MOVIE FACTS AND FIBS (MF^2): A BENCHMARK FOR LONG MOVIE UNDERSTANDING

Emmanouil Zaranis[1,2]*, **António Farinhas**[1,2]*, **Saul Santos**[1,2]*, **Beatriz Canaverde**[1,2]*, **Miguel Moura Ramos**[1,2]*, **Aditya K Surikuchi**[3], **André Viveiros**[1,2], **Baohao Liao**[4], **Elena Bueno-Benito**,[5] **Nithin Sivakumaran**[6], **Pavlo Vasylenko**[1,2], **Shoubin Yu**[6], **Sonal Sannigrahi**[1,2], **Wafaa Mohammed**[4], **Ben Peters**[18], **Danae Sánchez Villegas**[7], **Elias Stengel-Eskin**[17], **Giuseppe Attanasio**[2], **Jaehong Yoon**[16], **Stella Frank**[7,8], **Alessandro Suglia**[9], **Chrysoula Zerva**[1,2,13], **Desmond Elliott**[7,8], **Mariella Dimiccoli**[5,15], **Mohit Bansal**[6], **Oswald Lanz**[10,14], **Raffaella Bernardi**[10,14], **Raquel Fernández**[3,12], **Sandro Pezzelle**[3,12], **Vlad Niculae**[4,12], **André F. T. Martins**[1,2,11,13]
[1]Instituto Superior Técnico, Universidade de Lisboa [2]Instituto de Telecomunicações
[3]ILLC, University of Amsterdam [4]Language Technology Lab, University of Amsterdam
[5]Institut de Robòtica i Informàtica Industrial, CSIC-UPC [6]UNC Chapel Hill
[7]University of Copenhagen [8]Pioneer Center for AI [9]University of Edinburgh
[10]Free University of Bozen-Bolzano [11]TransPerfect [12]ELLIS Unit Amsterdam
[13]ELLIS Unit Lisbon [14]ELLIS Unit Trento [15]ELLIS Unit Barcelona
[16]Nanyang Technological University [17]The University of Texas at Austin [18]INESC-ID
{emmanouil.zaranis,andre.t.martins}@tecnico.ulisboa.pt

ABSTRACT

Holistic understanding of long-form video remains a challenge for vision-language models (VLMs). Unfortunately, current benchmarks cannot easily capture this limitation, since they mostly focus on "needle-in-a-haystack" details, rewarding context-insensitive retrieval over deep comprehension. Others rely on large-scale, semi-automatically generated questions (often produced by language models themselves) that are easier for models to answer but fail to reflect genuine understanding. In this paper, we address this gap by introducing **MF^2**, a new benchmark to evaluate how well models are able to comprehend, consolidate, and recall key narrative information from full-length movies (**50-170 minutes long**), requiring integration of *both* visual and language modalities. MF^2 includes over 50 full-length, **open-licensed** movies, each paired with manually constructed sets of claim pairs—one true (*fact*) and one plausible but false (*fib*), totalling over 850 pairs. These claims target core narrative elements such as character motivations and emotions, causal chains, and event order, and refer to **memorable moments** that humans can recall without rewatching the movie. Our experiments demonstrate that both open-weight and closed state-of-the-art models fall well short of human performance. Despite the relative ease of the task for humans who can effectively retain and reason over critical narrative information, current VLMs lack this ability and thus struggle.

1 INTRODUCTION

Vision-language models (VLMs) have demonstrated strong performance across a wide range of tasks involving both images and videos (Deitke et al., 2024; Chen et al., 2024; Liu et al., 2024b; Zhang et al., 2024; Bai et al., 2025b; Zhang et al., 2025; Xu et al., 2025; Li et al., 2025). As these models continue to scale and improve, a natural next frontier lies in long-form video understanding, essential for real-world applications such as education, storytelling, and other types of narrative video analysis—where success depends on integrating and reasoning over information that unfolds over extended periods.

*Equal contribution.

However, current video evaluation benchmarks present several shortcomings. Most rely on fairly short content (Lei et al., 2018; Xiao et al., 2021; Wu et al., 2021; Parmar et al., 2024; Rawal et al., 2024; Qiu et al., 2024; Fang et al., 2024), making them unsuitable for long-form analysis. When longer videos are available (Huang et al., 2020; Song et al., 2024; Chandrasegaran et al., 2024; Ataallah et al., 2024; Wang et al., 2024a; Fu et al., 2024; Wu et al., 2024), video understanding is often limited to retrieval capabilities. Such "needle-in-a-haystack"-like test suites (Kamradt, 2024; Wang et al., 2025a; 2024c; Zhao et al., 2025) focus on peripheral or low-level details such as *"What color is the liquid inside the bucket in the painting?"* (Wu et al., 2024). Success at these tasks may highlight strong cache-then-retrieve capabilities, but says little about abstractive, holistic understanding of long narrative components.[1] Moreover, existing non-retrieval long-form evaluations treat video content uniformly, ignoring the relative importance of different segments. While their synthetically generated questions may align with the overall narrative, they are not explicitly designed to assess core narrative elements or the memorization of critical content (Chandrasegaran et al., 2024). However, a core ability humans exhibit when consuming long video content is naturally identifying and memorizing salient, narrative key points. Rather than relying on questions that treat all video segments as equally important, regardless of their memorability or narrative salience, we argue that effective evaluation should probe whether VLM representations capture and maintain in "memory" the information that is prioritized by humans. This information should align with meaningful narrative moments that humans would naturally recall, without rewatching the video content (Zacks et al., 2009).

In this paper, we introduce **MF**^{2}, a benchmark to evaluate holistic narrative comprehension of long-form video. Building on existing literature (Huang et al., 2020; Song et al., 2024; Ataallah et al., 2024), we source MF^2 from movies, i.e., information-rich content with identifiable turning points that shape the narrative trajectory (Papalampidi et al., 2019; 2020), such as emotional arcs or causal relationships between characters and events. In total, we collect and release 53 open-licensed movies with an average and maximum duration of 88 and 170 minutes, respectively. We design our holistic tests around human perception and annotation of salient elements. Specifically, we manually curate over 850 claim pairs concerning single scenes, multiple scenes, or entire movies—all autonomously identified by annotators as important narrative elements. We follow Karpinska et al. (2024) and craft claims in pairs—one true statement (a *fact*) and one plausible but false counterpart (a *fib*) (see Fig. 1 and Appendix F for examples). For each such pair, models assess claims independently and must correctly identify both the true and false statements. Compared to prior work, this design i) overcomes biases due to option ordering and poorly constructed distractors of multiple-choice questions (Li & Gao, 2024; Loginova et al., 2025; Singh et al., 2025; Molfese et al., 2025), ii) allows for isolated statement-level reasoning and truth prediction, and does not involve contrastive reasoning between paired alternatives, and iii) avoids the use of automatic judges, which may struggle with rich long-form video (Bavaresco et al., 2025; Liu & Zhang, 2025; Ye et al., 2025b). Through an extensive experimental setup including different conditions, input configurations, and modalities, MF^2 proves challenging for both leading open-weight and commercial VLMs, which lag significantly behind human performance. While retrieval-oriented research demonstrates positive outcomes on increasingly longer contexts, our findings suggest that such learning may be shallower than expected. This gap calls for new research and development focused on holistic understanding of memorable facts in long-form, information-rich video content.

Our contributions are as follows:

1. We introduce MF^2, a benchmark and evaluation framework designed for evaluating VLMs holistic understanding of long video content. MF^2 comprises 53 openly-licenses movies alongside 850 human-crafted, high-quality claim pairs for unbiased evaluation. To facilitate reproducibility and support future research, we release MF^2's data and code under a CC-BY 4.0 license. We release the raw video too, preventing external links that become inaccessible over time.[2]

[1] Cognitive psychology research has demonstrated humans consuming continuous video stream show the opposite—we are generally able to follow narrative structures but fail to notice low-level cues (Levin & Simons, 1997).

[2] We ensure reproducibility by releasing the codebase and the dataset at https://github.com/deep-spin/MF2 and https://huggingface.co/datasets/sardinelab/MF2 respectively.

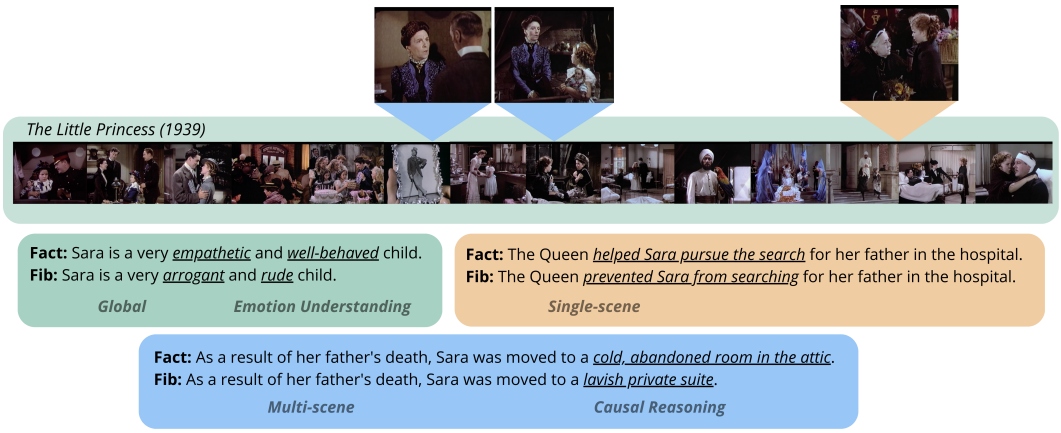


Figure 1: Illustration of three claim pairs (each with a *fact* and a *fib*) from the movie "The Little Princess". Our claims target memorable events, focusing on key turning points of the narrative such as emotional arcs and causal relationships between characters, and require reasoning across different granularities (single-scene, multi-scene and global).

Table 1: Comparison of video datasets across different aspects. MC stands for multiple-choice and OE for open-ended questions.

Dataset	Avg. Duration (mins)	Annotation	Evaluation Format	Source Availability
CausalChaos (Parmar et al., 2024)	-	Auto & Manual	MC & OE	Source link not available
CinePile (Rawal et al., 2024)	2.67	Auto & Manual	MC	YouTube links
EgoSchema (Mangalam et al., 2023)	3.00	Auto & Manual	MC	Videos
ViMuL (Shafique et al., 2025)	4.52	Auto & Manual	MC & OE	Videos
EgoPlan-Bench2 (Qiu et al., 2024)	up to 5	Auto & Manual	MC	Videos
LongVideoBench (Wu et al., 2024)	7.89	Manual	MC	Videos
Video-MMMU (Hu et al., 2025)	8.44	Manual	MC	Videos
MovieChat-1K (Song et al., 2024)	9.40	Manual	MC & OE	Videos
MLVU (Zhou et al., 2024)	12.00	Auto & Manual	MC & OE	Videos
Neptune (Nagrani et al., 2025)	up to 15	Auto & Manual	MC & OE	Videos
Video-MME (Long) (Fu et al., 2024)	39.76	Manual	MC	YouTube links
HourVideo (Chandrasegaran et al., 2024)	45.70	Auto & Manual	MC	Videos
InfiniBench (Ataallah et al., 2024)	52.59	Auto & Manual	MC & OE	Key frames
LVBench (Wang et al., 2024a)	68.35	Manual	MC	YouTube links
MF^2	**88.33**	**Manual**	**Claim pairs**	**Videos**

2. We perform an extensive evaluation of state-of-the-art open and closed models as well as a human evaluation to establish upper-bound performance, revealing a notable performance gap between models and humans.

2 MF^2: Movie Facts and Fibs

MF^2 includes 53 full-length, open-licensed movies, each accompanied by subtitles, and 868 human-authored contrastive claim pairs. Each pair tests whether a model can distinguish true from false information based on its understanding of the story. Fig. 1 shows some examples (see more in Appendix F). We now describe the dataset construction process in detail, covering movie selection (§2.1), annotation methodology including claim categorization and granularity (§2.2), and human quality control procedures used to filter ambiguous or low-quality claims (§2.3). Fig. 2 provides an overview of these three stages.

2.1 Movie Selection and Subtitles

We started by collecting a pool of movies from the Internet Archive,[3] an online repository of open-licensed media. We specifically selected titles released under the Public Domain 1.0 license to ensure

[3] https://archive.org

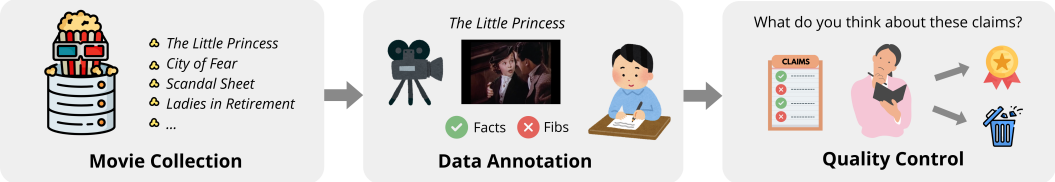


Figure 2: Dataset construction process involving three main stages: movie collection, data annotation, and quality control.

legal reusability and support open-access research. To reduce the risk of data contamination in modern foundation models (Jacovi et al., 2023), we focused on older films released between 1920 and 1970, prioritizing those with limited online visibility, measured by the number of user reviews on IMDb. We sourced original-language subtitles—the majority of which are in English—from OpenSubtitles.org,[4] a widely used platform that provides subtitles for a large collection of movies, TV shows, and other video content. This process yielded a final collection of 53 full-length movies with an average duration of 88.33 minutes, each accompanied by audio and aligned subtitles. Notably, the collection spans a wide temporal range of approximately five decades, a substantial span for capturing social norms across time, and covers a broad variety of genres, as detailed in Appendix A.

