Burn *After* Reading: Do Multimodal Large Language Models Truly Capture Order of Events in Image Sequences?

Anonymous ACL submission

Abstract

This paper introduces the new and challenging TempVS benchmark, which focuses on temporal grounding and reasoning capabilities of Multimodal Large Language Models (MLLMs) in image sequences. TempVS consists of three main tests (i.e., event relation inference, sentence ordering and image ordering), each accompanied with a basic grounding test, yielding a total of 2,085 annotated image sequences and 15k multiple-choice questions. TempVS requires MLLMs to rely on both visual and linguistic modalities to understand the temporal order of events. We extensively evaluate ***38*** state-of-the-art MLLMs, demonstrating that models struggle to solve TempVS. Our analysis reveals a substantial performance gap between current MLLMs and human capabilities, accompanied by fine-grained insights that suggest promising directions for future research.

1 Introduction

Multimodal Large Language Models (MLLMs) (Achiam et al., 2023; Gemini et al., 2024; Liu et al., 2024a) have demonstrated remarkable performance in various vision and language tasks. At the same time, the need for standardized evaluation frameworks has become increasingly critical in systematically assessing and comparing MLLMs' performance across different tasks, domains and settings. Most existing benchmarks focus on settings involving a single image (Fu et al., 2023; Li et al., 2024b; Yue et al., 2024; Liu et al., 2024c; Lu et al., 2024). While some also consider multi-image settings (Jiang et al., 2024; Fu et al., 2024; Ying et al., 2024; Li et al., 2024b), they mainly focus on cross-image recognition and reference. To date, relatively little attention has been paid to more complex tasks, such as temporal understanding and reasoning in multiple images.

Some recent studies have assessed MLLMs' temporal comprehension across multiple images, but certain limitations remain. First, some tasks can be resolved by relying on a single image rather than a sequence (Liu et al., 2024b; Ying et al., 2024). For example, determining whether "pulling a wind-up toy so it continues moving forward or quickly stops" could be answered only based on the final image. Second, some tasks depend heavily on commonsense or world knowledge (Wang et al., 2025; Meng et al., 2025), such as rearranging a set of shuffled images into the correct sequence of cooking steps. Third, some benchmarks (Song et al., 2024) use distractor options absent from the images, allowing the models to infer ground-truth answers based on the presence of objects. These factors may result in the benchmarks failing to truly assess the model's understanding of temporal sequences. In addition, none of the existing benchmarks are designed for multi-event scenarios, making them inadequate for evaluating complex temporal sequences and relations in MLLMs.

As a result, a question arises: *Do existing MLLMs really understand time by accurately aligning the order of events in language and image sequences?* To address this, we propose **TempVS**, a benchmark for multi-event **Temp**oral grounding and reasoning in **V**isual **S**tory image sequences. TempVS contains 2,085 image sequences (9,803 images) across cartoon animations, movies and daily-life albums, with 15,192 multiple-choice questions. TempVS features three temporal understanding and reasoning tasks: event relation inference, sentence ordering, and image ordering (as shown in Figure 1). These tasks are accompanied by grounding tasks to check whether the model can match the exact image consistent with a single text description. We select image sequences making up visual stories, each containing several events that are temporally related yet relatively independent, in the sense that no event could easily be predicted from preceding events. This makes it hard to resolve the tasks without considering both linguistic

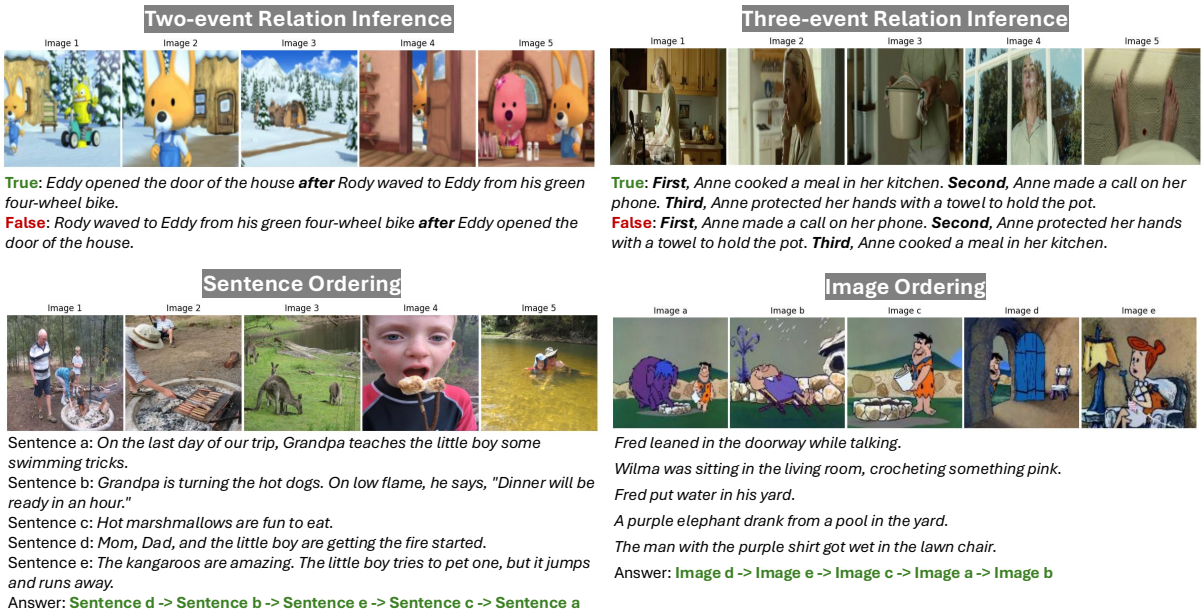


Figure 1: Illustrative examples from the main tests of TempVS benchmark. Additional examples are provided in Appendix E.

and visual modalities. TempVS challenges models to reason about event order in the image sequence and text (e.g., a sentence describing two events using *before* or *after*, or a story) and integrate both.

We extensively evaluate 38 MLLMs, including open-source models ranging from 0.5B to 78B (e.g., LLaVA-OneVision, InternVL2.5, Qwen2-VL, Phi-3.5-vision, DeepSeek-vl2, LLaVA-NeXT-Video) and the closed-source GPT-4o. We show that TempVS is highly challenging for SOTA models, especially on event relation inference and image ordering tasks. In particular, while models can accurately ground events to images, their performance on the main tasks, which require multimodal reasoning with sequences, remains unsatisfactory. We further analyze the impact of the choice of linguistic structure, distance between events, and Chain-of-Thought prompting. Our analysis sheds light on future directions for improvement in architectural design, training objectives, and/or post-training methods to enhance temporal reasoning.

Contributions. (1) We introduce a new benchmark TempVS, for evaluating multi-event temporal grounding and reasoning ability in image sequences for MLLMs. (2) We extensively evaluate 38 MLLMs from different model families and sizes, highlighting the performance gap compared to human annotators. (3) Our findings in evaluation results and fine-grained analysis suggest potential pathways for future improvements.

2 Related Work

Multimodal Large Language Models Progress in large language models (LLMs) (Achiam et al., 2023; AI@Meta, 2024; Touvron et al., 2023; Gemini et al., 2024) has provided impetus to the development of multimodal LLMs which process both visual and textual information. State-of-the-art MLLMs (Achiam et al., 2023; Dai et al., 2023; Gemini et al., 2024; Liu et al., 2024a; Abdin et al., 2024; Wang et al., 2024a; Chen et al., 2024) are built upon LLMs with an integrated visual encoder and a connection module. These models surpass the earlier generation of multimodal models, which were typically based on BERT-type architectures, on many downstream tasks (Bugliarello et al., 2023). While some studies have focused on training MLLMs to interpret multiple images using interleaved image-text datasets (Jiang et al., 2024; Huang et al., 2024; Li et al., 2024a,c), their capability to understand and reason about multi-event temporal relationships in sequential visual data remains largely unexplored.

Multi-image Benchmarks Multi-image understanding requires MLLMs to compare, analyze and interpret relationships across multiple images (Li et al., 2024d). Benchmarks such as NLVR2 (Suhr et al., 2019), BLINK (Fu et al., 2024), SEED-Bench-2 (Li et al., 2024b), Mantis-Eval (Jiang et al., 2024) and MMT-Bench (Ying et al., 2024) cover a subset of multi-image tasks focusing on assessing

models' ability to identify similarities and variations across multiple images. DEMON (Li et al., 2024e) evaluates the demonstrative instruction-following abilities of MLLMs. Mementos (Wang et al., 2024b) aims to detect object and behavior hallucinations in descriptive text generated for sequential images. MileBench (Song et al., 2024) evaluates MLLMs' performance in long contexts. MIBench (Liu et al., 2024b) assesses MLLM's ability in multi-image instruction, multimodal knowledge-seeking and in-context learning. MuirBench (Wang et al., 2025) is a comprehensive multi-image understanding benchmark with unanswerable counterparts to test the robustness of MLLMs. MMIU (Meng et al., 2025) incorporates a large number of test questions that cover a diverse array of multi-image tasks and relationships. We propose TempVS, the first benchmark specifically designed for multi-event temporal understanding and reasoning in image sequences. In particular, it is designed to avoid shortcuts such as reliance on single images/frames and commonsense reasoning to bypass full integration of the temporal information in both text and images.

3 The TempVS Benchmark

TempVS evaluates models' ability to understand and reason about temporal relations by evaluating the consistency between *textual descriptions of temporally related events* and *the visual event order in an image sequence*. To achieve this, we create three main tests (MT): **event relation inference**, **sentence ordering**, and **image ordering**. Additionally, to investigate whether a model's difficulty arises from challenges in temporal understanding or from more basic grounding issues, we create a corresponding **grounding test** (GT) (§3.1) for main tests. We present the task curation process and statistics of TempVS benchmark in §3.2.

TempVS is built from existing datasets pairing image sequences with narrative captions. Since captions may vary in their level of detail, we use both the original captions in the source data, and simplified versions in which the main event is extracted from the caption (we describe this process in §3.2). In what follows, we denote an image sequence S consisting of n images, their corresponding captions, and extracted events as:

$$\mathcal{S} = [(I_1, C_1, E_1), (I_2, C_2, E_2), \ldots, (I_n, C_n, E_n)],$$

where I_i denotes the i-th image, C_i its associated caption, and E_i the extracted event.

Statement Template ($i < j < k$ in image sequence)		
Two-event	Pos: E_j after E_i	Neg: E_i after E_j
	Pos: E_i before E_j	Neg: E_j before E_i
	Pos: E_j. Earlier, E_i	Neg: E_i. Ealier, E_j
	Pos: E_i. Then, E_j	Neg: E_j. Then, E_i
	Pos: E_i. E_j.	Neg: E_j. E_i.
Three-event	Pos: E_i before E_j, and after that, E_k Neg: E_k before E_i, and after that, E_j	
	Pos: E_i. Later, E_j.Finally, E_k. Neg: E_j. Later, E_i.Finally, E_k.	
	Pos: First, E_i. Second, E_j.Third, E_k. Neg: First, E_i. Second, E_k.Third, E_j.	
	Pos: E_i. E_j. E_k.	Neg: E_k. E_j. E_i.

