Video SimpleQA: Towards Factuality Evaluation in Large Video Language Models

Meng Cao[1*], Pengfei Hu[1,2*], Yingyao Wang[2], Jihao Gu[2], Haoran Tang[3], Haoze Zhao[1], Chen Wang[2], Jiahua Dong[1], Wangbo Yu[3], Ge Zhang[4], Xiang Li[2], Ian Reid[1], Xiaodan Liang[1,5]

[1]MBZUAI
[2]Alibaba Group
[3]Peking University
[4]ByteDance Inc.
[5]Sun Yat-sen University
xiaodan.liang@mbzuai.ac.ae

Abstract

Recent advancements in Large Video Language Models (LVLMs) have highlighted their potential for multi-modal understanding, yet evaluating their factual grounding in videos remains a critical unsolved challenge. To address this gap, we introduce `Video SimpleQA`, the first comprehensive benchmark tailored for factuality evaluation in video contexts. Our work differs from existing video benchmarks through the following key features: 1) **Knowledge required**: demanding integration of external knowledge beyond the video's explicit narrative; 2) **Multi-hop fact-seeking question**: Each question involves multiple explicit facts and requires strict factual grounding without hypothetical or subjective inferences. We also include per-hop single-fact-based sub-QAs alongside final QAs to enable fine-grained, step-by-step evaluation; 3) **Short-form definitive answer**: Answers are crafted as unambiguous and definitively correct in a short format with minimal scoring variance; 4) **Temporal grounded required**: Requiring answers to rely on one or more temporal segments in videos, rather than single frames. We extensively evaluate 33 state-of-the-art LVLMs and summarize key findings as follows: 1) Current LVLMs exhibit notable deficiencies in factual adherence, with the best-performing model o3 merely achieving an F-score of 66.3%; 2) Most LVLMs are overconfident in what they generate, with self-stated confidence exceeding actual accuracy; 3) Retrieval-augmented generation demonstrates consistent improvements at the cost of additional inference time overhead; 4) Multi-hop QA demonstrates substantially degraded performance compared to single-hop sub-QAs, with first-hop object/event recognition emerging as the primary bottleneck. We position Video SimpleQA as the cornerstone benchmark for video factuality assessment, aiming to steer LVLM development toward verifiable grounding in real-world contexts.

Introduction

The substantial advancements in Large Language Models (LLMs) (Achiam et al. 2023; Reid et al. 2024; Touvron et al. 2023) over the past few years have inaugurated a new frontier in artificial intelligence. Despite their remarkable capabilities, the factuality concern (Wang et al. 2024b; Akhtar et al. 2023; Wang et al. 2023) remains a critical challenge, *i.e.*, how to ensure that the generated contents are consistent with factual knowledge and grounded in credible sources.[1]

Existing research has primarily focused on evaluating factuality in text-based (Yu et al. 2022; Pan et al. 2024; Lin, Hilton, and Evans 2022; Chern et al. 2023; Gou et al. 2023) and image-based (Marino et al. 2019; Wang et al. 2015, 2017; Zellers et al. 2019; Jain et al. 2021) scenarios. Recently, the SimpleQA benchmark (Wei et al. 2024) introduced by OpenAI and its subsequent works (He et al. 2024b; Gu et al. 2025; Cheng et al. 2025) streamline the factuality evaluation by considering only concise and fact-seeking questions, which enables standardized and tractable assessments. However, extending this paradigm to video contexts is under-explored and presents unique challenges due to the inherent temporal dynamics and procedural knowledge. To bridge this gap, we present `Video SimpleQA`, a comprehensive factuality evaluation benchmark tailored for Large Video Language Models (LVLMs). As shown in Figure **??** and Figure 2, Video SimpleQA is composed of multi-hop fact-seeking questions and short-form definitive answers. Compared to previous video benchmarks, Video SimpleQA stands out with the following advancements:

- **Knowledge required**: Beyond comprehending the visual content, Video SimpleQA necessitates the integration of external knowledge that is not explicitly presented in the video narrative, *e.g.*, domain-specific information, contextual background, commonsense.
- **Multi-hop fact-seeking question**: Questions necessitate strict adherence to factual grounding principles, *eliminating any hypothetical or subjective inferences*. In addition, each question is constructed to involve *multiple explicitly identifiable facts*. To achieve this, beyond the final multi-hop question-answer (QA) pairs, we additionally provide per-fact specific sub-QA annotations, which fa-

*These authors contributed equally.

[1]Please refer to (Wang et al. 2023, 2024b) for the differentiation between the *factuality* and the similar *hallucination* concepts.