2.2 Data Annotation

Unlike prior benchmarks that rely on semi-automatically generated questions, often produced by language models and potentially reflecting model-specific biases, we opted for fully manual annotations, *prioritizing quality over quantity*. The annotation process involved 26 annotators (see Appendix A for details), all of whom are co-authors of this work, who watched the full movies, identified key narrative elements, and constructed pairs of constrastive claims: one factually correct statement (*fact*) and one minimally altered, false counterpart (*fib*). Following Karpinska et al. (2024), annotators were instructed to minimize lexical differences between the *fact* and the *fib*, changing only the parts needed to flip the truth value. The annotation guidelines are presented in Appendix D. This contrastive formulation serves two purposes: *(i)* it isolates the specific narrative element being tested, reducing the chance that models rely on superficial cues (e.g., sentence length, structure, or other lexical patterns); and *(ii)* it simplifies quality control (see §2.3) by making inconsistencies easier to detect.

Claim granularity. To capture different levels of reasoning, annotators labeled each *fact* according to the granularity required to verify its truth: *(i) single-scene*: answerable using information from one scene; *(ii) multi-scene*: requiring integration across multiple scenes; and *(iii) global*: relying on high-level understanding that spans the full movie, including accumulated or inferred information (cannot be easily tied to distinct scenes). As shown in Fig. 3 (left), the dataset includes a balanced distribution of single-scene and multi-scene *facts* (with a smaller proportion requiring global reasoning). Importantly, all claims test long-form comprehension irrespectively of the reasoning granularity: while global claims require reasoning across the entire movie, key events can also unfold within single or multiple scenes. Even single-scene claims are non-trivial, as they assess whether models can extract and retain salient localized information. While humans naturally focus on important elements, models may lack this ability .

Comprehension dimensions. In addition to the reasoning granularity, annotators also labeled each claim pair with one or more comprehension dimensions, indicating the specific aspects of narrative understanding being tested. These dimensions, informed by prior work (Xiao et al., 2021; Zhang et al., 2023b; Wang et al., 2024a), are defined in Appendix A, with their distribution is shown in Fig. 3 (right). Annotators could choose multiple dimensions for the same claim.

Grounding Timeframes For a subset of single-scene claim pairs (12%), annotators were instructed to mark the visual segment (timeframe) of the movie required to answer the claim pair correctly, providing a basis for evaluating models' use of relevant visual information.

[4] https://www.opensubtitles.org

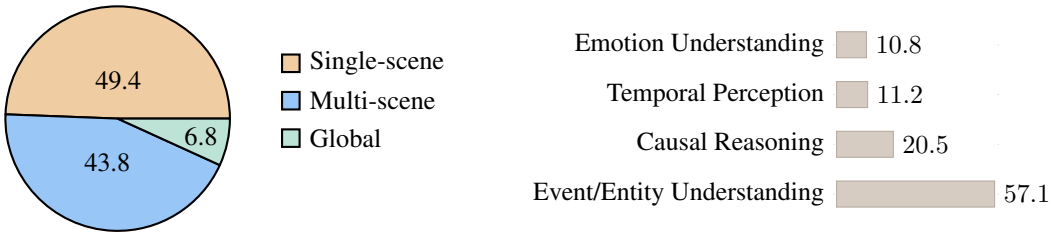


Figure 3: Distribution of claim pairs across reasoning granularities (**left**) and comprehension dimensions (**right**).

2.3 Quality Control

We conducted a human evaluation stage to establish a human baseline for model comparison (see § 3), which was also used to collect feedback on the quality of claims. For this round, annotators first selected a subset of movies they had not previously seen during the data annotation stage. After watching a movie, they classified the corresponding claims as either true or false using a custom annotation interface (see Appendix D for an example and full guidelines). *Claims were presented one at a time, and annotators were required to respond based solely on memory.* To support the identification of problematic claims, we encouraged annotators to leave comments whenever a claim was ambiguous, poorly phrased, open to interpretation, or too fine-grained to be meaningfully tied to narrative understanding (e.g., needle-in-a-haystack claims). The annotation guidelines emphasized the importance of paying close attention while watching the movie, as many claims require subtle reasoning or contextual understanding. Importantly, annotators were instructed not to use any external tools or take notes, ensuring that all responses reflected natural human memory and comprehension.

An optional second stage allowed annotators to revisit their previous responses with access to the movie. This stage was used exclusively to collect additional comments for validation: annotators used it to revise earlier answers after reflecting on the full context of a claim pair.

As part of filtering, two annotators reviewed all comments left during the stages described above. Without watching the corresponding movies, and solely based on the comments left, they identified problematic claims and removed them from the dataset. Importantly, no claims were rewritten at this stage. They were either accepted or discarded. This filtering step resulted in the removal of 104 pairs of claims, yielding a cleaner set of 868 high-quality pairs.

3 Experimental Setup

This section describes the setup used to evaluate a range of VLMs on the MF^2 benchmark. Our experiments include both closed and open-weight models, tested across multiple input modalities using a standardized evaluation protocol.

Modalities. We evaluate all models under a vision-language setup, where they receive visual input in the form of sampled movie frames. We also experiment with providing subtitles as additional input. For the ablation studies on Gemini 2.5 (see §4.2), we test two other configurations: one that includes movie synopses, and another that provides only the movie title and release year.

Baselines. We cover a wide range of state-of-the-art VLMs. Among the closed models, we include GPT-4o (OpenAI et al., 2024), Gemini 2.5 Pro (Team et al., 2023) and Claude 3.7 Sonnet (Anthropic, 2024). Our open-weight models include models from the QwenVL (Bai et al., 2025b;a) and InternVL (Zhu et al., 2025; Wang et al., 2025b) families, Ovis2 (Lu et al., 2024), Gemma3 (Team et al., 2025), GLM-4.5V (Team et al., 2026) and others. We also include models specialized for long videos, such as LongVILA-R1 (Chen et al., 2025), MiMo-VL-7B-RL (Xiaomi, 2025). For all models except GPT-4o, we first downsample videos to 1 frame per second, following each model's preprocessing approach. From these frames, we then uniformly sample a subset, adjusting the number of frames based on each model's maximum context length, original training settings, and GPU

Table 2: Performance of both open-weight and closed models on MF^2. We report pairwise and standard accuracy for video inputs with and without subtitles. Best values overall are **bolded**; best per group are underlined. References for each model are included. Asterisk (*) denotes models that failed to follow the instruction prompt in a substantial fraction of samples and their results are therefore not reported. Numbers in parentheses indicate the number of active parameters.

Method	#Params	#Frames	Pairwise Acc. (%)		Acc. (%)	
			w/o subs	w/ subs	w/o subs	w/ subs
Baselines						
Random	-	-	25.0	25.0	50.0	50.0
Human	-	-	-	84.1	-	90.5
Closed Models						
GPT-4o (OpenAI et al., 2024)	-	50	18.8	46.8	55.2	71.4
Claude 3.7 Sonnet (Anthropic, 2024)	-	100	3.8	44.6	51.4	71.5
Gemini 2.5 Pro (Team et al., 2023)	-	120	**37.2**	**60.6**	**64.2**	**77.6**
Open-weight Models						
mPLUG-Owl3* (Ye et al., 2025a)	7B	-	-	-	-	-
VideoLLaMA3 (Zhang et al., 2025)	7B	180	20.5	33.5	57.0	62.7
Qwen2.5-VL (Bai et al., 2025b)	7B	180	24.6	32.8	56.7	62.0
LLaVA-Video (Zhang et al., 2024)	7B	64	6.6	19.0	51.7	57.8
LongVILA-R1 (Chen et al., 2025)	7B	180	11.5	16.9	50.1	56.6
MiMo-VL-7B-RL (Xiaomi, 2025)	7B	256	25.81	38.59	57.55	65.21
Kangaroo* (Liu et al., 2024a)	8B	-	-	-	-	-
InternVL3 (Zhu et al., 2025)	8B	64	10.9	36.9	53.1	64.6
LLaVAOneVision1.5 (An et al., 2025)	8B	32	4.9	26.4	51.0	60.9
Molmo2 (Clark et al., 2026)	8B	256	28.3	43.0	59.0	67.7
Phi4* (Microsoft et al., 2025)	14B	-	-	-	-	-
Aria (Li et al., 2025)	25B(3.9B)	64	2.53	33.41	50	63.25
Gemma3 (Team et al., 2025)	27B	64	31.5	42.9	61.2	68.1
Qwen3VL-IT (Bai et al., 2025a)	30B	180	19.2	42.3	56.9	68.6
Ovis2 (Lu et al., 2024)	34B	10	18.8	45.6	53.3	69.5
InternVL3.5 (Wang et al., 2025b)	38B	64	26.6	46.2	60.0	70.3
Qwen2.5-VL (Bai et al., 2025b)	72B	180	29.7	45.9	58.8	70.4
LLaVA-Video (Zhang et al., 2024)	72B	64	15.6	41.8	54.6	69.1
InternVL3 (Zhu et al., 2025)	78B	64	22.1	51.3	58.0	72.7
GLM-4.5V-FP8 (Team et al., 2026)	108B	256	26.61	31.33	55.0	62.44

memory constraints.[5] For GPT-4o, frames are uniformly sampled directly from the original videos without prior downsampling. The exact number of frames sampled for each model is reported in Table 2. To extract predictions, we use regular expressions to identify True/False answers in the model outputs, selecting either the first or last valid match depending on the prompt structure. We include all prompt templates and answer parsing details in Appendix E.1 for reproducibility. We also include a human baseline where evaluators judged claims based on their memory, without rewatching scenes (see §2.3).

Evaluation protocol. We report two metrics: *(i)* pairwise accuracy, which measures how often models correctly classify both the true and the false claim in a pair (i.e., they receive credit only if both are labeled correctly; no points are awarded for partial correctness); and *(ii)* standard accuracy, which is computed over individual claims. The random baselines are 25% and 50%, respectively. Following prior work (Karpinska et al., 2024), both models and human annotators see and evaluate each claim independently, without access to the paired structure during prediction. Pairwise accuracy is computed post-hoc by grouping predictions from the same pair. We adopt pairwise accuracy as our

[5]Note that models always receive uniformly sampled frames from the full movie—not targeted scene windows. They must process the entire movie and transcript to identify relevant content, irrespectively of the reasoning granularity of the claim.

Table 3: Pairwise accuracy across different input configurations. Values in parentheses indicate the performance difference of the targeted setting with respect to the corresponding full-movie setting (frames drawn from the entire movie). Positive and negative deltas denote performance gains and drops, respectively. Results are computed on a subset of 103 single-scene claim pairs.

Model	Subs-only	Video-only (full)	Video+Subs (full)	Video-only (targeted)	Video+Subs (targeted)
Gemini 2.5 Pro	56.7	36.1	56.3	49.4 (+13.3)	74.0 (+17.7)
Gemma3-27B	39.8	27.2	46.6	36.9 (+9.7)	49.5 (+2.9)
Qwen3VL-30B	30.1	12.6	37.9	22.3 (+9.7)	52.4 (+14.5)
InternVL3.5-38B	51.5	29.1	51.5	41.7 (+12.6)	53.4 (+1.9)
Qwen2.5VL-72B	37.9	34.0	44.7	29.1 (-4.9)	53.4 (+8.7)
InternVL3-78B	54.4	27.2	49.5	40.8 (+13.6)	66.0 (+16.5)

primary metric, as it more accurately reflects models' ability to distinguish true from false claims; a discussion for this particular choice is provided in Appendix B.

4 Results and Analysis

In this section, we first present the main experimental results (§4.1), followed by ablation studies (§4.2) analyzing model performance across input modalities, reasoning granularities, and comprehension dimensions, and assessing the necessity of visual information.

4.1 Main Results

In Table 2, we report both standard and pairwise accuracy for humans, open-weight, and closed models across two input modalities: video-only and video with subtitles. Our results reveal that:

Both open-weight and closed models fall significantly short of human performance. Among the closed models, Gemini 2.5 Pro achieves the highest scores, with a pairwise accuracy of 60.6%, followed by the open-weight InternVL3-72B, which performs 9.3% lower, when evaluated on both video and subtitles. Despite their fair perfomance, *both models rank significantly behind humans*, with a 24.1% absolute gap. Among small scale models ($\leq$9B), Molmo2 achieves the highest performance, outperforming the random baseline by 18% in the video with subtitles setting, while InternVL3.5 demonstrates the best results within the medium scale category (25B–38B). These findings underscore the difficulty of the task for current models, but also highlight humans' superior ability to retain and reason over critical narrative information.

Subtitles provide a critical signal beyond visual information alone. When models rely on visual information alone, spanning the entire movie, performance remains far lower than when subtitles are available. In particular, medium and large scale models achieve remarkably higher results when textual information is integrated with visual inputs. While these findings highlight the strong contribution of textual cues, they also raise an important question: **are visual cues useful for answering the claims correctly or subtitles alone provide enough information?**

4.2 Ablations

Visual cues are indeed useful. To investigate the usefulness of visual information, we conduct an ablation study on Gemini 2.5 Pro and five mid and large scale open-weight models using a subset of single-scene claims with annotated timeframes specifying the relevant visual content (see Section 2.2). In Table 3, we report for this subset, model performance on subtitles-only, video-only, and video with subtitles settings, comparing results when frames are drawn from the entire movie versus only the targeted scenes, corresponding to the annotated timeframes. Across most models, performance improves when given the targeted visual content, both with and without subtitles, suggesting that subtitles alone, while important, are not sufficient and that visual cues are essential for correctly answering the constructed claims.

Table 4: Performance of Gemini 2.5 Pro across different input modalities. *Video* uses only the visual stream; *Subs* includes only subtitle information; *Synopsis* relies only on the synopsis of the movie obtained from Wikipedia; *Video w/ Subs* combines both visual and subtitles inputs; and *Movie Title* uses only the claim, along with the movie title and release year, without access to movie content.