Table 1: Templates of positive (Pos) and negative (Neg) statements used for **MT1** event relation inference.

3.1 Main Tests and Grounding Test

MT1: Event Relation Inference MT1 evaluates a model's understanding of the chronological order of events based on an image sequence and a textual description. The text describes the temporal relation of the events either explicitly through adverbial markers such as *after* or *before*, or implicitly through the natural order of sentences. From an image sequence of length n, we select either (1) two pairs $\{(I_i, E_i), (I_j, E_j)\}$ where $i < j \leq n$, ensuring that I_i appears before I_j and E_i occurs before E_j; or (2) three pairs $\{(I_i, E_i), (I_j, E_j), (I_k, E_k)\}$ where $i < j < k \leq n$, following the same ordering constraints. These event-image pairs are not necessarily adjacent in the sequence, resulting in varying distances between them. We then generate positive and negative statements by applying the templates in Table 1 to describe the temporal relations between these events. The negative statement retains the same event clauses and the temporal conjunction as the positive statement, while swapping the positions of these clauses, such that the text expresses a different temporal order.[1] Finally, we derive triples comprising an original image sequence, a positive or negative statement and the corresponding answer (True or False).

MT2: Sentence Ordering This task evaluates whether models can correctly reorder a shuffled set of sentences based on the temporal order of events in a given image sequence. Thus, MT2

[1]For the negative statements with three events, we randomly select one from the five combinations that is not the same as the positive statement.

requires not only an understanding of temporal relations between events but also consideration of the text's coherence and fluency. Models are tasked to select the correct sentence order from five given options, based on an ordered image sequence and a set of shuffled event descriptions. We create versions with both the original captions (C) and the extracted events (E).

MT3: Image Ordering This task adopts the conjugate form of MT2, requiring models to rearrange a set of shuffled images into the correct temporal order based on the given textual description. Similar to MT2, we also examine whether different text styles (i.e., original captions or extracted event descriptions) could affect the model's ability to determine the correct order.

GT: Grounding Test As a prerequisite for solving one of the main tests, we assume that MLLMs should be able to match an event description with the corresponding image in a multi-image sequence. Specifically, given an event description E_i or a caption C_i and image sequence $[I_1, I_2, \ldots, I_n]$, models are required to identify the index of the image (i.e., I_i) that best corresponds to the given textual description. The motivation for conducting grounding tests is as follows: If a model passes the grounding tests but fails the corresponding main tests, it indicates that the model struggles with understanding temporal order, even if it can accurately recognize and associate visual and linguistic elements. In contrast, if a model performs well on the main tests but fails the corresponding grounding tests, it may suggest that its success stems not from the true temporal grounding or reasoning but rather from leveraging statistical patterns, correlations, or systematic biases learned during training.

3.2 Benchmark Curation and Statistics

Data Source Given our objective of evaluating whether models understand the chronological order of events across image sequences and language, we require data containing multiple images that form a temporal sequence presenting events. We choose four visual story datasets: FlintstonesSV (Gupta et al., 2018), PororoSV (Li et al., 2019), VIST (Huang et al., 2016) and VWP (Hong et al., 2023). FlintstonesSV and PororoSV, designed for story visualization, contain annotated frames from cartoon animations.[2] VIST, built for visual storytelling, originates from Flickr albums with user-uploaded daily-life photos. VWP features movie scene sequences paired with aligned synopses. This collection with rich styles and diverse domains plays a crucial role in assessing MLLMs' multi-event temporal grounding and reasoning capabilities.

Dataset Filtering To ensure that each image sequence contains a sufficient number of character-centered visual events, we use Detectron2[3] to detect and retain image sequences where PERSON can be detected in at least 60% of the images. To avoid temporal overlap between any two events, we remove any image sequence whose captions contain stative verbs such as *'belong'*, *'love'* and *'exist'*. To minimize ambiguity, we remove captions starting with pronouns and compute the BERTScore (Zhang et al., 2020) between captions, omitting sequences with highly similar captions. Similarly, we remove image sequences with highly similar images based on the cosine similarity between their CLIP (Radford et al., 2021) embeddings. Simple events (E) are extracted from the original captions (C) using GPT-4 (Achiam et al., 2023); any sequences with captions from which no event can be extracted are removed. To ensure that each image-text pair in $\{(I_i, E_i), (I_j, E_j)\}$ remains distinct, we use CLIP to compute cross-modality similarity between different image-text combinations. Ambiguous pairs are filtered based on a threshold, ensuring that within-pair similarity is significantly higher than cross-pair similarity.[4] A similar process is applied to sets of three image-event pairs $\{(I_i, E_i), (I_j, E_j), (I_k, E_k)\}$. In Appendix A, we provide the details of stative verbs list, the similarity thresholds and statistics of the dataset after each filtering step.

Prompt and Option Generation After filtering the datasets, we create the positive and negative statements by concatenating the events with the templates shown in Table 1 for MT1. We sequentially apply each template to its corresponding temporal relation group in the dataset, ensuring an even distribution of statement types. Moreover, we use ChatGPT(OpenAI, 2022) to generate variations of different prompt components for different tasks including task instructions, answer requirements and

[2]For the major characters in FlintstonesSV and PororoSV, we provide descriptions of their appearances to match them with their names.

[3]https://ai.meta.com/tools/detectron2/

[4]That is, we guarantee that $sim(I_i, E_i) > sim(I_i, E_j)$ and $sim(I_i, E_i) > sim(I_j, E_i)$ and $sim(I_j, E_j) > sim(I_i, E_j)$ and $sim(I_j, E_j) > sim(I_j, E_i)$.

	MT1 (two)	MT1 (three)	MT2 (event)	MT2 (caption)	MT3
FlintstonesSV	2,104	916	501	485	565
PororoSV	864	172	320	326	395
VWP	850	208	274	256	316
VIST	3,742	830	708	551	809
TempVS	7,560	2,126	1,803	1,618	2,085

Table 2: TempVS benchmark statistics: In MT1, the number indicates the total statements; in MT2, the number of image sequences and corresponding shuffled sentence sets; in MT3, the number of textual events or captions and their associated shuffled image sets.

response formats (see Appendix C), which results in a total of 328 possible prompt variations. By incorporating sufficient diversity in prompts, we mitigate the risk of results being influenced by a specific prompt formulation (Sclar et al., 2024). All tests are formed as multiple-choice questions. In MT1, the options are "True" and "False" with positions alternated across samples (e.g., A. True; B. False. and A. False; B. True.) to prevent position bias (Zheng et al., 2024). In MT2 and MT3, one correct sequence is presented alongside four randomly shuffled incorrect sequences with options labeled "A" to "E". The Grounding Test uses image indices as answer choices. To ensure fair evaluation, correct answers are evenly distributed across options throughout the benchmark.

Quality Control To discourage "blind" models that leverage language biases, we filter examples in the benchmark that could be easily solved based only on the linguistic modality. We apply three unimodal LLMs Phi-3.5-mini-instruct [4B] (Abdin et al., 2024), Llama-3 [8B] (AI@Meta, 2024) and Qwen-2.5-instruct [72B] (QwenTeam, 2024), which are popular LLM backbones in current MLLMs. In MT1 and MT2, we discard a sample if at least two LLMs can answer the question correctly without visual inputs. A manual check was performed by the authors to exclude ambiguous images, grammatically incorrect and/or semantically implausible statements, and cases where image sequences did not match the text.

Benchmark Statistics TempVS consists of 2,085 distinct image sequences with corresponding original captions and extracted events. Table 2 shows the statistics of each task from each data source in TempVS. Most image sequences in the benchmark contain 5 images each, except for 61 sequences from VWP, which have 6 to 9 images.

4 Experiments

4.1 Experimental Setup

Models We evaluate a diverse family of state-of-the-art models with various sizes (ranging from 0.5B to 78B) and different vision and LLM backbones. We select DeepSeek-vl2 [3B/16B] (Wu et al., 2024), InternVL2.5 [1B/8B/26B/78B] (Chen et al., 2024), Janus-Pro [1B/7B](Chen et al., 2025), LLaVA-NeXT-Interleave [0.5B/7B] (Li et al., 2024c), LLaVA-OneVision [0.5B/7B/72B] (Li et al., 2024a), LLaVA-NeXT-Video [7B/34B] (Zhang et al., 2024b), LongVA [7B] (Zhang et al., 2024a), Mantis[8B] (Jiang et al., 2024), Phi-3-vision [4B], Phi-3.5-vision [4B] (Abdin et al., 2024), and Qwen2-VL [2B/7B/72B] (Wang et al., 2024a). We also evaluate GPT-4o [2024-11-20]. The implementation details are provided in Appendix D.1.

Evaluation Metrics For multiple-choice questions, we benchmark MLLMs' performance using accuracy as the metric for predicted options. Accuracy scores are reported for both the main tests (MT) and their corresponding grounding tests (GT). In MT1, where each question relates to either two or three events, there is a GT for each separate event. In MT2 and MT3, each question corresponds to a number of grounding tests equal to the number of images in the sequence. Additionally, we introduce a stricter metric, GT_{strict}, which assesses the number of image sequences where a model passes all corresponding grounding tests.[5] To examine the relationship between main test and grounding test performance, we further report MT|GT_{strict}, where a model's success on a main test is considered valid only if it passes all corresponding grounding tests.

4.2 Main Results

Results for 21 state-of-the-art MLLMs are shown in Table 3. For the full quantitative results of all tested 38 MLLMs, we refer to Appendix D.2. Our main observations of the empirical results are as follows:

TempVS is challenging even for SOTA MLLMs InternVL2.5-26B-MPO achieves the highest performance on the two-event relation inference, while GPT-4o leads in the three-event relation inference.

[5]Appendix D.2 provides overall grounding test results, showing models' ability to precisely locate the correct image within a sequence based on textual input in general.

