ construction in Section 3.2.3.

Default Prompt for Multi-Choice Q&As Generation

You are an AI visual assistant specialized in analyzing videos.

Given a human-annotated video description, and several error-prone points, your task is to propose five challenging multi-choice QAs about complex temporal-dynamic fine-grained understanding and reasoning. These questions should be require integrating information across multiple events and frames.

Requirements
- Questions can include temporal understanding and reasoning, counterfactual reasoning, causal reasoning, etc.
- You only need to focus on key events and objects. Ensure that the question can be answered from the given video description. Ensure that a single event or frame cue cannot answer these questions.
- The four options should be confusing and of similar length. Ensure that it is unable to judge the correct option just by the textual question and four textual options. For example, if there are multiple elements in an answer, then each correct element should not be the one that appears the most times among the four options.
- If the options are an disordered list of some elements (at least 3), normalise the elements to "(a) element_i\n(b) element_j\n..." as a part of the question, and answer should be formatted as something like "(b)(c)(a)(d)".

Task Type
1. camera motion; 2. camera transition; 3. scene description; 4. scene transition; 5. action recognition; 6. action sequence; 7. action-subject matching; 8. object recognition; 9. object appearance; 10. object location.

Output a JSON formed as:
[
  {"question": "", "task_type": "", "answer": "", "options": {"A": "", "B": "", "C": "", "D": ""}, "correct_option", ""}
  ...
]

Reference Examples (for reference only, not limited)
{"question": "What is the order of camera state changes throughout the video?", "task_type": "camera motion", "answer": "rotate left.", "options": {"A": "stationary.", "B": "zoom out.", "C": "rotate left.", "D": "pan left."}, "correct_option", "C"}
{"question": "How many times does the camera switch in the video? How many of these camera shots are close-ups of the woman?", "task_type": "camera transition", "answer": "4, 2.", "options": {"A": "3, 2.", "B": "4, 2.", "C": "4, 0.", "D": "4, 1."}, "correct_option", "B"}
{"question": "Reasoning which team will score based on the video?","task_type": "action recognition", "answer": "The team in red uniforms.", "options": {"A": "Not Sure.", "B": "The team in yellow uniforms.", "C": "The team in blue uniforms.", "D": "The team in red uniforms."}, "correct_option", "D"}
{"question": "What is the order in which this person picks up the objects?\n(a) a book\n(b) a pen\n(c) an apple", "task_type": "action sequence", "answer": "11", "options": {"A": "(c) (b) (a)", "B": "(a) (b) (c)", "C": "(b) (c) (a)", "D": "(b) (a) (c)"}, "correct_option", "D"}
{"question": "Which number player scored the goal?", "task_type": "action-subject matching", "answer": "11", "options": {"A": "11", "B": "17", "C": "7", "D": "1"}, "correct_option", "A"}
{"question": "What is the temporal order of occurrence of the following objects?\n(a) an apple\n(b) a guava\n(c) a banana\n(d) a loaf of bread", "task_type": "object recognition", "answer": "(c) (d) (b) (a)", "options": {"A": "(c) (b) (d) (a)", "B": "(d) (c) (a) (b)", "C": "(c) (d) (b) (a)", "D": "(d) (b) (c) (a)"}, "correct_option", "C"}

Input

Video Description
{video_caption}

Error-prone Points
{error_prone_points}

Output
DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only output the List. Output:

Figure 26: The prompt used to generate multi-choice QAs for the TUNA-MCQ construction in Section 3.2.3.