	Input Modality				
Metric	**Video**	**Subs**	**Synopsis**	**Video w/ Subs**	**Movie Title**
Pairwise Acc.	37.2	56.7	25.5	60.6	43.7

We now turn to Gemini 2.5 Pro, the best-performing model, to examine its behavior across *input modalities*. Additional ablation studies are provided in Appendix C, where we analyze performance across *reasoning granularities* and *comprehension dimensions* for Gemini 2.5 Pro as well as for other large-scale open-weight models.

Beyond vision: the role of textual and world knowledge for Gemini 2.5. Table 4 presents an ablation study of Gemini 2.5 Pro, highlighting its strong reliance on subtitles and parametric (internal) knowledge. Notably, the model appears to underutilize the visual information: performance improves marginally when visuals are combined with subtitles compared to using subtitles alone, something that likely contributes to its gap relative to human performance. Examples of such cases are provided in Appendix F (Examples 5 and 6). This observation aligns with previous work showing that VLMs often fail to effectively leverage available visual information and instead overrely on language priors and textual input (Fu et al., 2025; Zheng et al., 2025). When provided solely with the movie title and release year, model's performance is above random, reflecting its reliance on prior knowledge likely acquired during pretraining. In contrast, performance declines when the model is given only the movie synopsis, indicating that not all forms of textual context are equally informative. Overall, these results suggest that (1) models leverage textual information from subtitles, without fully integrating successfully it with visual input (Fu et al., 2025; Zheng et al., 2025), while also (2) relying on prior knowledge obtained during pretraining. Crucially, performance remains substantially below human level, highlighting that simply encoding the full video content with standard methods (e.g., brute-force encoding) is insufficient. Future research should develop strategies enabling models to focus on relevant visual information while ignoring irrelevant details, an ability current VLMs lack.

5 Related Work

Vision and long-context LLMs. The field of VLMs has progressed rapidly, with recent models achieving strong video–language understanding (Deitke et al., 2024; Bai et al., 2025b; Zhu et al., 2025). Early approaches focused on short clips using spatio-temporal modules or temporal pooling (Zhang et al., 2023a; Li et al., 2023a; Maaz et al., 2024), while more recent systems rely on projection layers for efficient visual–language alignment (Li et al., 2023b; Liu et al., 2023; Li et al., 2023a; Liu et al., 2024b). For long-video understanding, methods focus on token compression, extended context windows, or memory consolidation (Li et al., 2023b; Liu et al., 2025; Balazevic et al., 2024; Song et al., 2024), while other pipelines convert video to captions before reasoning (Zhang et al., 2024; Wang et al., 2024b).

Long video understanding benchmarks. Existing benchmarks focus on short clips (Nagrani et al., 2025) or domain-specific videos (Mangalam et al., 2023; Qiu et al., 2024), with most under a few minutes or solvable using a few keyframes (Yu et al., 2019; Zhang et al., 2023b). Benchmarks for longer content (Parmar et al., 2024; Wu et al., 2024; Hu et al., 2025) remain limited in duration, scale, or annotation quality. Even extended video datasets (Chandrasegaran et al., 2024; Ataallah et al., 2024) often rely on synthetic questions or multiple-choice formats (Fu et al., 2024; Wang et al., 2024a), which introduce ordering biases and emphasize contrastive over isolated reasoning. Other benchmarks focus on retrieval (Zohar et al., 2025; Wang et al., 2025a; Zhao et al., 2025), while non-retrieval ones treat segments uniformly (Chandrasegaran et al., 2024; Tan et al., 2024), without explicitly testing memorization of critical content. Our benchmark addresses these gaps with long-form, manually annotated movies and evaluates claims independently using a pairwise protocol.

6 CONCLUSIONS

In this paper, we introduce MF^2, a comprehensive multimodal benchmark designed to evaluate VLMs on deep narrative understanding in the context of long movie comprehension. Our benchmark adopts a binary evaluation protocol and covers a diverse range of claim categories, including emotion understanding, temporal perception, causal reasoning, and event/entity understanding. These claims span varying levels of granularity—single-scene, multi-scene, and global— and focus on memorable core narrative elements. All examples are annotated by humans to ensure high-quality and reliable labels. Our extensive evaluation of both open-weight and closed state-of-the-art models reveals a significant performance gap between models and humans, underscoring the challenges and importance of our benchmark. Commercial models such as Gemini 2.5 Pro outperform others, including GPT-4o and other open-weight variants, yet still fall short of human-level performance. While transcripts provide a critical signal, brute-force encoding of visual content remains insufficient as models fail to identify, prioritize, and retain the most relevant information in extended video content, as humans naturally do. Addressing this limitation, we hope MF^2 will drive the development of methods that enable models to selectively extract, consolidate, and reason over relevant information while automatically discarding irrelevant details, a capability current VLMs still lack.

7 ETHICS STATEMENT

We adhered to established scientific and ethical standards in constructing and releasing MF^2. All source movies are released under the permissive Public Domain 1.0 license. Claims and annotations were created and validated exclusively by the authors; no external crowdworkers were employed. To encourage a plurality of perspectives in the annotation process, the annotation team consists of individuals from diverse demographic, institutional, and geographic backgrounds. Since MF^2 is derived entirely from fictional movies, it contains no personally identifiable information (PII) of real individuals. Nonetheless, some fictional content may reflect cultural stereotypes or outdated social norms. We caution researchers that models evaluated on MF^2 may inherit such biases, and we recommend appropriate safeguards when interpreting or deploying results. We advise users to employ MF^2 strictly within the scope of this work, namely as a benchmark for evaluating vision–language models on long movie understanding, and discourage its use beyond it.

8 REPRODUCIBILITY STATEMENT

We ensure reproducibility by releasing the full dataset and the codebase at https://huggingface.co/datasets/sardinelab/MF2 and https://github.com/deep-spin/MF2 respectively. The repository includes extended instructions to replicate all experimental settings. To facilitate long-term accessibility, and in accordance with the Public Domain 1.0 license, we additionally host a copy of the raw movie data. Detailed annotation protocols are provided in Appendix D, while Appendix E outlines additional experimental details. We encourage independent verification of our results and welcome contributions from the community to extend or stress-test MF^2 over time.

ACKNOWLEDGMENTS

We gratefully acknowledge Miguel Graça, Evan Paces-Wiles, Maya Nachesa, Daniil Larionov, Bryan Sukidi, and José Pombal for their participation in the human evaluation process. This work was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI), by EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), by the project DECOLLAGE (ERC-2022-CoG 101088763), by a research grant (VIL53122) from VILLUM FONDEN, by the Pioneer Center for AI DNRF grant number P1, by FCT/MECI through national funds and when applicable co-funded EU funds under UID/50008: Instituto de Telecomunicações, by DARPA ECOLE Program No. HR00112390060, NSF-AI Engage Institute DRL-2112635, ARO Award W911NF2110220, ONR Grant N00014-23-1-2356, and the Microsoft Accelerate Foundation Models Research (AFMR) grant program, by the project PID2023-151351NB-I00 funded by MCIN/ AEI /10.13039/501100011033 and by ERDF, UE. This collaboration resulted from an ELLIS workshop at MFO, the Oberwolfach Research Institute for Mathematics

in the German Black Forest. The event was funded by the state of Baden-Württemberg (Germany) and organised in collaboration with the ELLIS Institute Tübingen and the Max Planck Institute for Intelligent Systems.

REFERENCES

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. Llava-onevision-1.5: Fully open framework for democratized multimodal training. In *arXiv*, 2025.

Anthropic. The claude 3 model family: Opus, sonnet, haiku. 2024. URL https://api.semanticscholar.org/CorpusID:268232499.

Kirolos Ataallah, Chenhui Gou, Eslam Abdelrahman, Khushbu Pahwa, Jian Ding, and Mohamed Elhoseiny. Infinibench: A comprehensive benchmark for large multimodal models in very long video understanding, 2024. URL https://arxiv.org/abs/2406.19875.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

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

Ivana Balazevic, Yuge Shi, Pinelopi Papalampidi, Rahma Chaabouni, Skanda Koppula, and Olivier J Henaff. Memory consolidation enables long-context video understanding. In *Forty-first International Conference on Machine Learning*, 2024.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 238–255, Vienna, Austria, 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.acl-short.20/.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He, Cristobal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Fei-Fei Li. Hourvideo: 1-hour video-language understanding. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, Sifei Liu, Hongxu Yin, Yao Lu, and Song Han. Scaling rl to long videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 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*, pp. 24185–24198, 2024.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Winson Han, Ali Farhadi, and Ranjay Krishna. Molmo2: Open weights and data for vision-language models with video understanding and grounding, 2026. URL https://arxiv.org/abs/2601.10611.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Jen Dumas, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *arXiv preprint arXiv:2406.14515*, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2024.

Stephanie Fu, tyler bonnen, Devin Guillory, and Trevor Darrell. Hidden in plain sight: VLMs overlook their visual representations. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=qQb1JLrwol.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pp. 709–727. Springer, 2020.

Alon Jacovi, Avi Caciularu, Omer Goldman, and Yoav Goldberg. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5075–5084, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.308. URL https://aclanthology.org/2023.emnlp-main.308/.

Greg Kamradt. Needle in a haystack - pressure testing LLMs, 2024. URL https://github.com/gkamradt/LLMTest_NeedleInAHaystack.

Marzena Karpinska, Katherine Thai, Kyle Lo, Tanya Goyal, and Mohit Iyyer. One thousand and one pairs: A "novel" challenge for long-context language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17048–17085, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.948. URL https://aclanthology.org/2024.emnlp-main.948/.

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

Daniel T Levin and Daniel J Simons. Failure to detect changes to attended objects in motion pictures. *Psychonomic Bulletin & Review*, 4(4):501–506, 1997.

Amanpreet Li, Rowan Zellers, Youngjae Yu, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *CVPR*, 2022.

Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Fan Zhou, Chengen Huang, Yanpeng Li, Chongyan Zhu, Xiaoyi Ren, Chao Li, Yifan Ye, Peng Liu, Lihuan Zhang, Hanshu Yan, Guoyin Wang, Bei Chen, and Junnan Li. Aria: An open multimodal native mixture-of-experts model, 2025. URL https://arxiv.org/abs/2410.05993.

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

Ruizhe Li and Yanjun Gao. Anchored answers: Unravelling positional bias in gpt-2's multiple-choice questions. *arXiv preprint arXiv:2405.03205*, 2024.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models, 2023b.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=HN8V0flwJF.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

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

Ming Liu and Wensheng Zhang. Is your video language model a reliable judge? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=m8yby1JfbU.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Vishwesh Nath, Jinyi Hu, Sifei Liu, Ranjay Krishna, Daguang Xu, Xiaolong Wang, Pavlo Molchanov, Jan Kautz, Hongxu Yin, Song Han, and Yao Lu. Nvila: Efficient frontier visual language models, 2024b.

Olga Loginova, Oleksandr Bezrukov, Ravi Shekhar, and Alexey Kravets. Addressing blind guessing: Calibration of selection bias in multiple-choice question answering by video language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3216–3246, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.162. URL https://aclanthology.org/2025.acl-long.162/.

Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural Embedding Alignment for Multimodal Large Language Model. *arXiv e-prints*, art. arXiv:2405.20797, May 2024. doi: 10.48550/arXiv.2405.20797.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL https://openreview.net/forum?id=JVlWseddak.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li,

Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025. URL https://arxiv.org/abs/2503.01743.

Francesco Maria Molfese, Luca Moroni, Luca Gioffré, Alessandro Scirè, Simone Conia, and Roberto Navigli. Right answer, wrong score: Uncovering the inconsistencies of LLM evaluation in multiple-choice question answering. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18477–18494, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.950. URL https://aclanthology.org/2025.findings-acl.950/.

Arsha Nagrani, Mingda Zhang, Ramin Mehran, Rachel Hornung, Nitesh Bharadwaj Gundavarapu, Nilpa Jha, Austin Myers, Xingyi Zhou, Boqing Gong, Cordelia Schmid, Mikhail Sirotenko, Yukun Zhu, and Tobias Weyand. Neptune: The long orbit to benchmarking long video understanding, 2025. URL https://arxiv.org/abs/2412.09582.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng,

Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. Movie plot analysis via turning point identification. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1707–1717, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1180. URL https://aclanthology.org/D19-1180/.

Pinelopi Papalampidi, Frank Keller, Lea Frermann, and Mirella Lapata. Screenplay summarization using latent narrative structure. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1920–1933, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.174. URL https://aclanthology.org/2020.acl-main.174/.

Paritosh Parmar, Eric Peh, Ruirui Chen, Ting En Lam, Yuhan Chen, Elston Tan, and Basura Fernando. Causalchaos! dataset for comprehensive causal action question answering over longer causal chains grounded in dynamic visual scenes. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=gP4aAi7q8S.

Lu Qiu, Yi Chen, Yuying Ge, Yixiao Ge, Ying Shan, and Xihui Liu. Egoplan-bench2: A benchmark for multimodal large language model planning in real-world scenarios. *arXiv preprint arXiv:2412.04447*, 2024.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Ruchit Rawal, Khalid Saifullah, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark. *arXiv preprint arXiv:2405.08813*, 2024.

Bhuiyan Sanjid Shafique, Ashmal Vayani, Muhammad Maaz, Hanoona Abdul Rasheed, Dinura Dissanayake, Mohammed Irfan Kurpath, Yahya Hmaiti, Go Inoue, Jean Lahoud, Md. Safirur Rashid, Shadid Intisar Quasem, Maheen Fatima, Franco Vidal, Mykola Maslych, Ketan Pravin More, Sanoojan Baliah, Hasindri Watawana, Yuhao Li, Fabian Farestam, Leon Schaller, Roman Tymtsiv, Simon Weber, Hisham Cholakkal, Ivan Laptev, Shin'ichi Satoh, Michael Felsberg, Mubarak Shah, Salman Khan, and Fahad Shahbaz Khan. A culturally-diverse multilingual multimodal video benchmark & model, 2025. URL https://arxiv.org/abs/2506.07032.