		Two-event Relation (MT1)			Three-event Relation (MT1)			Sentence Ordering - event (MT2)			Sentence Ordering - caption (MT2)			Image Ordering - event (MT3)		Image Ordering - caption (MT3)	
		GT_s	MT	MT|GT_s	GT_s	MT	MT|GT_s	GT_s	MT	MT|GT_s	GT_s	MT	MT|GT_s	MT	MT|GT_s	MT	MT|GT_s
Random		4	50		0.6	50		0.032	20		0.032	20		20		20	
DeepSeek-vl2	3B	20.8	49.7	49.7	10.5	49.8	50.6	0.8	19.5	25.0	1.5	18.2	17.9	20.4	42.9	20.7	12.5
	16B	14.2	43.1	42.2	6.7	44.4	44.4	0.4	15.7	14.3	0.6	17.1	18.2	16.6	0.0	15.5	0.0
InternVL2.5	26B	46.0	57.1	58.4	39.6	58.4	60.2	10.2	56.7	73.6	13.1	63.0	76.9	26.7	27.3	31.7	35.9
	26B-MPO	51.3	**60.3**	**62.1**	46.0	62.1	64.7	12.6	69.9	90.8	17.0	76.9	87.3	34.4	39.7	39.5	43.7
	78B	51.6	54.2	55.3	47.3	56.8	57.0	13.6	67.0	84.9	18.5	71.1	83.9	31.1	40.0	38.5	47.1
	78B-MPO	58.8	58.5	59.9	56.5	61.4	62.6	18.4	**79.8**	96.6	25.9	**86.3**	**96.4**	**41.0**	48.8	**53.8**	**69.7**
Janus-Pro	1B	2.7	48.3	48.1	0.7	46.5	42.6	0.0	18.6	-	0.1	19.8	0.0	22.5	0.1	22.3	0.0
	7B	4.3	35.1	34.1	0.4	32.9	39.3	0.0	17.1	-	0.0	15.3	-	20.9	-	21.0	-
LLaVA-NeXT-Interleave	0.5B	2.5	49.8	47.8	0.2	50.4	50.0	0.0	20.7	-	0.0	20.7	-	20.2	-	20.8	-
	7B	13.0	51.6	52.4	7.2	50.1	49.8	0.3	25.1	16.7	0.4	27.0	0.0	20.9	16.7	20.0	22.2
LLaVA-OneVision-ov	0.5B	8.6	45.3	45.5	2.9	48.1	47.6	0.1	18.8	0.0	0.1	18.4	0.0	19.4	**100.0**	19.0	0.0
	7B	32.8	56.0	58.0	26.0	57.5	59.8	4.5	44.2	41.2	6.8	46.9	47.6	21.3	14.3	21.6	19.8
	72B	46.4	59.3	**62.1**	40.5	61.5	63.5	9.8	65.2	81.8	14.0	75.1	86.6	27.6	31.8	29.1	36.5
LLaVA-NeXT-Video	7B	5.8	46.0	46.0	1.4	44.9	47.0	0.0	19.0	-	0.0	18.2	-	21.0	-	21.3	-
	34B	6.5	58.5	58.4	1.6	59.5	57.6	0.1	31.8	**100.0**	0.0	33.4	-	19.8	-	20.0	-
LongVA	7B	8.7	54.7	56.1	2.1	56.0	61.7	0.2	34.2	66.7	0.2	35.3	50.0	19.5	0.1	19.0	-
Mantis-Idefics	8B	12.2	51.9	53.3	4.1	52.0	51.6	0.1	22.2	0.0	0.2	20.8	0.0	18.6	0.0	19.2	50.0
Phi-3.5-vision	3.4B	4.0	49.0	47.7	0.8	48.8	48.3	0.0	23.1	-	0.0	25.4	-	19.2	-	18.3	-
Qwen2-VL-Instruct	7B	32.4	54.0	55.4	21.2	53.6	55.3	3.4	42.5	64.6	4.4	44.6	61.3	23.1	20.4	24.6	35.9
	72B	31.7	54.0	56.4	20.6	55.6	60.6	3.7	46.3	64.3	5.0	55.1	70.0	26.5	47.3	28.1	47.4
GPT-4o	API	**60.3**	58.3	60.1	**57.0**	**64.5**	**66.4**	**18.6**	53.4	53.9	**28.6**	61.5	55.3	22.6	23.5	23.0	23.5

Table 3: Zero-shot average accuracy performance of 11 popular MLLMs families with 21 variants on TempVS benchmark on strict grounding test (GT_s), main test (MT) and the main test when all corresponding grounding tests pass (MT|GT_s). Best models per metric are marked in boldface and the second best models are underlined.

InternVL2.5-78B-MPO outperforms other models in both sentence ordering and image ordering tasks. However, most models with parameters less than or equal to 7B exhibit random chance accuracies of approximately 50% (for MT1) and 20% (for MT2 and MT3). Most MLLMs perform similarly on two-event and three-event relation inference tasks, while for the strongest models (such as InternVL2.5[26B/78B], LLaVA-OneVision-ov-72B and GPT-4o), their performance is even slightly better on the understanding three-event statements. In addition, sentence ordering is a relatively simple task (with the highest accuracy at 86.3%), while image ordering is a significantly more challenging task (with the highest accuracy only at 53.8%). GPT-4o performs substantially worse than several of the best-performing open-source models in both sentence and image ordering tasks. In terms of language type, we find that sentence and image ordering are easier with original captions than with extracted events. This may indicate that models might leverage additional contextual details and temporal cues from the original captions, which are unavailable in the simpler extracted event descriptions.

Both scale and post-training help As shown in Figure 2, accuracy generally improves with model size. However, the marginal gains diminish as models get larger. In the two-event and three-event relation inference, this effect is more evident, with smaller models (7B or 26B) already attaining competitive and in some cases higher accuracy compared to larger models ($>$ 70B). Additionally, the two ordering tasks exhibit larger performance gaps between smaller and larger models, which can be attributed to the superior long-range reasoning capabilities of the more powerful LLM backbones in larger MLLMs. Surprisingly, DeepSeek-VL2 [3B/16B] and Janus-Pro [1B/7B] are exceptions, as their smaller models outperform the larger ones in most cases.

Our results also highlight the importance of high-quality post-training: under the same model sizes, InternVL2.5-MPO consistently outperforms InternVL2.5 across all evaluated tasks, especially when the model parameters exceed 7B. These results indicate that Mixed Preference Optimization (MPO) (Chen et al., 2024) effectively enhances the overall multimodal temporal understanding and reasoning capabilities. Similarly, models fine-tuned using Direct Preference Optimization (DPO) are consistently better than their SFT-only counterparts. Other results comparing LLaVA-NeXT-Video and LongVA families are provided in Appendix D.2. We also see a positive impact of instruction tuning on complex reasoning tasks such as image ordering (e.g., Qwen2-VL-Instruct outperforms Qwen2-VL across all tasks).

Event grounding does not guarantee understanding of temporal relations In MT1, we observe a slight difference between MT and MT|GT_{strict}. This suggests that even if a model can accurately identify the image corresponding to every single event within a sequence, it may still lack the ability to understand the chronological order of events in text. In short, grounding textual descriptions and reasoning about their temporal relations require different capabilities. Among all grounding tests, GPT-4o performs the best, but it lags significantly behind top-tier open-source models like InternVL2.5-MPO[26B/78B] especially in the two ordering tasks (MT2 and MT3). In MT2

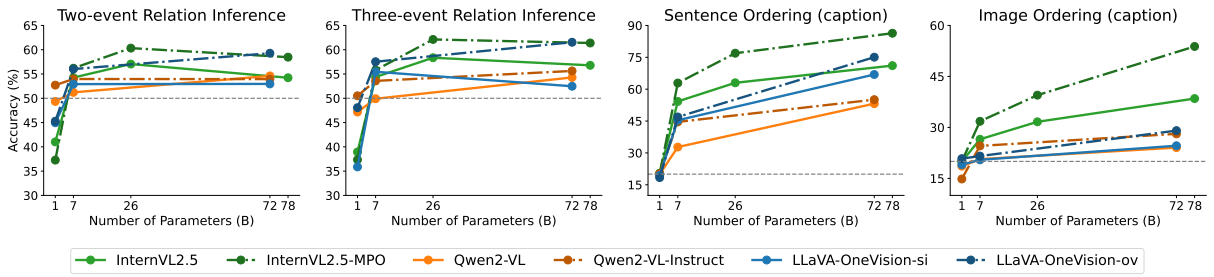


Figure 2: Illustrative TempSV benchmark results of selected models with different number of parameters.

	# Image seqs	Accuracy	Fleiss' kappa
Two-event Relation	60	82.5	0.728
Three-event Relation	60	81.6	0.689
Sentence Ordering (event)	40	81.2	0.751
Sentence Ordering (caption)	40	89.3	0.764
Image Ordering (event)	40	79.1	0.827
Image Ordering (caption)	40	77.9	0.742

Table 4: Results of human evaluation on all main tests.

and MT3, MT|GT$_{strict}$ accuracy improves substantially over MT for most large MLLMs, confirming the fundamental role of visual grounding in sentence or image ordering tasks. For instance, the MT|GT$_{strict}$ accuracy scores of InternVL2.5-78B-MPO are higher than MT by 16.8% (event) and 10.1% (caption) in MT2, and by 7.8% (event) and 15.9% (caption) in MT3. Meanwhile, for small models like DeepSeek-vl2[3B], MT|GT$_{strict}$ accuracy is sometimes even smaller than MT accuracy. This indicates some dependency of these models' reasoning abilities on their grounding capability, though this is unlikely to be the only factor affecting performance.

Gap with human performance To evaluate the quality and estimate the difficulty of TempVS, we perform a human assessment on 280 randomly selected image sequences, covering all fine-grained statement types of MT1 as well as both language types (C and E) for MT2 and MT3. We recruited 36 annotators on the Prolific platform and collected three responses for each image sequence, yielding 840 responses collected in total. Further details are provided in Appendix B.

There is a large gap between human average performance (Table 4) and SOTA MLLMs. For the tasks of two-event and three-event relation inference and image ordering, there remains plenty of room for improvement. However, for the sentence ordering task, the strongest model InternVL2.5-78B-MPO is already close to human performance. Humans exhibit substantial agreement among themselves on the main tasks, with Fleiss' kappa (Landis and Koch, 1977) above 0.68 across the board.

4.3 Further Analysis

Impact of temporal expressions We further analyze how the models understand and reason about different types of statements in the two-event and three-event relation inference tasks (MT1). We select the top five models on these two tasks for comparison (Table 5). When comparing explicit and implicit temporal event statements (cf. Table 1), we observe that models consistently perform better on the former. This could be because the presence of temporal adverbs or conjunctions in a sentence helps clarify the order in which events happen. The top five models always achieve higher accuracy on positive examples than on negative ones. Despite the even distribution of "True" and "False" across options "A" and "B" in our benchmark, models exhibit a tendency to predict "True" more frequently. By incorporating adversarial samples, our TempVS benchmark effectively reveals MLLMs' biases toward certain answers, providing a robust assessment of their compositional temporal reasoning capabilities.

We also observe the better performance of models on explicitly marked temporal relations involving *before* (resp. *after*) and on *then* (resp. *earlier*). The key difference between these pairs of complementary temporal adverbials is that with *before* and *then*, the order of events in text mirrors their order in the image sequence, whereas this is not the case with *after*/*earlier*. We therefore see some evidence of an *iconicity effect*: temporal relations are easier for models when the surface order of events mirrors their actual order (in the visual modality). This echoes similar findings in the discourse processing and psycholinguistic literature on narrative comprehension (Zwaan and Radvansky, 1998; Smith, 2003). It also points to an important avenue for future research in fine-grained multimodal bench-

Tasks	Models	InternVL2.5-26B-MPO		InternVL2.5-78B-MPO		llava-onevision-72b-ov		LLaVA-NeXT-Video-34B		GPT-4o		Top-5 Average		
	Statement Type	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Pos	Neg	Both
Two-event Relation	after	59.9	62.9	70.7	49.1	71.4	50.6	77.9	37.1	56.2	60.4	67.2	52.0	59.6
	before	66.4	66.7	71.6	57.9	70.6	59.0	77.1	41.4	53.3	73.4	67.8	59.7	63.8
	earlier	70.8	47.0	74.3	36.8	74.1	41.5	85.9	28.8	69.1	38.5	74.8	38.5	56.7
	then	65.5	59.3	73.3	44.8	71.3	50.0	86.4	35.2	61.0	63.1	71.5	50.5	61.0
	implict	65.3	38.7	74.4	31.6	74.4	30.1	86.4	28.3	64.0	43.4	72.9	34.4	53.6
Three-event Relation	before/after	66.7	72.3	70.7	64.0	77.7	54.6	71.2	72.0	55.7	78.2	65.1	67.2	66.2
	first/second/third	55.2	70.4	65.4	67.5	76.5	45.9	68.4	64.1	50.4	78.9	60.4	65.5	63.0
	later/finally	75.2	56.3	87.5	32.5	77.4	52.8	100.0	0.1	69.4	68.3	79.8	43.1	61.5
	implict	73.3	27.1	82.6	20.3	78.7	28.4	100.0	0.3	70.3	43.7	76.3	28.4	52.4

Table 5: Fine-grained accuracy of different statement types in the two-event and three-event relation inference tasks. "Pos" denotes positive examples, while "Neg" represents negative examples. The higher accuracy score is highlighted in gray for each pair of positive and negative statements.