Shrutika Singh, Anton Alyakin, Daniel Alexander Alber, Jaden Stryker, Ai Phuong S Tong, Karl Sangwon, Nicolas Goff, Mathew de la Paz, Miguel Hernandez-Rovira, Ki Yun Park, et al. It is too many options: Pitfalls of multiple-choice questions in generative ai and medical education. *arXiv preprint arXiv:2503.13508*, 2025.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. Moviechat: From dense token to sparse memory for long video understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18221–18232, 2024. doi: 10.1109/CVPR52733.2024.01725.

Chaolei Tan, Zihang Lin, Junfu Pu, Zhongang Qi, Wei-Yi Pei, Zhi Qu, Yexin Wang, Ying Shan, Wei-Shi Zheng, and Jian-Fang Hu. Synopground: A large-scale dataset for multi-paragraph video grounding from tv dramas and synopses, 2024. URL https://arxiv.org/abs/2408.01669.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia

Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Haochen Li, Jiale Zhu, Jiali Chen, Jiaxing Xu, Jiazheng Xu, Jing Chen, Jinghao Lin, Jinhao Chen, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Ruiliang Lyu, Shangqin Tu, Sheng Yang, Shengbiao Meng, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wei Jia, Wenkai Li, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyu Zhang, Xinyue Fan, Xuancheng Huang, Yadong Xue, Yanfeng Wang, Yanling Wang, Yanzi Wang, Yifan An, Yifan Du, Yiheng Huang, Yilin Niu, Yiming Shi, Yu Wang, Yuan Wang, Yuanchang Yue, Yuchen Li, Yusen Liu, Yutao Zhang, Yuting Wang, Yuxuan Zhang, Zhao Xue, Zhengxiao Du, Zhenyu Hou, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2026. URL https://arxiv.org/abs/2507.01006.

Hengyi Wang, Haizhou Shi, Shiwei Tan, Weiyi Qin, Wenyuan Wang, Tunyu Zhang, Akshay Nambi, Tanuja Ganu, and Hao Wang. Multimodal needle in a haystack: Benchmarking long-context capability of multimodal large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3221–3241, Albuquerque, New Mexico, April 2025a. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.166. URL https://aclanthology.org/2025.naacl-long.166/.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. Lvbench: An extreme long video understanding benchmark, 2024a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025b. URL https://arxiv.org/abs/2508.18265.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pp. 58–76. Springer, 2024b.

Yuxuan Wang, Cihang Xie, Yang Liu, and Zilong Zheng. Videollamb: Long-context video understanding with recurrent memory bridges, 2024c. URL https://arxiv.org/abs/2409.01071.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B. Tenenbaum, and Chuang Gan. STAR: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=EfgNF5-ZAjM.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. In *The Thirty-eight Conference on Neural Information*

Processing Systems Datasets and Benchmarks Track, 2024. URL https://openreview.net/forum?id=3G1ZDXOI4f.

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

LLM-Core-Team Xiaomi. Mimo-vl technical report, 2025. URL https://arxiv.org/abs/2506.03569.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mPLUG-owl3: Towards long image-sequence understanding in multi-modal large language models. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=pr37sbuhVa.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in LLM-as-a-judge. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=3GTtZFiajM.

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

Jeffrey M Zacks, Nicole K Speer, and Jeremy R Reynolds. Segmentation in reading and film comprehension. *Journal of Experimental Psychology: General*, 138(2):307, 2009.

Rowan Zellers, Ximing Lu, Youngjae Yu, Jae Sung Park, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. In *NeurIPS*, 2021.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. URL https://arxiv.org/abs/2501.13106.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21715–21737, 2024.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023a. URL https://arxiv.org/abs/2306.02858.

Hongjie Zhang, Yi Liu, Lu Dong, Yifei Huang, Zhen-Hua Ling, Yali Wang, Limin Wang, and Yu Qiao. Movqa: A benchmark of versatile question-answering for long-form movie understanding, 2023b. URL https://arxiv.org/abs/2312.04817.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video Instruction Tuning With Synthetic Data. *arXiv e-prints*, art. arXiv:2410.02713, October 2024. doi: 10.48550/arXiv.2410.02713.

Zijia Zhao, Haoyu Lu, Yuqi Huo, Yifan Du, Tongtian Yue, Longteng Guo, Bingning Wang, weipeng chen, and Jing Liu. Needle in a video haystack: A scalable synthetic evaluator for video MLLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ZJo6Radbqq.

Xu Zheng, Chenfei Liao, Yuqian Fu, Kaiyu Lei, Yuanhuiyi Lyu, Lutao Jiang, Bin Ren, Jialei Chen, Jiawen Wang, Chengxin Li, Linfeng Zhang, Danda Pani Paudel, Xuanjing Huang, Yu-Gang Jiang, Nicu Sebe, Dacheng Tao, Luc Van Gool, and Xuming Hu. Mllms are deeply affected by modality bias, 2025. URL https://arxiv.org/abs/2505.18657.

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

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv e-prints*, art. arXiv:2504.10479, April 2025. doi: 10.48550/arXiv.2504.10479.

Orr Zohar, Rui Li, Andres Marafioti, Xiaohan Wang, Stanford AI Team, and Hugging Face. Timescope: How long can your video large multimodal model go? https://huggingface.co/blog/timescope-video-lmm-benchmark, July 2025. Accessed: YYYY-MM-DD.

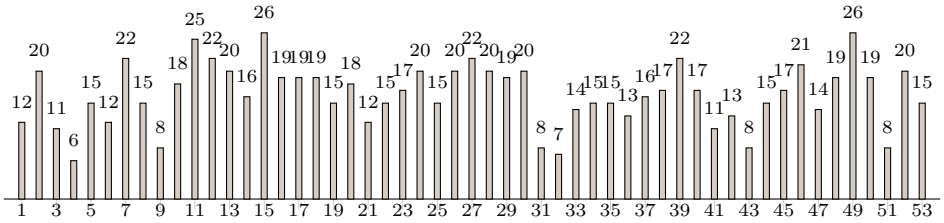


Figure 4: Number of claim pairs per movie.

A Data Annotation Details

Movie Collection Metadata We collected 53 full-length movies from the Internet Archive[6], focusing on titles released under the Public Domain 1.0 license to ensure legal reusability and support open-access research. Original-language subtitles, mostly in English, were obtained from OpenSubtitles.org[7]. For one movie without available subtitles, we used whisper-1 (Radford et al., 2023)[8] to generate a transcript and manually post-edited to ensure high quality. Each movie includes audio and aligned subtitles, with an average duration of 88 minutes. In Table 7, we provide detailed information on the 53 released movies, including their genre, original language, and duration. Collectively, the movies span a wide temporal range of approximately five decades (1920–1970) and cover a diverse set of genres.

Annotator Demographics All annotations were performed manually by 26 co-authors from 12 institutions across 7 countries, with a roughly balanced distribution of career stages (PhD students, postdocs, and faculty members) and gender. We did not rely on crowdsourcing or semi-automatic approaches, as constructing claims about central narrative events—often requiring reasoning across multiple scenes—demands careful human judgment. Having co-authors perform the annotations also provided strong intrinsic motivation and accountability, ensuring high-quality, consistent work that might not be achievable with anonymous or paid crowdworkers. Detailed guidelines, multiple rounds of discussion, and collective revisions were used to ensure consistency and high quality, while annotators were not informed of any specific hypotheses or expected outcomes, reducing the risk of bias.

Annotated Claim Pairs Figure 4 shows the number of annotated claim pairs per movie.

Comprehension Dimensions Each claim pair was labeled according to one or more comprehension dimensions, which capture specific aspects of narrative understanding. These dimensions are informed by prior work (Xiao et al., 2021; Zhang et al., 2023b; Wang et al., 2024a) and are defined in Table 6.

B Pairwise Accuracy vs Simple Accuracy

Our evaluation protocol is motivated as follows: multiple-choice settings introduce distractor-driven biases, where models exploit superficial cues or distractor order (Molfese et al., 2025). Following prior work on narrative understanding (Karpinska et al., 2024), we evaluate each minimally differing claim independently, removing distractors and reducing such shortcut behaviors. Under this setup, pairwise accuracy is stricter and more informative; a model is counted as correct only if it predicts both labels accurately, minimizing the chance that it appears "correct for the wrong reason" (Karpinska et al., 2024). In contrast, in simple accuracy each prediction contributes independently to the score making it less suitable for our case and more suitable for multiple-choice formats, which we intentionally avoid for the aforementioned reasons.

The task is equivalent to two independent binary classification tasks only at inference time. Although we obtain a true/false prediction for each claim independently, the evaluation protocol differs: it is

[6] https://archive.org
[7] https://www.opensubtitles.org
[8] https://platform.openai.com/docs/models/whisper-1

Table 5: Reporting pairwise accuracy, average accuracy, accuracy on facts and accuracy on fibs for some open-weight models and Gemini2.5. Results highlight that simple accuracy (by it's own) can be misleading.

Model	Pairwise Acc	Average Acc	Acc on Facts	Acc on Fibs
Video-only				
Gemini 2.5 Pro	37.2	64.2	47.6	80.9
Gemma3-27B	31.5	61.2	57.4	65.1
Qwen3VL-30B	19.2	56.9	24.3	89.5
InternVL35-38B	26.6	60.0	38.8	81.1
Qwen2.5VL-72B	29.7	58.8	45.4	72.1
InternVL3-78B	22.1	58.0	37.7	78.3
Video w/ Subtitles				
Gemini 2.5 Pro	60.6	77.6	69.1	86.2
Gemma3-27B	42.9	68.1	60.4	75.8
Qwen3VL-30B	42.3	68.5	51.5	85.6
InternVL35-38B	46.2	70.3	61.6	79.0
Qwen2.5VL-72B	45.9	70.4	55.6	85.1
InternVL3-78B	51.3	72.7	67.5	77.9

pairwise, not item-wise. This makes the protocol different from the standard binary classification evaluation accuracy in which each prediction contributes independently to the score. In our evaluation, the relationship between the two predictions is essential. **The model must implicitly distinguish what truly happened from a minimally edited alternative that is also plausible, even though it never sees the two statements jointly during inference.** That said, the evaluation depends on the relationship between the two predictions, not the predictions alone.

As a further illustration of why interpreting results based only on simple accuracy can be misleading, we report in Table 5 several metrics (pairwise accuracy, simple accuracy, accuracy on facts and accuracy on fibs) for some of the open-weight models and for Gemini 2.5 Pro, across both the video-only and video w/ subtitles settings. *Notably, simple accuracy can appear high simply because the model performs well on the fibs, while failing on the facts*, highlighting a bias towards predicting False. Taken together **these considerations motivate our choice of pairwise accuracy as the primary evaluation metric**.

C Extended Ablations on Reasoning Granularity and Comprehension Dimensions

Input modality contributions across comprehension dimensions and reasoning granularities for Gemini 2.5 Pro. In Fig. 5, we present ablation studies for Gemini 2.5 Pro, examining how different input modalities contribute to performance across comprehension dimensions and reasoning granularities. We observe that the model handles temporal perception more effectively than other comprehension aspects across all modalities, a trend that also holds for large-scale open-weight models as well (see Appendix C). This is likely because time-related information is often directly observable in visual and textual inputs, making it easier to track and interpret (Zellers et al., 2021; Li et al., 2022). Event and entity understanding is notably weaker under visual-only conditions, likely due to the need for linguistic disambiguation. This limitation becomes evident when subtitles are introduced: the most significant gain is observed in the aforementioned category, highlighting the complementary role of textual context. In contrast, emotional understanding benefits the least from subtitles, indicating challenges in affective comprehension. Beyond comprehension dimensions, reasoning performance under visual-only inputs remains relatively consistent across reasoning types. However, under the presence of textual cues, global reasoning becomes more challenging than single- and multi-scene reasoning.

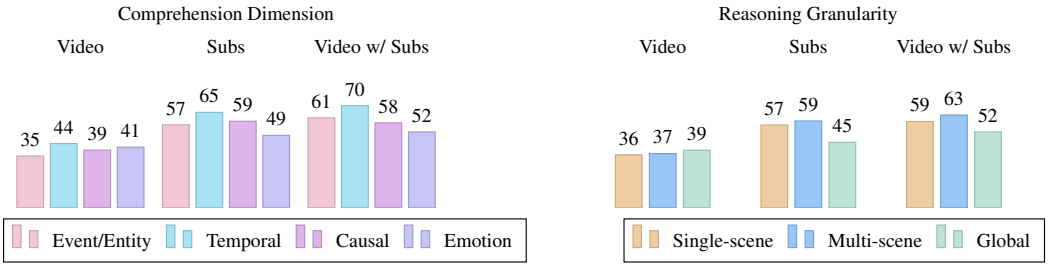


Figure 5: Pairwise accuracy for Gemini 2.5 Pro per comprehension dimension and reasoning granularity when varying the input modalities.

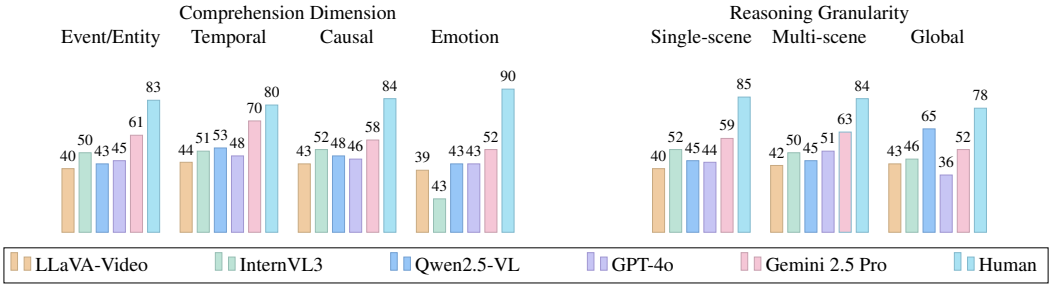


Figure 6: Pairwise accuracy for large-scale models with video and subtitles, and human baseline per comprehension dimension and reasoning granularity.