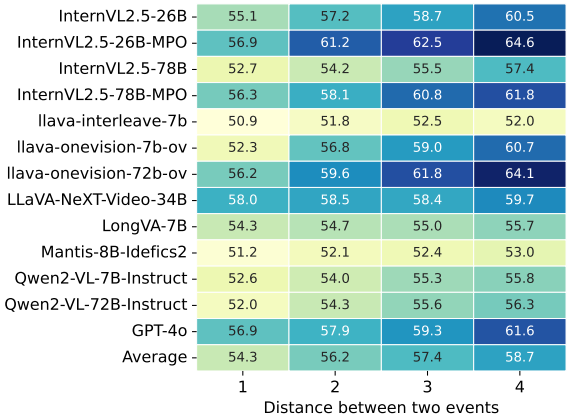


Figure 3: Accuracy of the two-event relation inference task with different distances between the two events in a sequence.

marking, namely, in cases where surface characteristics in two modalities are not perfectly aligned.

Distance between two events Figure 3 presents the accuracy of the two-event relation inference task as the distance between events in the original sequence increases from one to four. Nearly all models improve in performance as the distance increases. On the one hand, more distant visual events in the dataset are typically more visually distinct. On the other hand, the models may fail to effectively separate two closely spaced images, even though separators were inserted between images in the input.[6]

Prompting with Chain-of-Thought Chain-of-Thought (CoT, Wei et al., 2022) is a widely used approach to enhance models' reasoning ability, by allowing models to generate intermediate reasoning steps before producing the final answer. It may thus also enhance models' temporal reasoning skills on TempVS. We conduct CoT experiments using InternVL2.5-78B and GPT-4o. The detailed prompts are listed in Appendix C. As shown in Table 6, CoT yields large gains on sentence and image ordering, but limited improvement for event relation inference (MT1). This indicates the potential of step-by-step reasoning for ordering tasks. However, simple CoT does not help event relation inference. We leave the investigation of methods to enhance models' understanding of this complex task and improve its temporal reasoning capabilities for future work.

[6]In our experiments, we combine sequential images into one picture. In our preliminary experiments, we attempted to input the images into the model sequentially, one at a time, and observed little difference in performance compared to merging them into a single input. To mitigate the risk of exceeding the context length in some MLLMs, we chose to merge the images.

	InternVL2.5-78B			GPT-4o		
	w/o. CoT	w. CoT	Δ	w/o. CoT	w. CoT	Δ
Two-event relation (MT1)	54.23	56.13	**+1.90**	58.27	60.43	**+2.16**
Three-event relation (MT1)	56.79	57.20	**+0.41**	64.52	63.38	**-1.14**
Sentence Ordering (MT2)	71.09	85.42	**+14.33**	61.50	81.41	**+19.91**
Image Ordering (MT3)	38.49	47.54	**+9.05**	22.97	33.77	**+10.80**

Table 6: Model performance comparison with and without Chain of Thought (CoT).

5 Conclusion

In this paper, we present a novel and challenging benchmark TempVS, designed to assess multimodal, multi-event temporal reasoning abilities of MLLMs in image sequences. After evaluating 38 advanced MLLMs, we find that current models typically struggle with reasoning about temporal relations and rearranging shuffled images to the correct order based on a narrative. Further analysis of linguistic structures, event distance, and Chain-of-Thought reasoning has shed light on promising avenues for future work. Our study contributes to MLLM development by uncovering weaknesses in multi-event temporal reasoning in multi-image scenarios, while TempVS provides a valuable resource for further research.

Limitations Despite a careful examination of publicly available repositories, technical reports and papers, we find no evidence that the evaluated MLLMs were trained on the data included in TempVS. However, for models (such as GPT-4o) that have not fully disclosed or explicitly stated their training data, the possibility of data leakage and contamination remains unclear. This could potentially lead to an overestimation of their advantages. Additionally, we focus on multiple-choice questions to ensure structured evaluation and clear correctness criteria. However, other question types, such as open-ended questions, particularly the evaluation of open-ended generation in multi-image temporal understanding and reasoning, are also worth exploring. We leave these for future work.

Ethical Considerations In this study, we employ published datasets and pretrained multimodal large language models, with no known significant ethical concerns regarding their usage. However, we acknowledge that biases in the original image-caption data may influence both the models and their evaluations. Our research has received approval from our institution's Ethics Board, ensuring compliance with ethical guidelines for human annotation process. Additionally, all collected human-annotated data has been deidentified to protect participants' data privacy and security.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI@Meta. 2024. Llama 3 model card.

Emanuele Bugliarello, Laurent Sartran, Aishwarya Agrawal, Lisa Anne Hendricks, and Aida Nematzadeh. 2023. Measuring progress in fine-grained vision-and-language understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1559–1582, Toronto, Canada. Association for Computational Linguistics.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

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

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer.

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

Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. 2018. Imagine this! scripts to compositions to videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 598–613.

Xudong Hong, Asad Sayeed, Khushboo Mehra, Vera Demberg, and Bernt Schiele. 2023. Visual writing prompts: Character-grounded story generation with curated image sequences. *Transactions of the Association for Computational Linguistics*, 11:565–581.

Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1233–1239.

Yupan Huang, Zaiqiao Meng, Fangyu Liu, Yixuan Su, Nigel Collier, and Yutong Lu. 2024. Sparkles: Unlocking chats across multiple images for multimodal instruction-following models.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max W.F. Ku, Qian Liu, and Wenhu Chen. 2024. Mantis: Interleaved multi-image instruction tuning. *Transactions on Machine Learning Research*, 2024.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

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. *Preprint*, arXiv:2408.03326.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024b. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308.

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. *Preprint*, arXiv:2407.07895.

Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, et al. 2024d. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*.

Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. 2024e. Fine-tuning multimodal LLMs to follow zero-shot demonstrative instructions. In *The Twelfth International Conference on Learning Representations*.

Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. 2019. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6329–6338.

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

Haowei Liu, Xi Zhang, Haiyang Xu, Yaya Shi, Chaoya Jiang, Ming Yan, Ji Zhang, Fei Huang, Chunfeng Yuan, Bing Li, et al. 2024b. Mibench: Evaluating multimodal large language models over multiple images. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22417–22428.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2024c. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*.

Fanqing Meng, Chuanhao Li, Jin Wang, Quanfeng Lu, Hao Tian, Tianshuo Yang, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. 2025. MMIU: Multimodal multi-image understanding for evaluating large vision-language models. In *The Thirteenth International Conference on Learning Representations*.

OpenAI. 2022. Introducing chatgpt.

QwenTeam. 2024. Qwen2.5: A party of foundation models.

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

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Carlota Smith. 2003. *Modes of Discourse*. Cambridge University Press, Cambridge.

Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. 2024. Milebench: Benchmarking MLLMs in long context. In *First Conference on Language Modeling*.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2025. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *The Thirteenth International Conference on Learning Representations*.

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

Xiyao Wang, Yuhang Zhou, Xiaoyu Liu, Hongjin Lu, Yuancheng Xu, Feihong He, Jaehong Yoon, Taixi Lu, Fuxiao Liu, Gedas Bertasius, Mohit Bansal, Huaxiu Yao, and Furong Huang. 2024b. Mementos: A comprehensive benchmark for multimodal large language model reasoning over image sequences. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 416–442. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *Preprint*, arXiv:2412.10302.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, jiayi lei, Quanfeng Lu, Runjian Chen, Peng Xu, Renrui Zhang, Haozhe Zhang, Peng Gao, Yali Wang, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. 2024. MMT-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask AGI. In *Forty-first International Conference on Machine Learning*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024b. Llava-next: A strong zero-shot video understanding model.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Rolf A. Zwaan and Gabriel A. Radvansky. 1998. Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2):162–85.

A Details in Data Filtering

A.1 Stative Verbs

We filter out the samples where any form of a stative verb exists (as displayed in Table 7) in the captions to avoid describing a state. In contrast, we only keep the examples with dynamic verbs to describe actions.

Base Form	Present Participle	3rd Person Singular	Past Tense	Past Participle
wish	wishing	wishes	wished	wished
equal	equaling	equals	equaled	equaled
signify	signifying	signifies	signified	signified
feel	feeling	feels	felt	felt
involve	involving	involves	involved	involved
sense	sensing	senses	sensed	sensed
sound	sounding	sounds	sounded	sounded
detest	detesting	detests	detested	detested
want	wanting	wants	wanted	wanted
see	seeing	sees	saw	seen
forget	forgetting	forgets	forgot	forgot
matter	mattering	matters	mattered	mattered
contain	containing	contains	contained	contained
own	owning	owns	owned	owned
taste	tasting	tastes	tasted	tasted
dislike	disliking	dislikes	disliked	disliked
remember	remembering	remembers	remembered	remembered
suppose	supposing	supposes	supposed	supposed
resemble	resembling	resembles	resembled	resembled
think	thinking	thinks	thought	thought
envy	envying	envies	envied	envied
depend	depending	depends	depended	depended
hate	hating	hates	hated	hated
know	knowing	knows	knew	known
require	requiring	requires	required	required
love	loving	loves	loved	loved
appreciate	appreciating	appreciates	appreciated	appreciated
need	needing	needs	needed	needed
concern	concerning	concerns	concerned	concerned
span	spanning	spans	spanned	spanned
appear	appearing	appears	appeared	appeared
owe	owing	owes	owed	owed
weigh	weighing	weighs	weighed	weighed
disagree	disagreeing	disagrees	disagreed	disagreed
become	becoming	becomes	became	become
fear	fearing	fears	feared	feared
measure	measuring	measures	measured	measured
possess	possessing	possesses	possessed	possessed
like	liking	likes	liked	liked
look	looking	looks	looked	looked
imagine	imagining	imagines	imagined	imagined
mind	minding	minds	minded	minded
belong	belonging	belongs	belonged	belonged
loathe	loathing	loathes	loathed	loathed
lack	lacking	lacks	lacked	lacked
deserve	deserving	deserves	deserved	deserved
mean	meaning	means	meant	meant
promise	promising	promises	promised	promised
believe	believing	believes	believed	believed
prefer	preferring	prefers	preferred	preferred
cost	costing	costs	costed	costed
hope	hoping	hopes	hoped	hoped
recognize	recognizing	recognizes	recognized	recognized
include	including	includes	included	included
support	supporting	supports	supported	supported
understand	understanding	understands	understood	understood
comprise	comprising	comprises	comprised	comprised
agree	agreeing	agrees	agreed	agreed
realize	realizing	realizes	realized	realized
value	valuing	values	valued	valued
seem	seeming	seems	seemed	seemed
hear	hearing	hears	heard	heard
doubt	doubting	doubts	doubted	doubted
consist	consisting	consists	consisted	consisted
smell	smelling	smells	smelled	smelled

Table 7: Full list of stative verbs.

A.2 Threshold Values

Due to the differences in image and text styles across datasets caused by their respective domains, we determined the threshold values for text similarity and image similarity for each dataset through manual inspection and empirical tuning, as shown in the Table 8. For example, in FlintstonesSV, for text similarity, two texts are considered dissimilar if their BERTScore precision and recall are both below 0.98, and their F1 score is below 0.96. Similarly, for image similarity, two images are considered dissimilar if their CLIP similarity score is below 0.94.

	BERTScore			CLIP Similarity
	precision	recall	f1	
FlintstonesSV	<0.98	<0.98	<0.96	<0.94
PororoSV	<0.96	<0.96	<0.95	<0.90
VWP	<0.98	<0.98	<0.97	<0.95
VIST	<0.92	<0.92	<0.90	<0.88

Table 8: The similarity threshold values used in data filtering.

A.3 Data Statistics after Each Filtering Step

In Table 9, we show the statistics of image sequences left after each data filtering step.

	FlintstonesSV	PororoSV	VWP	VIST
Original image sequences	24,433	11,444	12,627	49,700
No stative verbs	10,378	3,114	1,995	20,294
No starting pronoun	10,105	2,952	914	12,216
No similar text	3,092	1,952	815	2,906
No similar image	636	644	809	2,315
No ambiguous image-text	633	535	686	2,284
No repetitive image sequences	633	535	498	1,880
Without No_EVENT	612	535	417	1,292
Final image sequences after manual check	565	395	316	809
Final two-event groups	2104	864	850	3742
Final three-event groups	916	172	208	830

Table 9: The number of image sequences of each step in data filtering process.

B Human Performance Survey

We designed three questionnaires for the human performance survey, corresponding to the three main tasks: event relation inference, sentence ordering, and image ordering. To ensure that participants fully understand the tasks, we provided task instructions and two sample questions at the beginning of each questionnaire (as shown in Figure 4). Additionally, Figure 5 presents sample questions that participants were required to answer. We randomly selected 280 image sequences from TempVS benchmark and collected three responses for each sequence from different annotators. Participants in MT1 were required to complete 30 questions, while participants in the other two ordering tasks completed 20 questions each. The median

completion time was approximately 20 minutes, ensuring that long-time focus did not negatively impact participants' judgment. We recruited 36 annotators (18 females, 18 males) via Prolific, all of whom were proficient in English and had at least a college-level education.

Task Introduction: Multi-event Relation Inference

In this task, you will evaluate the accuracy and consistency of textual statements in relation to a given sequence of images. Your goal is to determine whether the provided statement is fully supported by the visual evidence presented in the images.

Each task will present you with:

- A sequence of images displayed from left to right in the timeline.
- A textual statement describing an event or situation.

The key is to focus on whether the sequence of events described in the text matches the order of events depicted in the image sequence from left to right. If they are consistent, select **True**; otherwise, select **False**.

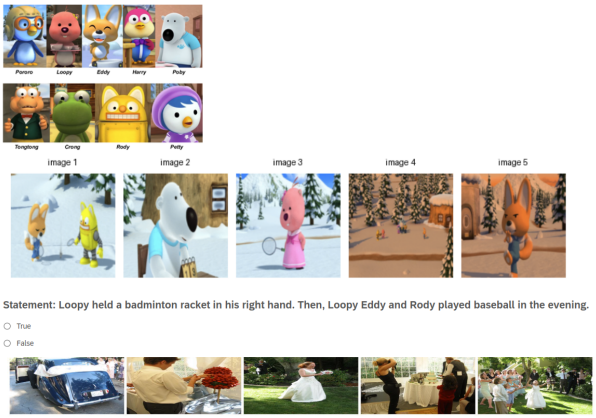
You will see two examples of the tasks in the next page.

Statement: **Betty talked to Barney in a ballroom in front of a sign advertising an awards show after Wilma angrily talked to an unseen person outside.**
Answer: **True**
Explanation: 'Betty talked to barney in a ballroom in front of a sign advertising an awards show" is located in Image 3, 'Wilma angrily talked to an unseen person outside' is located in Image 1, so it is True. Note that Wilma is the woman with red hair, Betty is the woman with black hair, Barney is the man with yellow hair.

Figure 4: An example of task instruction and example question shown in the beginning of human questionnaire

C Prompts

Prompt Template The prompt examples we used for each task are shown below. Our prompt consists of Character Description (only for image sequences from FlintstonesVS and PororoVS), Task Instruction, Question Text, Response Format and Options (only for main tests), and a prefix indicating the beginning of the answer.



Following the text to give the correct order of the shuffled images: The florist set up the wedding early. The photographer took a picture of the cake. The bride waited nervously outside before beginning. It's time to catch the bouquet. They went off by a car.

- A. Image e -> Image b -> Image a -> Image d -> Image c
- B. Image b -> Image c -> Image d -> Image e -> Image a
- C. Image e -> Image c -> Image a -> Image d -> Image b
- D. Image b -> Image d -> Image c -> Image e -> Image a
- E. Image b -> Image d -> Image e -> Image a -> Image c

Figure 5: Example question interface used by the human annotators

GT: Grounding Test
User:
[Character Description] Description of character appearance in the images: Fred is chubby, has black hair, a large nose and wears an orange and black spotted short-sleeved loincloth with a blue scarf. Barney is short, has yellow hair, oval eyes and wears a brown loincloth with a black X-shaped shoelace on the top.
[Task Instruction] Pick the image from the image sequence that accurately represents the event. When making the choice, focus on the evidence presented in the sequence of images from left to right.
[Question Text] The event is: Fred was sitting in a room on a stool.
[Response Format] Submit the right number of the image in the sequence as your answer only without additional reasoning or repetition of the instructions.
The answer is:

MT1: Event Relation Inference

[Character Description] Description of character appearance in the images: Pororo is a gentoo penguin with an orange-yellow beak wearing a tan aviator's helmet and orange goggles.

[Task Instruction] Is the statement completely accurate and consistent with the content in the sequence of images? When making the choice, focus on the evidence presented in the sequence of images from left to right.

[Question Text] The statement is: A snowball dropped down a small snow hill after Pororo threw a snowball.

[Response Format] Submit only the right option letter as your answer, e.g., Option [Letter].

[Options] Options are: A. True; B. False.

The answer is:

MT2: Sentence Ordering

[Task Instruction] You are given an ordered image sequence and several sentences in a random order. Your task is to analyze the content of the sequence of images from left to right and rearrange the sentences into the correct chronological or logical order. Read the image sequence from left to right. Use the images' content to guide your sentence ordering. Avoid assumptions not supported by the image sequence or sentences.

[Question Text] The shuffled sentences are:

Sentence a: John prepares to shoot again and fires.

Sentence b: The shadowy figure is hit with the rifle blast.

Sentence c: John takes aim at a shadowy figure up above.

Sentence d: Ken stops the horses and prepares to leave the stagecoach.

Sentence e: Ken is approaching a house out on the prairie on his stagecoach with his two horses.

[Response Format] Do not provide explanations or repeat the prompt. Select from the following options and your answer should only be in the format: Option [Letter].

[Options]

A. Sentence c -> Sentence b -> Sentence a -> Sentence e -> Sentence d;

B. Sentence b -> Sentence e -> Sentence d -> Sentence a -> Sentence c;

C. Sentence e -> Sentence b -> Sentence c -> Sentence d -> Sentence a;

D. Sentence c -> Sentence a -> Sentence e -> Sentence d -> Sentence b;

E. Sentence d -> Sentence c -> Sentence e -> Sentence b -> Sentence a.

The answer is:

MT3: Image Ordering
[Task Instruction] The text narrates a story or event sequence. Use your vision-language reasoning to reorder the images to reflect the narrative structure. Carefully read the provided text. Focus on the events, actions, and details described to reorder the images logically.The images are labeled in order as Image a, Image b, Image c, Image d, and Image e, and so on if there are more photos.
[Question Text] The events are: Some people came for the family gathering today. The girls enjoyed some fruit. We played on the swings. The boys lounged in the chair. Grandpa put his grandson on his knee.
[Response Format] Select from the following options and your answer should be in the format: 'Option [Letter]'. Respond with the correct option only, avoiding any explanations or repetition.
[Options]
A. Image b -> Image c -> Image d -> Image a -> Image e;
B. Image e -> Image d -> Image c -> Image b -> Image a;
C. Image b -> Image e -> Image d -> Image a -> Image c;
D. Image a -> Image c -> Image e -> Image b -> Image d;
E. Image c -> Image b -> Image a -> Image e -> Image d.
The answer is:

Variations of Each Prompt Component For each component in the prompt, we generate multiple variations (as listed in Table 10). By combining different components, we ultimately generate 328 distinct prompts.

Chain-of-Thought Prompts

Task	CoT Prompt
MT1	Analyze the provided image sequence to determine whether the following statement is True or False. When making the choice, carefully examine the evidence presented in the sequence of images from left to right. First, describe the key details and changes observed in each image. Then, explain how these details support or contradict the given statement. Finally, based on your step-by-step reasoning, conclude whether the statement is True or False. Ensure your response follows this format: the reasoning process should be enclosed within <think> and </think> tags, and the final answer should be enclosed within <answer> and </answer> tags.
MT2	You are presented with an ordered image sequence and several sentences in random order. Your task is to determine the correct order of the sentences based on the context, events, or details observable in the images. Multiple-choice options are provided, with each option representing a possible sequence of the sentences. Think step by step, using the content of the images to guide your reasoning. Avoid assumptions not supported by the image sequence or the sentences. Format your response as follows: Enclose your step-by-step reasoning process within <think> and </think> tags. Enclose the selected option (e.g., Option A, Option B) within <answer> and </answer> tags.
MT3	Order the following images in the correct sequence based on the content of the story. Compare each image with the text description, carefully analyzing the sequence of events to determine the proper order of the unordered images. Response should be in this format: <think> First, examine the text description to identify key events and their chronological order. Next, analyze each image to match it with the corresponding event described in the text. Consider visual cues, actions, and details in the images that indicate the progression of the story. Arrange the images accordingly to reflect the correct sequence of events. </think> <answer> Option [Your Choice] </answer>

Table 11: Chain-of-Thought prompts for different tasks.