A fine-grained view of large-scale model performance across comprehension dimensions and reasoning granularities. Fig. 6 shows that, among the large-scale models, Gemini 2.5 Pro still demonstrates inferior performance, ranking second to humans in various categories. Other models like LLaVA-Video and InternVL3 generally show lower scores, suggesting areas for improvement. The results also highlight varying degrees of difficulty across the tasks, with emotion comprehension appearing to be a strong point for humans, while temporal perception is a strong point for models. Interestingly, the analysis on reasoning granularity reveals an interesting pattern between humans and models: as reasoning shifts from single-scene to multi-scene and eventually to global, model performance tends to oscillate across models, while human performance declines. Notably, Qwen2.5-VL shows improved accuracy on claims requiring global reasoning compared to the other granularities. This may suggest that global narrative information is more frequently represented in pretraining corpora (e.g., Wikipedia summaries of movies), whereas single-scene questions demand localized details that are less likely to be encountered in such sources. In contrast, humans may face increased cognitive load or memory limitations when reasoning across multiple scenes, which could explain the drop in performance in some cases.

D Detailed Guidelines for Data Annotation and Human-Eval

D.1 Data Annotation Guidelines

In Figs. 7 and 8, we present the detailed guidelines provided to annotators during the data annotation process. These include instructions for constructing contrastive claim pairs, and labeling each pair with the appropriate reasoning granularity and comprehension dimensions. Furthermore, in Figs. 9 and 10, we include a subset of illustrative examples shown to annotators to guide their annotations of reasoning granularity and comprehension dimensions, respectively.

We note that among the comprehension dimensions annotators could assign to each claim pair, an "Other" category was included to account for cases that did not clearly align with any of the predefined dimensions. As this label was selected rarely (0.49% of the data), it is excluded from the figures presented in the main text.

D.2 Human Evaluation Guidelines

In Figs. 11 and 12, we provide the full set of guidelines shared to participants during the human evaluation process, which consists of two stages: an initial stage in which evaluators respond without revisiting the movie, and an optional second stage that allows revisiting. While we only analyze the results from Stage 1—as our goal is to assess movie understanding based on memorable events without allowing participants to rewatch parts of the film—we include the complete instructions for both stages to offer full context. Additionally, we provide an illustration of the evaluation interface to clarify the evaluation setup.

E Details on Experimental Setup

E.1 Prompt Templates

In Figs. 13 and 14 we present the direct and explanation prompt templates used for open-weight and closed models, respectively. The former requests only a True/False response, while the latter additionally asks for a brief justification before the final answer. We found that the direct prompt yielded better performance for open-weight models, while the explanation prompt proved more effective for closed models. When experimenting with different input modalities—such as adding the synopsis, subtitles, or movie title—we adapt the prompts accordingly by incorporating this information into the instruction prompt. The number of frames is selected such that the model's context length is not exceeded

E.2 Resources

Our infrastructure consists of a single machine equipped with 2 NVIDIA H200 GPUs (140GB each) and 12 Intel Xeon Gold 6348 CPUs (2.60GHz, 1TB RAM). Most experiments were conducted on a single GPU, except for evaluations involving mid and large scale open-weight models, where all 2 GPUs were used to accelerate inference.

F Illustrative Examples

In this part, we include a small set of examples (Tables 8, 9, 10, 11, 12, 13, 14, 15, 16, 17) showing when visual or text cues are needed to resolve the claims and how Gemini 2.5 Pro (best performing model) succeeds or fails on them, along with its corresponding explanations.

AI Assistance

We would like to note that large language models (ChatGPT) were used to assist in drafting and polishing the writing of this work.

Table 7: Metadata of collected movies.

Movie (Year)	**Genre (IMDB)**	**Language**	**Duration (mins)**
The Last Chance (1945)	Drama, War	en, it	93.84
They Made Me a Criminal (1939)	Boxing, Film Noir, Crime, Drama, Sport	en	91.21
Tokyo After Dark (1959)	Drama	en	81.23
The Sadist (1963)	Horror, Thriller	en	91.63
Suddenly (1954)	Film Noir, Psychological Thriller, Crime, Drama, Thriller	en	76.71
Sabotage (Hitchcock) (1936)	Psychological Thriller, Spy, Crime, Thriller	en	75.92

Murder By Contract (1958)	Film Noir, Crime, Drama, Thriller	en	80.45
Pushover (1954)	Film Noir, Crime, Drama, Thriller	en	87.77
Go for Broke (1951)	Drama, History, War	en	90.85
Meet John Doe (1941)	Political Drama, Satire, Comedy, Drama, Romance	en	122.87
Scarlet Street (1945)	Film Noir, Tragedy, Crime, Drama, Thriller	en	102.39
Little Lord Fauntleroy (1936)	Period Drama, Drama, Family	en	100.72
Deadline - U.S.A. (1952)	Film Noir, Crime, Drama	en	87.06
My Favorite Brunette (1947)	Hard-boiled Detective, Comedy, Crime, Mystery, Romance, Thriller	en	87.34
Woman in the Moon (1929)	Adventure, Comedy, Drama, Romance, Sci-Fi	de	168.73
Lonely Wives (1931)	Comedy, Romance	en	85.35
Nothing Sacred (1937)	Satire, Screwball Comedy, Comedy, Drama, Fantasy, Romance	en	73.57
Fingerman (1955)	Film Noir, Crime, Drama, Thriller	en	82.06
Borderline (1950)	Film Noir, Crime, Drama, Thriller	en	88.16
Babes in Toyland (1934)	Screwball Comedy, Slapstick, Comedy, Family, Fantasy, Musical	en	77.26
The Man From Utah (1934)	Drama, Western	en	51.49
The Man With The Golden Arm (1955)	Drug Crime, Psychological Drama, Crime, Drama, Romance	en	119.07
A Star Is Born (1937)	Tragic Romance, Drama, Romance	en	110.98
Africa Screams (1949)	Farce, Action, Adventure, Comedy	en	79.13
Dementia 13 (1963)	Slasher Horror, Horror, Thriller	en	74.94
Fear and Desire (1952)	Drama, Thriller, War	en	70.19
The Little Princess (1939)	Costume Drama, Comedy, Drama, Family, Musical	en	92.77
Father's Little Dividend (1951)	Comedy, Drama, Romance	en	81.74
Kansas City Confidential (1952)	Conspiracy Thriller, Film Noir, Heist, Crime, Drama, Thriller	en	99.27

Of Human Bondage (1934)	Dark Romance, Film Noir, Medical Drama, Tragedy, Tragic Romance, Drama, Romance	en	82.77
Half Shot at Sunrise (1930)	Comedy, Musical	en, fr	78.04
Bowery at Midnight (1942)	B-Horror, Crime, Horror, Thriller	en	62.05
The Emperor Jones (1933)	Drama, Music	en	76.29
The Deadly Companions (1961)	Adventure, Drama, Western	en	93.62
The Red House (1947)	Film Noir, Drama, Mystery, Thriller	en	100.39
Trapped (1949)	Film Noir, Crime, Drama, Thriller	en	79.4
City of Fear (1959)	Crime, Drama, Thriller	en	75.18
Kid Monk Baroni (1952)	Action, Drama, Sport	en	79.56
Tight Spot (1955)	Film Noir, Crime, Drama, Thriller	en	95.99
Captain Kidd (1945)	Costume Drama, Swashbuckler, Adventure, Biography, Drama, History	en	87.53
The Front Page (1931)	Dark Comedy, Satire, Screwball Comedy, Comedy, Crime, Drama, Mystery, Romance	en	101.14
The Hitch-Hiker (1953)	Film Noir, Crime, Drama, Thriller	en	70.8
Obsession (1949)	Film Noir, Psychological Thriller, Crime, Thriller	en	92.39
Thunderbolt (1929)	Film Noir, Crime, Drama, Music, Romance	en	91.27
Cyrano de Bergerac (1950)	Swashbuckler, Adventure, Drama, Romance	en	112.87
Scandal Sheet (1952)	Film Noir, Crime, Drama, Romance, Thriller	en	81.75
Ladies in Retirement (1941)	Film Noir, Crime, Drama	en	92.31
Detour (1945)	Film Noir, Crime, Drama	en	69.09
The Crooked Way (1949)	Film Noir, Crime, Drama, Thriller	en	85.95
A Bucket of Blood (1959)	Comedy, Crime, Horror	en	65.84
Love Affair (1939)	Holiday Romance, Comedy, Drama, Romance	en	89.62
The Jackie Robinson Story (1950)	Biography, Drama, Sport	en	76.82

The Last Time I Saw Paris (1954)	Tragedy, Tragic Romance, Drama, Romance	en	116.02

Guidelines for Data Annotation (Part 1)

We are conducting a research study on long movie understanding as part of a broader effort to explore how well viewers comprehend and recall complex narratives. Your task is to create claims that test a viewer's comprehension of a movie after watching it. These claims will be used in a human evaluation study to assess how well participants understand and recall key events from the movie. We appreciate your participation in this data collection process.

General Task Instructions Select a movie from the current "Pool" of movies (the "Pool" can be found in <LINK>). Make sure this movie is not selected by another annotator.

- Watch the entire movie carefully.
- We highly recommend reading the example claims provided to gain a better understanding of the task you need to fulfil.
- Start writing down your claims following the template available in <LINK> (you will find two tabs available: the "Examples" tab contains claim examples, and the "Annotations Template" tab is the template you should follow). Please create another sheet with your claims–do not directly use the current template–and send it to us once it is completed.

Annotation Process

1. Writing Claims You are asked to create pairs of contrastive claims, where one claim is true (fact) and the counterfactual version is false (fib). The two claims should differ by minimal edits, meaning they should be as similar as possible while maintaining contrast. Each claim should differ in a subtle but meaningful way, challenging comprehension without being overly obvious.
Example:
Fact: The first bomb exploded in the bus.
Fib: The first bomb exploded in the aquarium.
Why this works: The counterfactual claim is created with minimal edits, maintaining contrast while testing the understanding of a key event.

2. Select Claim Granularity For each pair of claims you constructed, indicate whether answering them correctly requires reasoning based on a single scene, multiple scenes, or globally within the movie.
Definition of scene:
A scene in film refers to a complete unit of storytelling, usually consisting of a sequence of events and dialogue taking place in a specific location and time. It often involves one or more characters and is usually shot in one continuous take or consisting of a sequence of shots.
Reasoning Granularity Labels:

- **Single-scene:** Claims that are answerable using information from a single scene.
- **Multi-scene:** Claims falling into this granularity require information/evidence from multiple distinct scenes, but not from the whole film. In this case, details are usually spread out between the multiple scenes. The supporting information/evidence is distributed, but explicit and locatable (timestamps/scenes can be clearly identified and referenced)
- **Global:** Claims falling into this granularity require a holistic understanding of the movie narrative. They cannot be easily tied to specific scenes or timestamps, and need to infer or accumulate information/evidence that emerges across the entire narrative (timestamps/scenes can not be clearly identified and referenced).

Note: Reasoning granularity labels should be selected based on the fact (true claim). Check the examples provided in the "Examples for Reasoning Granularity" part.

Figure 7: Guidelines provided for the data annotation procedure (Part 1).

Guidelines for Data Annotation (Part 2)

3. Claim Categorization Identify the comprehension dimensions the constructed pair of claims examines. Sometimes more than one dimension is examined, so we allow for multiple labels.

Comprehension Dimension Labels:

- **Event/Entity Understanding:** it refers to claims that require the identification of key entities (such as people, places, or objects) and understanding of actions or events involving those entities throughout the narrative. Understanding these claims involves tracking the presence and role of entities across scenes, extracting relationships among them, observing and interpreting their actions, and linking them to relevant events in the narrative.
- **Temporal Perception:** temporal perception refers to claims that require understanding of the timeline of events. It involves reasoning about the order in which events or actions occur—e.g., determining whether an event/action takes place before, after or at the same time as another—and may also require counting the number of specific actions or events. Unlike tasks focused on localizing a specific action in time, temporal perception emphasizes comprehension of broader temporal relationships within the evolving storyline.
- **Emotion Understanding:** emotional understanding refers to claims that involve recognizing and interpreting the emotional development of characters throughout the narrative.
- **Causal Reasoning:** causal reasoning refers to claims that require identifying cause-and-effect relationships between events or actions, where the relationship may be either direct or implicit.
- **Other:** If none of the above fit, select "Other" and suggest a new category.

Note: The categorization is based on both claims (fact and fib). Check the examples provided in the "Examples for Comprehension Dimensions" part.

Important Points To Consider

- **Ensure claims assess the viewer's understanding of the movie**. To put it simply, claims should refer to **significant moments** in the movie, **avoiding trivial details or Needle in a Haystack (NIAH)-style claims**, such as: "The detective wears a red T-shirt" (if this detail is not important in the movie).
- **Claims must be clear and unambiguous in isolation**, meaning they should be understandable without requiring additional context but should still require reasoning based on the movie. **Each claim should be self-contained and make sense independently**, without referencing its counterfactual version. Also, **avoid highly subjective or interpretive claims**. Each claim should still have a definitive answer based on the movie's content.
- **Avoid providing unnecessary contextual details**. For example, do not use phrases like "in the beginning of the movie, . . . ", "in the final scene, . . . " unless such information is essential to understanding the claim.
- Ensure that claims focus on memorable and salient narrative content that can be recalled without rewatching the movie.
- Once you finish the annotation process, please **go through your claims and confirm that they are in line with the points raised above** (these points are important to be covered to ensure good quality of annotations).

Figure 8: Guidelines provided for the data annotation procedure (Part 2).