D Experiments

D.1 Implementation Details

We evaluate 38 existing MLLMs for multi-event temporal grounding and reasoning, including both proprietary and open-source models. All open-source models are assessed using their official pre-trained versions available on HuggingFace. Detailed configurations of the open-source models evaluated are listed in Table 12. To minimize randomness, we set the temperature to 0. Experiments on open-source models are conducted using a single NVIDIA H100 GPU. For models exceeding 70B parameters, we adopt the mixed precision (FP16), otherwise the full precision. Following Jiang et al. (2024) and Wang et al. (2025), we merge sequential images into a single input, separating them with a thin white band and labeling each with an index (e.g., "Image 1" when images are sequentially ordered, or "Image a" when images are shuffled). Initially, we tested sequentially inputting images one at a time but observed minimal performance differences compared to merging them.

Task Instruction - definition	
MT1	- Is the statement completely accurate and consistent with the content in the sequence of images? - Analyze the provided image sequence to determine whether the following statement is True or False. - Review the sequential images and decide whether the provided statement is True or False. - Is the statement entirely consistent and supported by the images in the sequence? - Examine the visual evidence in the provided sequence of images to determine whether the statement is True or False. - Does the statement fully match the information presented in the ordered images? Taking the content in the image seqeunce into account, can you decide whether the statement is True or False? - Given the multiple images provided in order, can you select the correct answer from True or False considering statement?
MT2	- You are given an ordered image sequence and several sentences in a random order. Your task is to analyze the content of the sequence of images from left to right and rearrange the sentences into the correct chronological or logical order. - Here is an left-to-right image seqeunce and some sentences in random order. Your goal is to determine the correct order of the sentences based on the context, events, or details observable in the images. Consider the visual elements in the sequential images to infer the logical or temporal sequence. - The images are presented in a correct order, but the sentences are not. Your task is to reorder the sentences to match the sequence of events or details in the images from left to right. - You are provided with a sequence of images and several randomly ordered sentences. Your task is to: 1. Understand the context of the image sequence; 2. Identify how each sentence relates to the image; 3. Rearrange the sentences to form a coherent order. - Using your multimodal understanding and reasoning of the ordered image sequence and the randomly shuffled sentences , arrange the sentences in the correct sequence to match the flow of events in the image seqeunce. - Presented is an image sequence in the order from left to right along with several unordered sentences. Your task is to determine the correct sequence of the sentences by analyzing the context, events, or details depicted in the images. Use the visual elements in the images to deduce the logical or chronological order.
MT3	- Rearrange the following images in the correct order based on content in the story. - The provided paragraph describes a sequence of events. Arrange the images in the correct chronological order to match the story. - The images are out of order compared to the text. Identify and reorder them to match the described sequence. - The text narrates a story or event sequence. Use your vision-language reasoning to reorder the images to reflect the narrative structure. - Using your understanding of the text and image content, arrange the images in the correct sequence to match the flow of events in the text. - The paragraph describes events in a specific timeline. Use multimodal reasoning to reorder the images in the correct sequence.
GT	- Based on the event described, select the image that best matches it from the following options. - Pick the image from the image sequence that accurately represents the event. - Which is the most fitting image for the described event? Pick from the choices below. - If you were to illustrate the event, which picture would you use? - Find the image from the sequence of images that represents the event most precisely. - Identify the picture that best represents the given event.
Task Instruction - requirement	
MT1	- When making the choice, focus on the evidence presented in the sequence of images from left to right. - Choose the correct option based on the content in the sequential images from left to right. - Only use the left-to-right content of the image sequence to inform the decision. - Evaluate the statement strictly based on the information shown in the sequence of images from left to right.
MT2	- Read the image sequence from left to right. Use the images' content to guide your sentence ordering. Avoid assumptions not supported by the image seqeunce or sentences. - Understand images from left to right in the order. Focus on the visual content of the images to determine the correct order of the sentences. Make sure the reordered sentences form a clear and coherent narrative or description. - Follow the sequence of images from left to right and use their content to determine the correct sentence order. Do not rely on assumptions that are not supported by the images or sentences. - Interpret the images in their left-to-right order, focusing on their visual details to arrange the sentences correctly. Ensure that the reordered sentences create a logical and coherent narrative or description.
MT3	- Carefully read the provided text. Focus on the events, actions, and details described to reorder the images logically. - Focus on matching the actions and events shown in the images with the details described in the text. - Identify key events and details from the text and use them to determine the proper order of the images. - Compare the images with the text description, focusing on the sequence of events to arrange unordered images correctly.
GT	- Make your choice by considering the evidence shown in the sequence of images from left to right. - Select the correct option using only the content presented in the sequential images from left to right. - Rely on the left-to-right order of the image sequence to guide your decision. - Assess the statement exclusively based on the information depicted in the sequence of images from left to right.
Response Format	
MT1	- No need to give reasoning process. Submit only the right option letter as your answer, e.g., Option [Letter]. - Do not tell the reasons of your decision. Provide the most suitable choice letter in the format of 'Option [Letter]' as your response only. - Do not repeat the prompt or include reasons. Only return the correct option letter in the form of 'Option [Letter]' as your response. - Please exclude explanations in the response. Offer the most proper choice letter in the format of 'Option [Letter]' as your answer only.
MT2	- Do not provide explanations or repeat the prompt. Select from the following options and your answer should only be in the format: Option [Letter]. - Provide your answer from the following choices only in the format: 'Option [Letter]' without explanations or repeating the instructions. - Choose the correct option from the choices provided below and output your answer only as 'Option [Letter]', avoiding any explanations or repetition.
MT3	- Select from the following options and your answer should be in the format: 'Option [Letter]'. Respond with the correct option only, avoiding any explanations or repetition. - Do not provide explanations or restate the question. Provide your answer from the following choices only in the format: Option [Letter]. - Choose the correct option from the choices provided below and submit your answer only as 'Option [Letter]', without any justifications or repetition of the prompt.
GT	- Submit the right number of the image in the sequence as your answer only without additional reasoning or repetition of the instructions. - Provide only the most suitable image number as your response, avoiding any explanations or repetition. - Only return the correct image number in the provided sequence without additional reasoning details or repetition of the instructions.

Table 10: All variations of different components we used to generate the prompts.

HuggingFace Model ID	# Params	Vision Backbone	Base LLM
deepseek-ai/deepseek-vl2-tiny	3.4B	SigLIP-SO400M-384	DeepSeekMoE
deepseek-ai/deepseek-vl2-small	16.1B	SigLIP-SO400M-384	DeepSeekMoE
OpenGVLab/InternVL2_5-1B	0.9B	InternViT-300M-448px-V2_5	Qwen2.5-0.5B-Instruct
OpenGVLab/InternVL2_5-1B-MPO	0.9B	InternViT-300M-448px-V2_5	Qwen2.5-0.5B-Instruct
OpenGVLab/InternVL2_5-8B	8.1B	InternViT-300M-448px-V2_5	internlm2_5-7b-chat
OpenGVLab/InternVL2_5-8B-MPO	8.1B	InternViT-300M-448px-V2_5	internlm2_5-7b-chat
OpenGVLab/InternVL2_5-26B	25.5B	InternViT-6B-448px-V2_5	internlm2_5-20b-chat
OpenGVLab/InternVL2_5-26B-MPO	25.5B	InternViT-6B-448px-V2_5	internlm2_5-20b-chat
OpenGVLab/InternVL2_5-78B	78.4B	InternViT-6B-448px-V2_5	Qwen2.5-72B-Instruct
OpenGVLab/InternVL2_5-78B-MPO	78.4B	InternViT-6B-448px-V2_5	Qwen2.5-72B-Instruct
deepseek-ai/Janus-Pro-1B	1.0B	ViT-L-16-SigLIP-384	DeepSeek-LLM-1.5b-base
deepseek-ai/Janus-Pro-7B	7.0B	ViT-L-16-SigLIP-384	DeepSeek-LLM-7b-base
llava-hf/llava-interleave-qwen-0.5b-hf	0.9B	SigLIP-400M	Qwen1.5-0.5B-Chat
llava-hf/llava-interleave-qwen-7b-hf	8.1B	SigLIP-400M	Qwen1.5-7B-Chat
llava-hf/llava-interleave-qwen-7b-dpo-hf	8.1B	SigLIP-400M	Qwen1.5-7B-Chat
lmms-lab/llava-onevision-qwen2-0.5b-ov	0.9B	siglip-so400m-patch14-384	Qwen2-0.5B
lmms-lab/llava-onevision-qwen2-0.5b-si	0.9B	siglip-so400m-patch14-384	Qwen2-0.5B
lmms-lab/llava-onevision-qwen2-7b-ov	8.0B	siglip-so400m-patch14-384	Qwen2-7B
lmms-lab/llava-onevision-qwen2-7b-si	8.0B	siglip-so400m-patch14-384	Qwen2-7B
lmms-lab/llava-onevision-qwen2-72b-ov-sft	73.2B	siglip-so400m-patch14-384	Qwen2-72B
lmms-lab/llava-onevision-qwen2-72b-si	73.2B	siglip-so400m-patch14-384	Qwen2-72B
llava-hf/LLaVA-NeXT-Video-7B-hf	7.1B	SigLIP-400M	vicuna-7b-v1.5
llava-hf/LLaVA-NeXT-Video-7B-DPO-hf	7.1B	SigLIP-400M	vicuna-7b-v1.5
llava-hf/LLaVA-NeXT-Video-34B-hf	34.8B	SigLIP-400M	vicuna-33b-v1.3
llava-hf/LLaVA-NeXT-Video-34B-DPO-hf	34.8B	SigLIP-400M	vicuna-33b-v1.3
lmms-lab/LongVA-7B	7.9B	SigLIP-400M	Qwen2-7B-Instruct
lmms-lab/LongVA-7B-DPO	7.9B	SigLIP-400M	Qwen2-7B-Instruct
TIGER-Lab/Mantis-8B-Idefics2	8.4B	idefics2-8b	Mistral-7B-v0.1
TIGER-Lab/Mantis-8B-siglip-llama3	8.5B	SigLIP	LLaMA-3-8B
microsoft/Phi-3-vision-128k-instruct	3.8B	CLIP ViT-L/14	Phi-3-mini-128k-instruct
microsoft/Phi-3.5-vision-instruct	4.2B	CLIP ViT-L/14	Phi-3.5-mini-instruct
Qwen/Qwen2-VL-2B	2.2B	CLIP ViT-L/14	Qwen2-1.5B
Qwen/Qwen2-VL-2B-Instruct	2.2B	CLIP ViT-L/14	Qwen2-1.5B
Qwen/Qwen2-VL-7B	7.6B	CLIP ViT-L/14	Qwen2-7B
Qwen/Qwen2-VL-7B-Instruct	7.6B	CLIP ViT-L/14	Qwen2-7B
Qwen/Qwen2-VL-72B	72.7B	CLIP ViT-L/14	Qwen2-72B
Qwen/Qwen2-VL-72B-Instruct	72.7B	CLIP ViT-L/14	Qwen2-72B

Table 12: Details of all evaluated open-source MLLMs: HuggingFace model id, number of parameters, vision backbone, base LLM.

D.2 Evaluation Results of 38 MLLMs

In this section, we present the complete quantitative results of 38 state-of-the-art multimodal large language models (MLLMs) on the TempVS benchmark. The results for the MT1 event relation reasoning task are shown in Table 13, while Table 14 presents the results for the MT2 sentence ordering task and Table 15 reports the results for the MT3 image ordering task. It is important to note that GT represents the overall grounding evaluation, which measures the number of events correctly matched to their corresponding images across the entire benchmark. GT_{strict} denotes the strict grounding evaluation, which calculates the number of image sequences in which every event within a sequence is correctly matched to its corresponding image.

	Two-event Relation Inference				Three-event Relation Inference			
	GT	GT_{strict}	MT	MT|GT_{strict}	GT	GT_{strict}	MT	MT|GT_{strict}
deepseek-vl2-tiny	0.4159	0.2082	0.4971	0.4965	0.4292	0.1052	0.4976	0.5064
deepseek-vl2-small	0.4116	0.1416	0.4311	0.4223	0.4306	0.0669	0.4436	0.4435
InternVL2_5-1B	0.3636	0.1702	0.4102	0.4073	0.3896	0.085	0.3895	0.4270
InternVL2_5-1B-MPO	0.3879	0.1881	0.3727	0.3709	0.4002	0.0953	0.3736	0.3569
InternVL2_5-8B	0.5655	0.3976	0.5426	0.5540	0.6246	0.3278	0.5438	0.5556
InternVL2_5-8B-MPO	0.6230	0.4697	0.5619	0.5736	0.6871	0.4118	0.558	0.5714
InternVL2_5-26B	0.6181	0.4604	0.5705	0.5843	0.6824	0.3958	0.5835	0.6022
InternVL2_5-26B-MPO	0.6521	0.5132	0.6032	0.6212	0.7096	0.4595	0.6212	0.6474
InternVL2_5-78B	0.6555	0.5157	0.5423	0.5532	0.7154	0.4733	0.5679	0.5698
InternVL2_5-78B-MPO	0.7046	0.5881	0.5847	0.5992	0.7771	0.5653	0.6139	0.6258
Janus-Pro-1B	0.2359	0.0273	0.4827	0.4814	0.2558	0.0073	0.4645	0.4259
Janus-Pro-7B	0.2399	0.0429	0.3509	0.3406	0.2555	0.0038	0.3287	0.3929
llava-interleave-qwen-0.5b-hf	0.2172	0.0245	0.4976	0.4779	0.2182	0.0019	0.5036	0.5012
llava-interleave-qwen-7b-hf	0.3521	0.1300	0.5161	0.5239	0.4112	0.0723	0.5007	0.4981
llava-interleave-qwen-7b-dpo-hf	0.3531	0.1358	0.5198	0.5357	0.4102	0.0845	0.5173	0.5319
llava-onevision-qwen2-0.5b-ov	0.304	0.0859	0.4528	0.4545	0.3343	0.0286	0.4807	0.4764
llava-onevision-qwen2-0.5b-si	0.2153	0.0235	0.4496	0.4429	0.2156	0.0013	0.3587	0.3731
llava-onevision-qwen2-7b-ov	0.5176	0.3278	0.5602	0.5804	0.578	0.2604	0.5753	0.5979
llava-onevision-qwen2-7b-si	0.4575	0.2505	0.5292	0.5419	0.4895	0.1403	0.5546	0.5654
llava-onevision-qwen2-72b-ov-sft	0.6215	0.4641	0.5928	0.6213	0.6844	0.405	0.6154	0.6349
llava-onevision-qwen2-72b-si	0.5418	0.3593	0.5296	0.5393	0.6015	0.2804	0.5248	0.5298
LLaVA-NeXT-Video-7B-hf	0.2731	0.0575	0.4603	0.4603	0.2972	0.0135	0.4486	0.4723
LLaVA-NeXT-Video-7B-DPO-hf	0.2718	0.0596	0.4672	0.466	0.2981	0.014	0.4520	0.4615
LLaVA-NeXT-Video-34B-hf	0.2742	0.0652	0.5847	0.5839	0.3053	0.0159	0.5947	0.5763
LLaVA-NeXT-Video-34B-DPO-hf	0.3028	0.0825	0.533	0.5386	0.3289	0.0205	0.5248	0.5263
LongVA-7B	0.3015	0.0874	0.5468	0.5613	0.3138	0.0208	0.5596	0.6169
LongVA-7B-DPO	0.3241	0.1132	0.5319	0.5582	0.3449	0.0386	0.5233	0.535
Mantis-8B-Idefics2	0.3452	0.1218	0.5193	0.533	0.3586	0.041	0.5197	0.5164
Mantis-8B-siglip-llama3	0.2444	0.0443	0.5238	0.5368	0.2392	0.0065	0.5251	0.5833
Phi-3-vision-128k-instruct	0.2226	0.0379	0.5196	0.5219	0.2332	0.0065	0.5132	0.4583
Phi-3.5-vision-instruct	0.2235	0.0400	0.4904	0.4774	0.2316	0.0078	0.4877	0.4828
Qwen2-VL-2B	0.3053	0.088	0.4935	0.4842	0.3188	0.0313	0.4717	0.4923
Qwen2-VL-2B-Instruct	0.3815	0.1738	0.5271	0.5314	0.4003	0.0777	0.5051	0.5052
Qwen2-VL-7B	0.4032	0.1664	0.5122	0.4982	0.4233	0.0899	0.4991	0.4603
Qwen2-VL-7B-Instruct	0.5144	0.3239	0.5397	0.5542	0.5415	0.2124	0.5358	0.5534
Qwen2-VL-72B	0.4906	0.2997	0.5461	0.5738	0.5167	0.1935	0.5429	0.5900
Qwen2-VL-72B-Instruct	0.5078	0.3167	0.5395	0.5643	0.5427	0.2059	0.5563	0.6062
GPT-4o	0.7043	0.6034	0.5827	0.6005	0.7807	0.5704	0.6452	0.6644

Table 13: Full results for MT1: event relation inference

	Ordering Sentences (events)				Ordering Sentences (captions)			
	GT	GT_{strict}	MT	MT| GT_{strict}	GT	GT_{strict}	MT	MT| GT_{strict}
deepseek-vl2-tiny	0.3701	0.0084	0.1953	0.25	0.4415	0.0154	0.1823	0.1786
deepseek-vl2-small	0.3021	0.0037	0.1574	0.1429	0.3735	0.0061	0.1707	0.1818
InternVL2_5-1B	0.3462	0.0089	0.2253	0.1176	0.3589	0.0105	0.2037	0.2105
InternVL2_5-1B-MPO	0.3636	0.0084	0.2168	0.4375	0.3833	0.0099	0.2032	0.3889
InternVL2_5-8B	0.5294	0.0711	0.4589	0.5778	0.5574	0.0765	0.5424	0.6259
InternVL2_5-8B-MPO	0.5747	0.0963	0.5663	0.7377	0.6183	0.1415	0.6289	0.7704
InternVL2_5-26B	0.5765	0.1016	0.5674	0.7358	0.6137	0.1311	0.63	0.7689
InternVL2_5-26B-MPO	0.6089	0.1258	0.6989	0.9079	0.6491	0.1696	0.7693	0.8734
InternVL2_5-78B	0.6121	0.1363	0.6695	0.8494	0.6536	0.1845	0.7109	0.8388
InternVL2_5-78B-MPO	0.6604	0.1842	0.7984	0.9657	0.7064	0.2594	0.8634	0.9639
Janus-Pro-1B	0.2344	0	0.1858	-	0.2497	0.0006	0.1982	0
Janus-Pro-7B	0.2409	0	0.1705	-	0.2623	0	0.1525	-
llava-interleave-qwen-0.5b-hf	0.2161	0	0.2068	-	0.2269	0	0.2065	-
llava-interleave-qwen-7b-hf	0.3363	0.0032	0.2505	0.1667	0.3477	0.0039	0.2704	0
llava-interleave-qwen-7b-dpo-hf	0.3332	0.0047	0.2679	0.3333	0.3409	0.0099	0.2996	0.3333
llava-onevision-qwen2-0.5b-ov	0.2931	0.0005	0.1884	0	0.3118	0.0006	0.1839	0
llava-onevision-qwen2-0.5b-si	0.2106	0	0.1895	-	0.2199	0	0.1944	-
llava-onevision-qwen2-7b-ov	0.4839	0.0447	0.4421	0.4118	0.5118	0.0683	0.4692	0.4758
llava-onevision-qwen2-7b-si	0.426	0.0168	0.4284	0.6875	0.4119	0.0116	0.4537	0.7619
llava-onevision-qwen2-72b-ov-sft	0.5838	0.0984	0.6516	0.8182	0.6179	0.1399	0.7506	0.8661
llava-onevision-qwen2-72b-si	0.5115	0.0411	0.61	0.8077	0.5464	0.0743	0.6691	0.7852
LLaVA-NeXT-Video-7B-hf	0.2645	0	0.1895	-	0.2742	0	0.1823	-
LLaVA-NeXT-Video-7B-DPO-hf	0.2666	0	0.1963	-	0.2736	0	0.1938	-
LLaVA-NeXT-Video-34B-hf	0.2646	0.0005	0.3184	1	0.2831	0	0.3343	-
LLaVA-NeXT-Video-34B-DPO-hf	0.2882	0.0016	0.3095	0	0.2984	0.0006	0.3133	0
LongVA-7B	0.286	0.0016	0.3421	0.6667	0.2909	0.0022	0.353	0.5
LongVA-7B-DPO	0.3089	0.0026	0.3547	0.8	0.2997	0.0028	0.3618	0.6
Mantis-8B-Idefics2	0.3326	0.0011	0.2216	0	0.2868	0.0022	0.2081	0
Mantis-8B-siglip-llama3	0.2406	0.0005	0.2142	0	0.2461	0	0.2153	-
Phi-3-vision-128k-instruct	0.2243	0	0.2316	-	0.2214	0	0.2329	-
Phi-3.5-vision-instruct	0.2262	0	0.2311	-	0.2331	0	0.2539	-
Qwen2-VL-2B	0.2873	0.0004	0.1957	0	0.3274	0.0011	0.1995	0
Qwen2-VL-2B-Instruct	0.3638	0.0058	0.1658	0.0909	0.333	0.0015	0.2014	0.5
Qwen2-VL-7B	0.4592	0.0028	0.3097	0.25	0.4879	0.0076	0.3276	0.5455
Qwen2-VL-7B-Instruct	0.4813	0.0342	0.4253	0.6462	0.5098	0.0441	0.446	0.6125
Qwen2-VL-72B	0.4539	0.0258	0.4363	0.7143	0.4914	0.0435	0.5319	0.7215
Qwen2-VL-72B-Instruct	0.4756	0.0368	0.4632	0.6429	0.5085	0.0496	0.5507	0.7
GPT-4o	0.6708	0.1863	0.534	0.5389	0.7231	0.2357	0.615	0.5532

Table 14: Full results for MT2: sentence ordering with events/captions.