Guidelines for Data Annotation (Part 3)

Examples for Reasoning Granularity In this part, we provide examples to illustrate how to assign reasoning granularity labels.

Example 1:
Fact: According to the Hattley, the individual shown in the photograph (Marakelli) worked with Constain.
Fib: According to Hattley, the individual shown in the photograph (Marakelli) had no connection or working relationship with Constain.
Reasoning Granularity: Single-scene.
Justification: This event is categorized as single-scene because it takes place within one specific scene: Hattley shows the photograph to Conley, they are having a discussion and it is implied that Marakelli worked with Constain in the mafia.

Example 2:
Fact: Hattley appeared visibly bothered with the discussion he had in his office with Constain's attorney.
Fib: Hattley appeared pleased with the discussion he had in his office with Constain's attorney.
Reasoning Granularity: Single-scene.
Justification: That is again a single scene event. Constain's attorney enters the office and they are having a discussion. After a while, Hattley kicks him out.

Example 3:
Fact: Miss Conley received a dress as a personal gift from the policeman.
Fib: Miss Conley received a dress as a gift from the government, delivered by the policeman.
Reasoning Granularity: Multi-scene.
Justification: That is a multi-scene event, that we need to ground on 2 independent scenes to answer the question correctly. In the first scene Miss Conley receives a gift from the policeman, who says that the gift is from the government. After a while (some scenes are interleaved), she understands that the policeman bought the gift for her and not the government. So to answer correctly, we need to ground on these 2 specific scenes.

Example 4:
Fact: Conley's statement about her occupation, describing herself as a "gang buster," implicitly refers to Constain.
Fib: Conley's statement about her occupation, describing herself as a "gang buster," implicitly refers to Pete Tinelli.
Reasoning Granularity: Global
Justification: There is a single scene in the end of a movie during which Conley characterises herself as a "gang buster". Although it is a single scene, it is impossible to understand solely by this scene why she said it and to whom she is referring to. We need to watch a big part of the movie (if not all of it) to understand that refers to Constain.

Figure 9: Guidelines provided for the data annotation procedure (Part 3). This part of the guidelines provides examples given to annotators to illustrate how to assign reasoning granularity labels. While more examples were shared during the annotation process, we include a selection here for illustrative purposes.

Guidelines for Data Annotation (Part 4)

Examples for Comprehension Dimensions In this part we provide examples to illustrate how to assign comprehension dimension labels.

Example 1:
Fact: At Jim's bar, the Connel keeps drinking as he talks to the fake John Doe, expressing his frustration and concern.
Fib: At Jim's bar, the Connel keeps drinking as he talks to the fake John Doe, expressing hope and happiness.
Comprehension Dimension: emotion understanding
Justification: We need to understand what emotion Connel expressed, to answer the pair of claims correctly.

Example 2:
Fact: Conley's statement about her occupation describing herself as a "gang buster", implicitly refers to Constain.
Fib: Conley's statement about her occupation describing herself as a "gang buster", implicitly refers to Pete Tinelli.
Comprehension Dimension: entity/event understanding
Justification: We need to understand to whom the expression "gang buster" refers to. So, the comprehension dimension is entity understanding.

Example 3:
Fact: Hallet brought Conley's sister to the hotel with the intent to make Conley testify in the trial.
Fib: Hallet brought Conley's sister to the hotel with the intent to make her feel safe.
Comprehension Dimension: causal reasoning
Justification: Here we need to understand why Hallet brought Conley's sister to the hotel. So it examines a causal-and-effect relationship.

Example 4:
Fact: Conley decided to testify only after Wiloughby's death.
Fib: Conley had already decided to testify before Wiloughby's death.
Comprehension Dimension: temporal perception
Justification: that pair examines the temporal dimension (if the decision was taken before or after Wiloughby's death).

Figure 10: Guidelines provided for the data annotation procedure (Part 4). This part of the guidelines provides examples given to annotators to illustrate how to assign comprehension dimension labels. While more examples were shared during the annotation process, we include a selection here for illustrative purposes.

Guidelines for Human Evaluation (Part 1)

This evaluation study aims to assess how well people comprehend and recall key events from a movie. You will watch a movie and then evaluate a series of claims about its content. Your goal is to determine whether each claim is True or False, based solely on what was shown in the movie. We appreciate your participation in this study.

Task Instructions

- Assign to yourself the movies you want to watch and do the test (we expect 2 movies per person). Please add your name to the Human-Eval column, on this LINK.
- Visit the platform for evaluation LINK.
- Provide your email to receive access to the movie (it will be used as your unique identifier).
- Once you submit your email, you should carefully select from the drop-down list the corresponding movie you assigned yourself and proceed with the evaluation. You will be shown with the movie link. Please open it in a new tab.

The test is divided in **2 stages**: The **first stage** is **mandatory** and should be completed by everyone (*during this stage you are not allowed to go back to the movie while answering the questions*). The **second stage** is **optional** (*during this stage you are allowed to go back to the movie while answering the questions*).

Stage 1:

1. **Watch the entire movie carefully before proceeding to the evaluation**. Pay attention to details and context in the movie, as some claims may be subtle or require careful reasoning.
2. After watching, it's time to proceed to Stage 1. **Please do not go back to the movie until Stage 1 of the test is completed.** Press the "Start Classifying Claims" button, and you will be shown with **one claim at a time**. For each claim shown, you need to do the following:
    - **Classify the claim as True/False** (you should always answer truthfully, without aiming to maximise you score).
    - Mark your **confidence** about your answer. This is helpful for stage 2, where you will have the opportunity to revise your claims (by looking back at the movie).
    - Leave a comment if any of the following applies: If a claim is **ambiguous, unclear, open to interpretation, has a bad phrasing or typos, you may leave an optional comment explaining your concerns**. You can also comment on the claim in case it is **needle-in-a-haystack style** and you think it is too detailed and doesn't test the understanding of the movie.
    - Once you answered, click "Save" to submit your response and move on to the next claim.

Important details: Once you submit an answer, you cannot go back and change it. At this stage, you are **strictly prohibited from searching back in the movie, rewinding, or rewatching scenes while answering the claims**. Your responses **should be based on your memory** and **understanding**. You must **NOT use any AI tools or external sources to verify or generate answers**. The goal of this study is to assess human understanding of long movies, not automated retrieval or AI-assisted responses. Also you are not allowed to take any paper notes, while watching the movie.

Figure 11: Guidelines provided for human evaluation (Part 1).

Guidelines for Human Evaluation (Part 2)

Stage 2:
Once you complete Stage 1, you will see a message asking you if you want to proceed to Stage 2 (Stage 2 is optional).

During Stage 2, you will be shown again with the choices you selected during Stage 1, but now **you can revise your answers by looking back to the movie** (you can reuse the movie link we provided you). You will be shown for each claim with the choices you did in Stage 1. You are free to change them and proceed to the next claims. Don't worry your answers will not be overwritten. Once you finish with Stage 2, you will be shown with a confirmation message.

If you have any questions or encounter any technical issues, please report them to our team! Thank you for your time and effort!

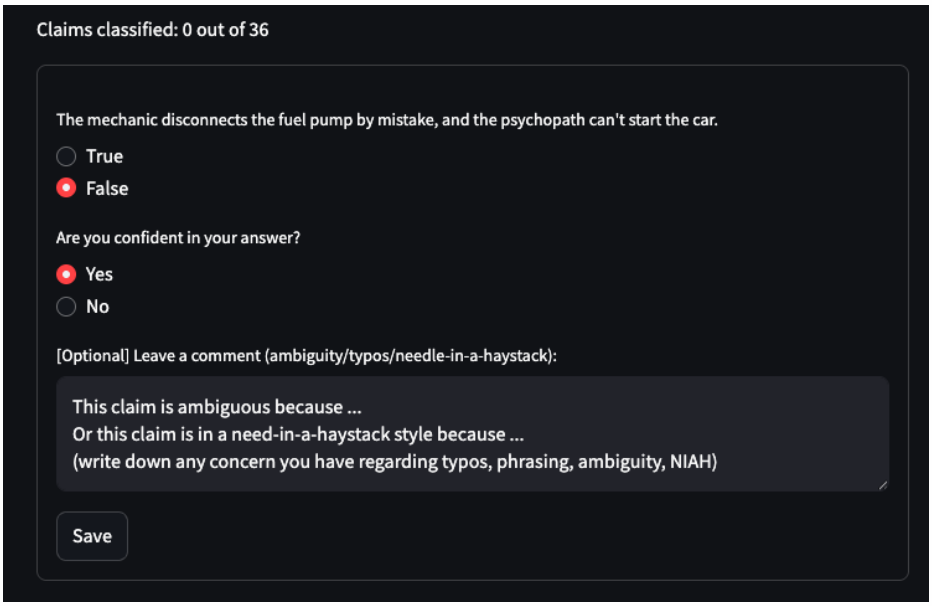


Illustration of the human evaluation interface.

Figure 12: Guidelines provided for human evaluation (Part 2).

Table 6: Definitions of comprehension dimensions.

Comprehension Dimension	Definition
Event/Entity Understanding	Involves identifying key entities (e.g., people, places, or objects) and understanding the events they participate in. This includes tracking entities across scenes, interpreting their roles, and recognizing their interactions and relationships throughout the narrative.
Temporal Perception	Requires reasoning about the timeline of events—determining whether actions occur before, after, or simultaneously—and may also include counting or sequencing events. The focus is on broader temporal relationships within the narrative.
Emotion Understanding	Involves recognizing the emotional states of characters and interpreting how these emotions evolve throughout the story.
Causal Reasoning	Focuses on identifying cause-and-effect relationships between events or actions, including both explicit and implicit dependencies that may span multiple scenes.

Direct Prompt Template

System: You are a helpful AI assistant. Your task is to carefully analyze the provided content and determine whether statements made about it are true or false based on the available information.

User: You are provided with a movie and a statement. Your task is to carefully watch the movie and then determine whether the statement is true or false.
Answer TRUE if the statement is true in its entirety based on the movie.
Answer FALSE if any part of the statement is false based on the movie.

Statement: {claim}
Based on the movie, is the above statement TRUE or FALSE?
Provide only your final answer.

Figure 13: Direct prompt template used for **open-weight** models.

Explanation Prompt Template

System: You are a helpful AI assistant. Your task is to carefully analyze the provided content and determine whether statements made about it are true or false based on the available information.

User: You are provided with a movie and a statement. Your task is to carefully watch the movie and then determine whether the statement is true or false.
Answer TRUE if the statement is true in its entirety based on the movie.
Answer FALSE if any part of the statement is false based on the movie.

Statement: {claim}
Based on the movie, is the above statement TRUE or FALSE?
First provide an explanation of your decision-making process in at most one paragraph, and then provide your final answer.

Figure 14: Explanation prompt template used for **closed** models.

Table 8: **Illustrative Example 1 from the movie "Suddenly"**. Gemini fails to predict both claims correctly in every modality setting.

Example 1 (Part 1) - Movie: ***"Suddenly"***


Pidge brings the water


Setting up the circuit


Pop spills the water


The electrocution

Fact (✓): The grandfather, Peter Benson, intentionally spilled the water on the floor.
Fib (✗): The grandfather, Peter Benson, accidentally spilled the water on the floor.

Granularity: Multi-scene

Context (Spoiler alert !): Early in the film, the TV breaks. Later, when the criminals have already broken into the house, pop (the grandfather) orchestrates the setup to save everyone (without explicitly stating it). He feigns a heart attack and asks for water, Pidge brings it. The TV technician pretends to repair the broken TV as part of a sabotage plan, while the circuit. When the moment comes, Pop *intentionally* spills the water on the floor (placing it on the edge of the table). Shortly after, the electrocution occurs when one of the criminals steps on the spilled water. *Both visual and textual information are needed to understand that Pop indeed spilled the water and what was his intention. Gemini fails to predict both claims correctly in every modality setting.*

Gemini's Predictions:
Subs only setting

Prediction on Fact: The statement is "The grandfather, Peter Benson, intentionally spilled the water on the floor." To determine if this is true, we need explicit evidence in the transcript: 1. That Peter Benson spilled water; 2. That he did so intentionally. Transcript evidence: Peter Benson (Pop) feigns a heart attack. He asks Pidge: "Pidge, put down that gun and get some water for your grandfather." Pidge presumably brings the water. Later, just before the climax, the line "Oh, butterfingers!" appears. This implies something was dropped or spilled, likely the water. However, the transcript does not explicitly state who exclaimed "Oh, butterfingers!" or who spilled the water. Since the transcript does not explicitly confirm that Peter Benson spilled the water, the statement cannot be verified. **FALSE** *(Ground-truth: TRUE)*

Prediction on Fib: The transcript mentions Pidge being asked to get water for his grandfather, Peter Benson (Pop), when Pop feigns a heart attack [..] However, the transcript does not state that Peter Benson subsequently spilled this water, or any water, on the floor. [...] it is not explicitly stated who Jud is referring to or that the fumbling involved spilling water. The context is the electrocution and shooting, not an earlier incident of spilling water. There is no direct statement or clear implication in the transcript that Peter Benson accidentally spilled water on the floor. **FALSE** *(Ground-truth: FALSE)*

Table 9: **Illustrative Example 1 from the movie "Suddenly"**. Gemini fails to predict both claims correctly in every modality setting.