	Ordering Images (events)		Ordering Images (captions)	
	MT	MT| GT_{strict}	MT	MT| GT_{strict}
deepseek-vl2-tiny	0.2042	0.4286	0.2072	0.125
deepseek-vl2-small	0.1663	0	0.1553	0
InternVL2_5-1B	0.2107	0.1429	0.1997	0.0625
InternVL2_5-1B-MPO	0.2052	0.1538	0.2002	0.2667
InternVL2_5-8B	0.2766	0.2783	0.2651	0.4113
InternVL2_5-8B-MPO	0.2986	0.3841	0.318	0.3816
InternVL2_5-26B	0.2666	0.2727	0.3165	0.3589
InternVL2_5-26B-MPO	0.344	0.3971	0.3949	0.4369
InternVL2_5-78B	0.311	0.4	0.3849	0.4709
InternVL2_5-78B-MPO	0.4099	0.4883	0.5382	0.6966
Janus-Pro-1B	0.2247	-	0.2227	0
Janus-Pro-7B	0.2092	-	0.2097	-
llava-interleave-qwen-0.5b-hf	0.2017	-	0.2077	-
llava-interleave-qwen-7b-hf	0.2087	0.1667	0.1997	0.2222
llava-interleave-qwen-7b-dpo-hf	0.2167	0.3846	0.2127	0.1905
llava-onevision-qwen2-0.5b-ov	0.1942	1	0.2082	0
llava-onevision-qwen2-0.5b-si	0.2032	-	0.1902	-
llava-onevision-qwen2-7b-ov	0.2132	0.1429	0.2157	0.1983
llava-onevision-qwen2-7b-si	0.2077	0.35	0.2047	0.4375
llava-onevision-qwen2-72b-ov-sft	0.2756	0.3182	0.2906	0.3647
llava-onevision-qwen2-72b-si	0.2546	0.3553	0.2461	0.2677
LLaVA-NeXT-Video-7B-hf	0.2102	-	0.2132	-
LLaVA-NeXT-Video-7B-DPO-hf	0.2067	-	0.2082	-
LLaVA-NeXT-Video-34B-hf	0.1977	0	0.1997	-
LLaVA-NeXT-Video-34B-DPO-hf	0.1887	0.3333	0.1932	0
LongVA-7B	0.1952	-	0.1902	-
LongVA-7B-DPO	0.1962	1	0.1937	0
Mantis-8B-Idefics2	0.1862	0	0.1922	0.5
Mantis-8B-siglip-llama3	0.2002	0	0.1987	-
Phi-3-vision-128k-instruct	0.1857	-	0.1897	-
Phi-3.5-vision-instruct	0.1922	-	0.1832	-
Qwen2-VL-2B	0.1917	-	0.1859	0
Qwen2-VL-2B-Instruct	0.1438	0.125	0.1483	1
Qwen2-VL-7B	0.1917	0	0.2067	0.1818
Qwen2-VL-7B-Instruct	0.2312	0.2041	0.2456	0.3585
Qwen2-VL-72B	0.2416	0.3548	0.2406	0.4561
Qwen2-VL-72B-Instruct	0.2651	0.4727	0.2811	0.4737
GPT-4o	0.2257	0.2353	0.2297	0.2349

Table 15: Full results for MT3: image ordering with events/captions.

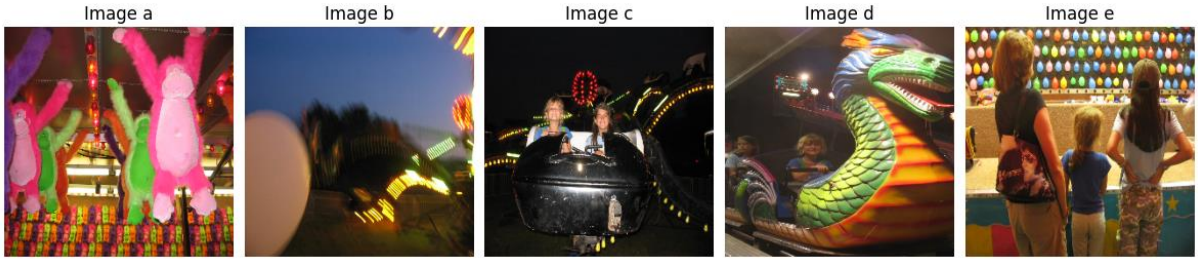
E Additional Examples in TempVS

We provide more examples in TempVS benchmark.

Captions: Fred is in his car driving somewhere. Fred is holding groceries at a fence outside. He puts the groceries on the fence and then jumps over the gate. Fred is standing in the doorway with two bags of groceries and yelling someone in the house. Fred is in the kitchen. He is emptying a bag of groceries on the table. He is speaking to someone off-screen. Wilma is standing in the room. She is speaking while having her elbows bent.

Events: Fred drove in his car. Fred put the groceries on the fence. Fred stood in the doorway with two bags of groceries. Fred emptied a bag of groceries on the table. Wilma stood in the room and spoke to someone.

Correct image sequence: **Image c -> Image e -> Image d -> Image a -> Image b**



Captions: We had fun riding the black roller coaster at the fair. The lights lit up the night and the rides made us all dizzy. The dragon coaster was mom's favorite. The arcade games had the funniest stuffed monkeys as prizes. We threw a million darts trying to win one!

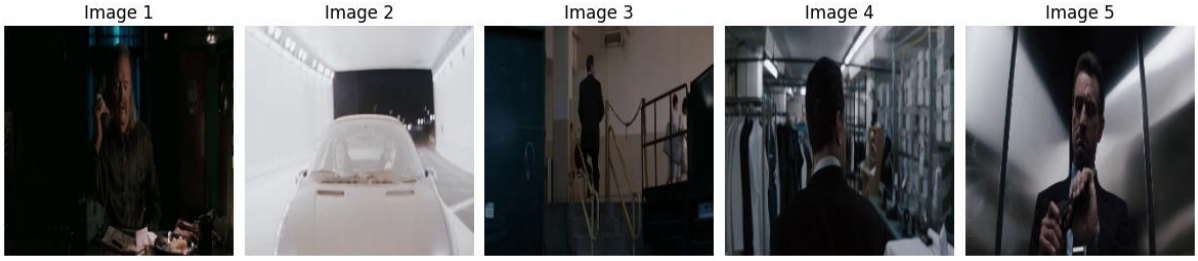
Events: We took the black roller coaster. The lights lit up the night. Mom favored the dragon coaster. The arcade games had stuffed monkeys as prizes. We threw the darts.

Correct image sequence: **Image c -> Image e -> Image d -> Image a -> Image b**

Event a: Eddy used a screwdriver on a green frog toy.
Event b: Loopy walked to a house.
Event c: Eddy turned around and waved to Loopy.
Event d: Loopy came in with a basket.
Event e: Loopy walked away from Crong and Pororo.

Caption a: Eddy is using a screwdriver on a green frog toy.
Caption b: Loopy is walking to a house surrounded by snowed fields and trees.
Caption c: Eddy turns around and waves to Loopy. Rody, Eddy and Loopy are together in a room.
Caption d: Loopy comes in with a basket. Loopy and Rody wave hands to each other.
Caption e: Loopy walks away and Crong and Pororo wave hands to loopy.

Correct sentence sequence: **Sentence e -> Sentence b -> Sentence d -> Sentence c -> Sentence a**



Event a: Robert found many different shops and services inside the building.
Event b: Jon sat in his room planning his next move.
Event c: Amy and Robert came in their car to meet Jon.
Event d: Robert loaded his gun in the lift.
Event e: Robert entered the building.

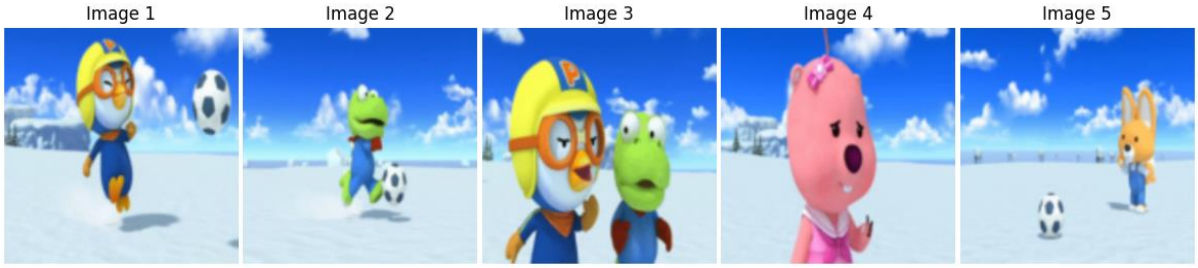
Caption a: Robert finds many different shops and services inside building. He is confused.
Caption b: Jon is sitting in his room, planning his next move.
Caption c: Amy and Robert are coming in their car to meet Jon.
Caption d: Robert gets inside lift and loads his gun. He intends to kill Jon as soon as find him.
Caption e: Robert goes inside the building and begins to search for Jon.

Correct sentence sequence: **Sentence b -> Sentence c -> Sentence e -> Sentence a -> Sentence d**

True: The King and his hunting party pursued the deer after the parents told Anna they were in the middle of the King's forest.
False: The parents told Anna they were in the middle of the King's forest after the King and his hunting party pursued the deer.

True: The King and his hunting party pursued the deer. Earlier, Anna sat with her parents beneath a large oak tree.
False: Anna sat with her parents beneath a large oak tree. Earlier, the King and his hunting party pursued the deer.

True: Anna sat with her parents beneath a large oak tree. Anna asked her parents where they were. The King and his hunting party pursued the deer.
False: The King and his hunting party pursued the deer. Anna sat with her parents beneath a large oak tree. Anna asked her parents where they were.



True: Pororo kicked the ball high. Crong received a ball from Pororo.
False: Crong received a ball from Pororo. Pororo kicked the ball high.

True: Pororo kicked the ball high. Then, Loopy failed to kick the ball.
False: Loopy failed to kick the ball. Then, Pororo kicked the ball high.

True: Pororo and Crong shouted to Loopy before Loopy failed to kick the ball, and after that, Eddy kicked the ball.
False: Eddy kicked the ball before Pororo and Crong shouted to Loopy, and after that, Loopy failed to kick the ball.

True: Granpa made some last-minute repairs before the kids had a tickle fight in the family room.
False: The kids had a tickle fight in the family room before Granpa made some last-minute repairs.

True: The mother, father, and son listened to a story at dinner. Earlier, Granpa made some last-minute repairs.
False: Granpa made some last-minute repairs. Earlier, the mother, father, and son listened to a story at dinner.

True: First, Granpa made some last-minute repairs. Second, the kids had a tickle fight in the family room. Third, Nina took out a chocolate cake.
False: First, Nina took out a chocolate cake. Second, Granpa made some last-minute repairs. Third, The kids had a tickle fight in the family room.

True: Wilma and betty talked in the living room. Earlier, Fred sat at the kitchen table with a boy.
False: Fred sat at the kitchen table with a boy. Earlier, Wilma and Betty talked in the living room.

True: Fred and Barney talked to each other while driving in the car after Fred choked the men in the blue dress with a cap in the dining room.
False: Fred choked the men in the blue dress with a cap in the dining room after Fred and Barney talked to each other while driving in the car.

True: Fred choked the men in the blue dress with a cap in the dining room. Fred and Barney talked to each other while driving in the car. Wilma and Betty talked in the living room.
False: Fred and Barney talked to each other while driving in the car. Fred choked the men in the blue dress with a cap in the dining room. Wilma and Betty talked in the living room.