Example 1 (Part 2) - Movie: *"Suddely"*

Video only setting

Prediction on Fact: The grandfather, Peter Benson, is seen at 2:17 reaching for a glass of water. He then appears to deliberately drop the glass, causing the water to spill on the floor. This creates a distraction, allowing the woman to crawl towards a gun. The context of the scene, where the family is being held hostage and looking for an opportunity to fight back, combined with the grandfather's actions and the subsequent events, strongly suggests the spill was intentional. **TRUE** *(Ground-truth: TRUE)*

Prediction on Fib: The video introduces an older character who is later identified as Peter Benson (around 0:46-0:47). Much later in the video, during a tense hostage situation, Peter Benson is seen sitting in an armchair holding a glass of water. At timestamp 2:18, he is startled and gets up abruptly, causing the water in his glass to spill onto the floor. The context of the scene, where he reacts to the unfolding events, suggests the spill was accidental due to his sudden movement or reaction. **TRUE** *(Ground-truth: FALSE)*

Video w/ subs setting

Prediction on Fact: The grandfather, Peter Benson, is seen getting up with a glass of water around timestamp 55:20. As he walks, he stumbles and drops the glass, spilling the water on the floor near the table leg and electrical wiring. Immediately after, he clutches his chest and feigns a heart attack. The transcript reflects this with dialogue about his pain. There is no visual or textual evidence to suggest the spill was intentional; rather, it appears accidental, possibly due to his feigned or real physical distress, and he then uses the moment to further distract the captors by exaggerating his heart condition.**TRUE** *(Ground-truth: TRUE)*

Prediction on Fib: The grandfather, Peter Benson, takes a glass of water and walks towards the table where the assassin has set up the rifle. He places the glass on the edge of the metal table (around 55:23). As he moves to talk to Jud, who is working on the TV, the glass tips over and spills the water on the floor (around 55:24). The spilling appears to be an unintentional consequence of his actions and the precarious placement of the glass. Therefore, it was an accident.**TRUE** *(Ground-truth: FALSE)*

Table 10: **Illustrative Example 2 from "Sabotage (Hitchcock)"**. Gemini predicts the pair of claims incorrectly in every modality setting.

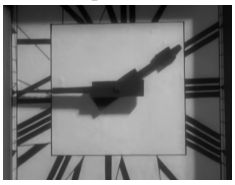
Example 2 (Part 1) - Movie: ***"Sabotage (Hitchcock)"***

Aquarium


Planning the sabotage


Meet in cinema


Kid enters the bus


The time arrives


Bus explosion


People in the cinema


Explosion in cinema

Fact (✓): The second bomb exploded in the cinema.
Fib (✗): There was a second bomb about to explode, but it didn't.

Granularity: Global
Context (**Spoiler alert !**): Mr. Verloc and the terrorist meet at the aqaurium, where the terrorist first mentions the plan to explode a bomb (he mentions "a parcel of fireworks" and asks Mr. Verloc to meet his friend the bombmaker). The terrorist meet again in the cinema, to discuss about the sabotage. Later in the movie the explosion happens inside the bus. While the second explosion happens inside the cinema. *Gemini fails to predict both claims correctly in every modality setting.*

Gemini's Predictions:

Subs only setting

Prediction on Fact: To determine if the statement is true or false, we need to identify the sequence of bomb-related incidents in the transcript. 1. **First Incident (Sabotage at the Cinema):** The transcript begins with a power outage at the cinema, described as "Sand. Sabotage. Wrecking. Deliberate." This is the first act of sabotage or a device causing disruption that Verloc is connected with. [...] This incident occurred at the cinema. 2. **Second Incident (Stevie's Bomb on the Bus):** Mr. Verloc is then tasked with a more serious act: to leave a "small parcel... a parcel of fireworks" (a bomb) in Piccadilly Circus. His young brother-in-law, Stevie, is sent to deliver this parcel. This bomb explodes prematurely. [...] This is the second distinct sabotage/bomb event in the plot, and this bomb exploded on a bus, not in the cinema. 3. **Third Incident (Bomb Explosion at the Cinema):** Towards the end of the transcript, the old man who makes the bombs is cornered in the cinema. [...] This bomb then explodes: [...] This explosion occurs at the cinema. Based on this chronological sequence of events: * The first destructive "bomb-like" event was the sabotage at the cinema (power outage). * The second bomb (Stevie's parcel) exploded on a bus in the West End. * The third bomb (detonated by the old man) exploded in the cinema. [...] Therefore, the statement is false. **FALSE** *(Ground-truth: TRUE)*

Prediction on Fib: The transcript describes two main bomb-related incidents. The first is the explosion that kills Stevie. The second occurs at the cinema where the old man (the bomb maker) threatens to "blow the place sky high." The transcript later confirms an explosion there: "The roof's gone. The whole of the back of the place - clean blown off." This indicates the second bomb *did* explode. Therefore, the statement that a second bomb was "about to explode, but it didn't" is false because it did explode. **FALSE** *(Ground-truth: FALSE)*

Table 11: **Illustrative Example 2 from "Sabotage (Hitchcock)"**. Gemini predicts the pair of claims incorrectly in every modality setting.

Example 2 (Part 2) - Movie: *"Sabotage"*

Video-only setting

Prediction on Fact: The video shows one clear explosion involving a bus on a street around the 1:46 mark. Following this, there are various scenes, including one where people are watching a cartoon in a cinema. However, there is no depiction or indication of a second bomb exploding, nor is there any explosion shown occurring inside the cinema. **FALSE** *(Ground-truth: TRUE)*

Prediction on Fib: The video montage shows scenes from Alfred Hitchcock's "Sabotage." The primary bomb plot shown results in an explosion on a bus around 1:46. Later in the montage, around 2:17-2:19, the Professor (the bomb maker) is seen suspiciously entering the cinema building. This implies he might be carrying a second bomb and intending to cause another explosion. However, the video montage ends shortly after this, at 2:23, without showing what happens to the Professor or if there was indeed a second bomb that was "about to explode." Crucially, the video does *not* show this second bomb failing to explode or being prevented from exploding. Since the video does not provide any information about the outcome of this potential second bomb (neither showing it explode nor showing it *not* explode), the assertion "but it didn't" is not supported by the available video content. Therefore, the statement cannot be confirmed as true based on the provided video. **FALSE** *(Ground-truth: FALSE)*

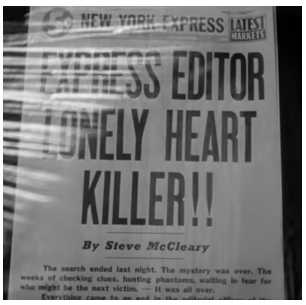
Video w/ subs setting

Prediction on Fact: The first act of sabotage was the power station blackout. The second planned act of sabotage, referred to as "a parcel of fireworks", was to be left in the cloakroom at Piccadilly Circus underground station [...] This parcel, carried by Stevie on a bus, explodes on the bus at 1:45. A newspaper headline later confirms "WEST END BOMB SENSATION. Fully-laden bus was blown up in Piccadilly this afternoon". [...] An explosion then occurs at the cinema, as described by an officer: "The roof's gone. The whole of the back of the place - clean blown off". This is a separate, third bomb event. Therefore, the second bomb (the one Stevie was carrying) exploded on the bus, not in the cinema. The explosion in the cinema was a distinct event caused by a different bomb wielded by the bird shop owner. **FALSE** *(Ground-truth: TRUE)*

Prediction on Fib: The statement claims: "There was a second bomb about to explode, but it didn't." Let's analyze the bombs in the film: 1. The first act of sabotage is the power station blackout. While not explicitly shown as a bomb explosion, it is an act of "wilful destruction." 2. The second major plot device is the bomb intended for Piccadilly Circus on Lord Mayor's Show day. Verloc gives this bomb to Stevie. This bomb explodes prematurely on the bus, killing Stevie and many others [...]. So this bomb *did* explode. 3. The third instance involving a bomb is at the cinema, where the Professor (the bomb maker) is cornered by the police. [..] This bomb *does* explode [...]. In all instances where a bomb is primed or ready for its purpose within the narrative, it either explodes (bus bomb, Professor's bomb in the cinema) or fulfills its sabotage (power station). There is no scene in the film where a bomb is "about to explode" and then "didn't." The bombs that are set up to go off, do go off. Therefore, the statement is false. **FALSE** *(Ground-truth: FALSE)*

Table 12: **Illustrative Example 3 from "Scandal Sheet"**. The pair of claims relies primarily on visual information. Gemini predicts them incorrectly when given only a single modality, but when provided with both the transcript and the video, it correctly identifies the pair of claims.

Example 3 (Part 1) - Movie: ***"Scandal Sheet"***



News about the killer

Sales increasing..

Sales increasing..

Fact (✓): The sales of the newspaper increased after it was revealed that the editor was the killer.
Fib (✗): The sales of the newspaper decreased after it was revealed that the editor was the killer.

Granularity: Single-scene

Context (Spoiler alert !): In the film's climax, Steve McCleary discovers that his editor-in-chief, Mark Chapman, is the "Lonely Hearts Killer." After Chapman is exposed and confronted, McCleary—now acting as editor—publishes the full story, revealing Chapman's true identity and crimes. Following the revelation, the newspaper's sales surge dramatically. *This pair of claims relies primarily on visual information. Gemini predicts them incorrectly when given only a single modality, but when provided with both the transcript and the video, it correctly identifies the pair of claims.*

Gemini's Predictions:
Subs only setting

Prediction on Fact: The transcript ends with Mark Chapman, the editor, being apprehended as the killer. He tells the reporter McCleary to "Write it up big, kid. It'll sell a lot of papers." This indicates an expectation that the story will boost sales, but the transcript concludes before any information about the actual sales figures *after* this revelation is provided. Therefore, the statement that sales *increased* after the revelation cannot be verified from the given text. **FALSE** *(Ground-truth: TRUE)*

Prediction on Fib: The transcript ends with Mark Chapman, the editor, being apprehended as the killer. His last words are, Ẅrite it up big, kid. It'll sell a lot of papers.T̈his is a prediction about future sales, not a statement of what actually happened to the sales. The transcript provides no information about the newspaper's sales figures *after* it was revealed that the editor was the killer. Therefore, the statement cannot be confirmed as true based on the provided text. **FALSE** *(Ground-truth: FALSE)*

Table 13: **Illustrative Example 3 from "Scandal Sheet"**. The pair of claims relies primarily on visual information. Gemini predicts them incorrectly when given only a single modality, but when provided with both the transcript and the video, it correctly identifies the pair of claims.

Example 3 (Part 2) - Movie: *"Scandal Sheet"*

Video only setting

Prediction on Fact: The video shows a chart of the "NET PAID CIRCULATION NEW YORK EXPRESS" at various points. Initially, at 0:10, the circulation is shown with an upward trend. After the editor is revealed to be the killer and is subsequently shot (around 2:26-2:28), the final shot of the movie (2:29) shows a newspaper headline "KILLER!!" with the byline "By Steve McCleary" superimposed over the same circulation chart. At this point, the line on the chart indicating sales has reached its highest point, implying a significant increase in sales coinciding with or immediately following the revelation and capture of the killer editor. **TRUE** *(Ground-truth: TRUE)*

Prediction on Fib: The movie shows a chart of the "New York Express" net paid circulation at 0:10, indicating an upward trend in sales from October to September, reaching 650,000. Later, after the editor (the man standing by the chart at 0:10) is revealed as the killer and is shot (around 2:27), the final shot at 2:29 shows a newspaper headline "KILLER!!" superimposed over the *same* circulation chart. The numbers and the upward trend on this chart are identical to the one shown at 0:10. The movie does not provide any information or visual evidence suggesting that the newspaper's sales decreased after the editor was revealed as the killer. The chart displayed at the end is the same one that showed increasing sales. **TRUE** *(Ground-truth: FALSE)*

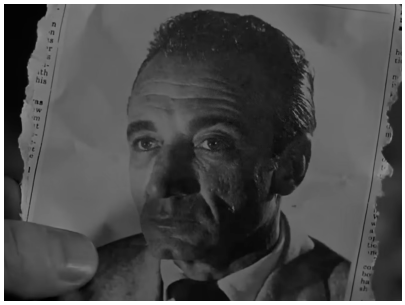
Video w/ subs setting

Prediction on Fact: The statement's truthfulness hinges on whether newspaper sales increased after the revelation that the editor, Mark Chapman, was the killer. 1. At 3:40, Mark Chapman discusses the newspaper's circulation, showing a graph and mentioning that when circulation reaches 750,000, a big bonus will be paid. The graph shows circulation rising, with the last point around June/July showing approximately 620,000-650,000. 2. The revelation that Mark Chapman is the killer unfolds towards the end of the film. 3. At 59:39, the newspaper headline "EXPRESS EDITOR LONELY HEART KILLER!!" by Steve McCleary is shown. 4. Immediately following this (59:45 - 59:49), a large circulation meter for the "N.Y. EXPRESS" is shown. A hand physically moves the needle on the meter, stopping at 750,000. "The End" is then superimposed over this image. This sequence strongly implies that the story revealing the editor as the killer led to the circulation reaching 750,000. Earlier, Chapman had also told McCleary to write the story big because "It'll sell a lot of papers" (59:06). The visual of the circulation hitting 750,000 immediately after the headline about the editor being the killer is the final image before "The End," suggesting this was the peak sales driven by this sensational story. **TRUE** *(Ground-truth: TRUE)*

Prediction on Fib: At the end of the movie, after the editor, Mark Chapman, is revealed as the killer and the headline "EXPRESS EDITOR LONELY HEART KILLER!!" is shown (59:40-59:43), the circulation meter for the New York Express is shown. A hand moves the needle to 750,000 (59:45). Earlier in the film (around 4:24), it was mentioned that reaching 750,000 in circulation would trigger a big bonus. The film ends with the circulation at this high number, indicating that sales reached this target *after* the revelation about the editor. There is no indication that sales decreased; in fact, the implication is that this sensational story would boost sales, as McCleary says, "Write it up big, kid. It'll sell a lot of papers." (59:14-59:16). **FALSE** *(Ground-truth: FALSE)*

Table 14: **Illustrative Example 4 from "Tight Spot"**. Gemini fails to predict both claims correctly when given only the subtitles or only the visual cues. However, when provided with both modalities together, it correctly predicts both claims.

Example 4 (Part 1) - Movie: ***"Tight Spot"***



Fact (✓): When Hallet shows Conley a photograph of an individual who was on the yacht, she immediately recognizes the person.
Fib (✗): When Hallet shows Conley a photograph of an individual who was on the yacht, she claims not to recognize the person.

Granularity: Single-scene

Context (Spoiler alert !): This pair of claims to be answered correctly requires *both* the *visual* and *text* cues. Hallet shows Conley a photograph (visual information) and asks whether she recognizes the person (Senior Maracelli). Conley recognizes the person. *When the model is given only the subtitles, it misclassifies both the fact and the fib (as visual cues are absent). In contrast, when provided with both the video and the subtitles, it predicts each one correctly.*

Gemini's Predictions:
Subs only setting

Prediction on Fact: The statement claims that "When Hallet shows Conley a photograph of an individual who was on the yacht, she immediately recognizes the person." We need to check two things in the transcript: 1. Does Hallett show Conley a photograph of an individual from the yacht? 2. If so, does Conley immediately recognize the person? In the transcript, the relevant exchange is: HALLETT: "And at one of these islands, this man came aboard the yacht." CONLEY: "Signor Maracelli. The Italian garden expert. He taught me to say his name himself." [...] *The transcript does not use the word "photograph" or describe the action of showing one in this specific interaction.* Since the statement's premise that Hallett shows Conley a photograph is not explicitly confirmed by the transcript, a part of the statement is not verifiably true based on the provided text. **FALSE** *(Ground-truth: TRUE)*

Prediction on Fib: When Mr. Hallett mentions Signor Maracelli coming aboard the yacht, Conley immediately identifies him as "The Italian garden expert" and adds, "He taught me to say his name himself." This indicates she clearly recognizes the person Hallett is referring to (and presumably showing a photograph of, though the act of showing the photo isn't explicitly stated, Hallett's line "this man came aboard the yacht" implies he's identifying someone specific). She does not claim not to recognize him. **FALSE** *(Ground-truth: FALSE)*

Table 15: **Illustrative Example 4 from "Tight Spot"**. Gemini fails to predict both claims correctly when given only the subtitles or only the visual cues. However, when provided with both modalities together, it correctly predicts both claims.

Example 4 (Part 2) - Movie: ***"Tight Spot"***

Video only setting

Prediction on Fact: The statement asks whether Conley immediately recognizes the person in the photograph shown to her by Hallet. In the video, at 00:30, Hallet takes out a photograph. At 00:31, he shows it to Conley. Conley looks down at the photo, then at 00:32, she looks up at Hallet with an expression that seems more contemplative or questioning rather than one of instant, clear recognition. She then glances back at the photo before looking at Hallet again and starting to speak at 00:33. Her reaction involves a pause and a sequence of looks, which does not convey *immediate* recognition. "Immediate" suggests an instantaneous, reflexive understanding, which is not evident in her more measured response. **FALSE** *(Ground-truth: TRUE)*

Prediction on Fib: The statement claims that when Hallet shows Conley a photograph of an individual from a yacht, Conley claims not to recognize the person. Observing the video, there is a scene from approximately 0:29 to 0:36 where a man (presumably Hallet, played by Edward G. Robinson) shows a photograph to a woman (presumably Conley, played by Ginger Rogers). The woman looks at the photograph (0:31-0:32). However, the video segment does not include any audible dialogue or clear visual cue (like shaking her head definitively or mouthing "I don't know") from the woman that would confirm she "claims not to recognize the person." Her facial expression is somewhat pensive or surprised but not an explicit denial of recognition. She then proceeds to make a phone call (0:34). The context that the photograph is of "an individual who was on the yacht" is also not explicitly provided by the visual information in this clip. Since the crucial part of the statement - her claiming not to recognize the person - is not depicted or audible in the provided video evidence, the statement cannot be verified as true based on the available information. **FALSE** *(Ground-truth: FALSE)*

Video w/ subs setting

Prediction on Fact: When Hallett shows Conley the photograph (around 00:40), he identifies the man as "Signor Maracelli." Conley immediately confirms recognition by saying, "The Italian garden expert. He taught me to say his name himself." This demonstrates she knew who the person was. **TRUE** *(Ground-truth: TRUE)*

Prediction on Fib: When Mr. Hallett mentions Signor Maracelli coming aboard the yacht, Conley immediately identifies him as "The Italian garden expert" and adds, "He taught me to say his name himself." This indicates she clearly recognizes the person Hallett is referring to (and presumably showing a photograph of, though the act of showing the photo isn't explicitly stated, Hallett's line "this man came aboard the yacht" implies he's identifying someone specific). She does not claim not to recognize him. **FALSE** *(Ground-truth: FALSE)*

Table 16: **Illustrative Example 5 from "Tight Spot"**. Gemini fails to predict the pair of claims correctly, in every modality setting.

Example 5 (Part 1) - Movie: ***"Tight Spot"***

Pete in taxi

Walking the stairs of the church

Pete being shot

Fact (✓): Pete Tinelli was shot.
Fib (✗): Pete Tinelli was poisoned.

Granularity: Multi-scene

Context (Spoiler alert !): Early in the film, we are introduced to Pete during a scene in which he rides in a taxi with two other men. Later, the group arrives in front of a church, and as they walk up the steps, Pete is suddenly shot while the other two men look around, trying to identify the shooter. *This pair of claims requires both visual and text cues to be resolved correctly, as the model needs to understand who is Pete (in the first scene) and connect it with his murder (being shot). Gemini fails to predict correctly the pair of claims in every modality setting*.

Gemini's Predictions:

Subs only setting

Prediction on Fact: The transcript states that "Pete's dead" and that "He was murdered by the Costain mob." [..] However, the transcript does not specify the method of murder. It does not explicitly say that Pete Tonelli was shot. Therefore, because the specific detail of "shot" is not confirmed in the transcript, the statement is false. **FALSE** *(Ground-truth: TRUE)*

Prediction on Fib: The transcript states, "Pete's dead... He was murdered by the Costain mob." It does not specify the method of murder. Since the transcript doesn't mention poisoning as the cause of Pete Tonelli's death, and only confirms he was murdered, the statement that he was poisoned cannot be confirmed as true based on the provided text. **FALSE** *(Ground-truth: FALSE)*

Video only setting

Prediction on Fact: A man is visibly shot around the 1:57-1:58 mark in the video. In this scene, a man is held by two others, another man (resembling Edward G. Robinson) nods, a gunshot sound is heard, and the man being held slumps. However, throughout the provided video clip, there is no dialogue, on-screen text, or any other information that identifies the man who was shot as "Pete Tinelli." [...] **FALSE** *(Ground-truth: TRUE)*

Prediction on Fib: The video contains several scenes that lead to the conclusion that Pete Tinelli was poisoned. At 1:58, a group of men are shown in what appears to be a garage, and one man is forcibly made to drink something. He subsequently collapses at 2:01. Later, at 2:05, a woman asks, "What happened to Pete Tinelli?" and is told, "He won't bother you anymore." Shortly after, at 2:08, one man informs another, "Pete Tinelli's been taken care of." When asked "How?", the response is "Poison." This sequence of events and dialogue directly confirms that Pete Tinelli was poisoned.**TRUE** *(Ground-truth: FALSE)*

Table 17: **Illustrative Example 5 from "Tight Spot"**. Gemini fails to predict the pair of claims correctly, in every modality setting.

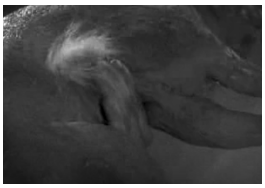
Example 5 (Part 2) - Movie: *"Tight Spot"*
Video w/ subs setting
Prediction on Fact: The provided transcript states at 00:54-00:55 that "Pete's dead" and "He was murdered by the Costain mob." However, neither the transcript nor the accompanying video footage specifies the method of his murder. The statement "Pete Tinelli was shot" asserts a specific cause of death (being shot) which is not explicitly mentioned or shown in the provided materials. Since the method of murder is not confirmed, we cannot ascertain that he was shot based solely on the given information. **FALSE** *(Ground-truth: TRUE)*
Prediction on Fib: The transcript states at 00:42, "He was murdered by the Costain mob." While this confirms Pete Tonelli was murdered, it does not specify the method of murder. The statement claims he was poisoned. Since the available information does not mention poisoning as the cause of death, the specific claim that he was poisoned cannot be verified as true from the provided movie and transcript. **FALSE** *(Ground-truth: FALSE)*

Table 18: **Illustrative Example 6 from the movie: "A Bucket of Blood"**. Gemini fails to predict the pair of claims correctly, in every modality setting.

Example 6 (Part 1) - Movie: *"A Bucket of Blood"*

Holding the cat sculpture

Drops it



The damage

Moment of realization

Fact (✓): The gallery owner dropped the cat sculpture, causing minor damage.

Fib (✗): The gallery owner dropped the cat sculpture, but it remained intact.

Granularity: Single-scene

Context (Spoiler alert !): When the gallery owner accidentally drops the cat sculpture in the cafe and sees real flesh beneath the clay, he realizes that the sculpture is not a realistic imitation but an actual dead cat covered in clay. *Gemini does not correctly predict the claim pair, as it fails to recognize that the sculpture was damaged.*

Gemini's Predictions:
Subs-only setting

Prediction on Fact: The transcript details the creation of the "Dead Cat" sculpture by Walter and its reception at the Yellow Door, owned by Leonard De Santis. Leonard agrees to display the cat sculpture. Later, a customer takes the sculpture home to show his wife and then makes an offer to buy it for $500. Leonard tells this customer, "Someone has the cat just now, but I'll have him back in a few days. Then you can have it for the $500." There is no mention in the transcript of Leonard, the gallery/club owner, dropping the cat sculpture or of it sustaining any damage. **FALSE**

Prediction on Fib: The transcript mentions that Mr. De Santis, the owner of the Yellow Door (where the art is displayed, akin to a gallery owner in this context), arranged for a potential buyer to take the cat sculpture home to show his wife. It states, "A man wanted to pay me $100 for the cat. In fact, he's taking it home to show to his wife." Later, the same man calls back to buy it, and Mr. De Santis says, "Someone has the cat just now, but I'll have him back in a few days." There is no mention in the transcript of Mr. De Santis or anyone else dropping the cat sculpture. Therefore, the premise that the gallery owner dropped it is false. **FALSE**

Table 19: **Illustrative Example 6 from the movie: "A Bucket of Blood"**. Gemini fails to predict the pair of claims correctly, in every modality setting.

Example 6 (Part 2) - Movie: *"A Bucket of Blood"*

Video-only setting

Prediction on Fact: The statement is: "The gallery owner dropped the cat sculpture, causing minor damage." To verify this, we need to identify a "gallery owner," a "cat sculpture," an act of "dropping" by the gallery owner, and resulting "minor damage." 1. **Gallery Owner:** * Leonard de Santis is the owner of "The Yellow Dog" cafe, which also functions as an art gallery where poetry is read and art is displayed [...] Walter Paisley creates a sculpture of a cat by encasing his landlady's dead cat in clay (0:16-0:20). This is the most prominent cat sculpture. * At 0:29, Leonard and Carla are seen holding other, more stylized white cat sculptures, indicating these might be common art pieces in their circle. 3. **Dropped the cat sculpture, causing minor damage:** * At 0:29, Leonard (the cafe/gallery owner) is shown holding one of the stylized white cat sculptures. He does not drop it. [...] No scene in the provided video shows a character identifiable as "the gallery owner" dropping a cat sculpture and causing minor damage. The closest event involves a police officer tapping Walter's cat sculpture, causing a tiny chip. Therefore, the statement is false. **FALSE**

Prediction on Fib: The statement says: "The gallery owner dropped the cat sculpture, but it remained intact." Let's break this down: 1. **Identity of the "gallery owner"**: In the movie clip, the gallery owner is Maxwell Brock, often seen wearing a beret or a suit, and he runs "The Yellow Door" gallery. 2. **Did the gallery owner drop a cat sculpture?**: * At 0:27, the main character, Walter Paisley (a busboy who becomes an artist), is carrying a large abstract sculpture (not a cat) when he is bumped by Carla. Walter drops this sculpture, and it shatters. Walter is not the gallery owner. * Walter Paisley creates a sculpture of a cat (by encasing his landlady's dead cat in clay) around 0:22. This sculpture is later handled but not dropped by the gallery owner. * At 0:50, Maxwell Brock (the gallery owner) is seen handling a *different* white, stylized cat sculpture. He picks it up, examines it carefully, and places it back down. He does *not* drop this sculpture. Since there is no instance in the provided video where the gallery owner (Maxwell Brock) drops any cat sculpture, the first part of the statement ("The gallery owner dropped the cat sculpture") is false. Therefore, the entire statement is false. **FALSE**

Video w/ subs setting

Prediction on Fact: The video and transcript show the gallery owner, Leonard, handling the cat sculpture on a couple of occasions. First, when Walter initially brings it to "The Yellow Door" (around 13:36-13:55), Leonard is skeptical but eventually agrees to display it. He takes the sculpture from Walter and places it on a surface without dropping it. Later, after the sculpture gains attention (around 25:05), Leonard is seen picking up the cat sculpture, looking at it, and then placing it back down on a table, again without dropping or damaging it. The transcript contains no mention of the gallery owner dropping the sculpture. **FALSE**

Prediction on Fib: The statement claims that the gallery owner dropped the cat sculpture. Reviewing the video and transcript, at 13:47, Leonard De Santis, the owner of the Yellow Door gallery/cafe, picks up the cat sculpture made by Walter Paisley. Shortly after, at 13:48, Walter Paisley himself is holding the sculpture. At 13:49, Walter Paisley (the artist, not the gallery owner) drops the cat sculpture onto the floor. The sculpture appears to remain intact. Since it was Walter, the artist, and not Leonard, the gallery owner, who dropped the sculpture, the first part of the statement is false. **FALSE**