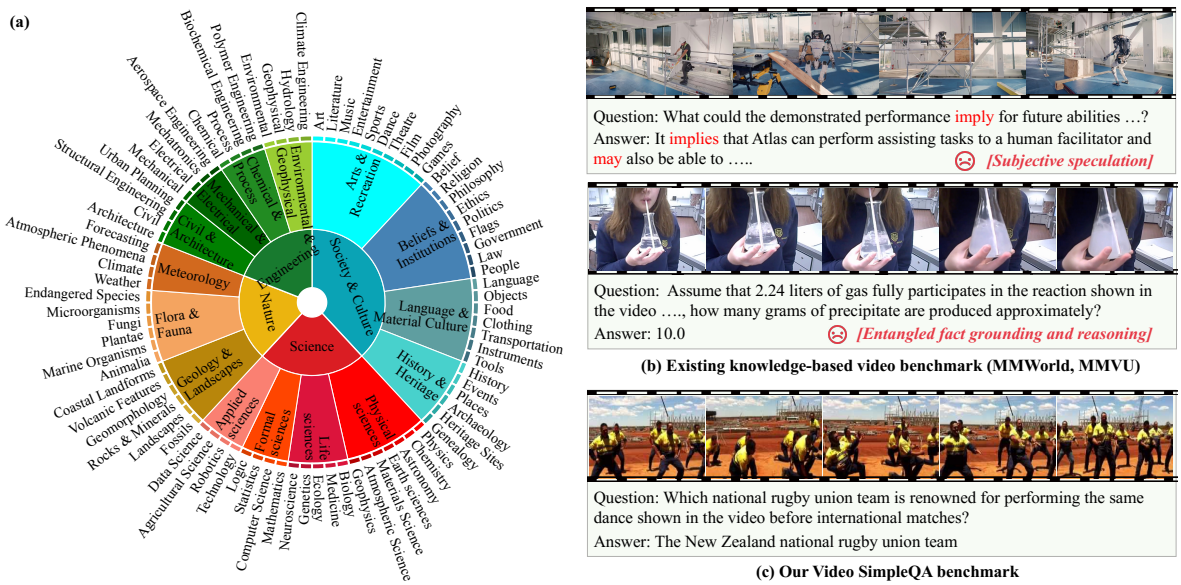


Figure 1: (a) The taxonomy of Video SimpleQA benchmark; (b) Illustrations of existing knowledge-based video benchmarks (Zhao et al. 2025; He et al. 2024a) which may contain subjective speculation or conflate factual grounding with reasoning skills (*i.e.*, mathematical calculation); (c) Illustrations of our Video SimpleQA benchmark with the fact-seeking question and definitive & short-form answer with multi-hop external facts verified.

Benchmarks	Video domain	Knowledge driven	Objective QA	Factuality exclusive	Multi-hop fact decomp.	Evidence source
Video-MME (Fu et al. 2024)	Open	✗	✗	✗	✗	✗
MMBench-Video (Fang et al. 2024)	Open	✗	✗	✗	✗	✗
Video-MMMU (Hu et al. 2025)	Professional	✓	✗	✗	✗	✗
MMVU (Zhao et al. 2025)	Discipline	✓	✗	✗	✗	✓
MMWorld (He et al. 2024a)	Discipline	✓	✗	✗	✗	✗
WorldQA (Zhang et al. 2024b)	Open	✓	✗	✗	✗	✗
KnowIT-VQA (Garcia et al. 2020)	TV shows	✓	✗	✗	✗	✗
Video SimpleQA	Open	✓	✓	✓	✓	✓

Table 1: Benchmark comparisons across key dimensions: video domain scope, knowledge-driven focus, objective factuality focus, exclusivity of factual evaluation, multi-hop fact decomposition, and external evidence source.

cilitate fine-grained evaluation of model performance at each fact-grounding hop and help pinpoint exactly which hop fails in factual grounding (*c.f.* Figure 2).

- **Short-form definitive answer**: All the answers are unambiguous, universally agreed upon, consistent over time, and invariant to individual perspectives. Following SimpleQA (Wei et al. 2024), the answers also advocate the short-form paradigm, which establishes reliable factual assessment with low run-to-run variance.
- **Temporal grounded:** Answering questions in Video SimpleQA should refer to one or more temporal segments in the video, rather than relying on a single frame.

While existing knowledge-based (Garcia et al. 2020; Zhang et al. 2024b; Hu et al. 2025) and recent discipline-based (Zhao et al. 2025; He et al. 2024a) benchmarks may appear similar to our Video SimpleQA , our benchmark features several distinct characteristics (*c.f.* Table 1):

- **Open-domain**: While KnowIT-VQA (Garcia et al. 2020) is constrained to TV shows, and MMVU (Zhao et al. 2025) as well as MMWorld (He et al. 2024a) focus on discipline-specific knowledge, our Video SimpleQA encompasses open-domain video types and questions.
- **Objective QA**: Our benchmark is explicitly designed for factuality evaluation through **objective** factual assertions, in contrast to existing benchmarks that often involve varying degrees of subjectivity, even those focusing on disciplinary knowledge. For instance, as shown in Figure 1(b) top, MMWorld (He et al. 2024a) includes cases requiring predictions about a robot's future capabilities—introducing subjective speculation and personal judgment, which deviates from our goal of evidence-based and objective evaluation.[2]
- **Factuality exclusive**: Discipline-based benchmarks of-

[2]More examples are available in the supplementary material.

sub-question-1: Which short science-fiction film is excerpted in the video?
sub-answer-1: Tears of Steel
ref. source-1: https://commons.wikimedia.org/wiki/Template:Motd/2013-01-26

sub-question-2: Who is the director of Tears of Steel?
sub-answer-2: Ian Hubert
ref. source-2: https://en.wikipedia.org/wiki/Tears_of_Steel

sub-question-3: Which country is Ian Hubert from?
sub-answer-3: USA
ref. source-3: https://baike.baidu.com/item/Ian%20Hubert

sub-question-4: Who is the first holder of Chief Justice of USA?
sub-answer-4: John Jay
ref. source-4: https://en.wikipedia.org/wiki/John_Jay

multi-hop-question: Who is the first holder of Chief Justice of the country which is the director of the short science-fiction excerted in the video from?
multi-hop-answer: John Jay

sub-question-1: Which country is the tourist attraction in the video located in?
sub-answer-1: Greece
ref. source-1: https://commons.wikimedia.org/wiki/Template:Motd/2011-01-16

sub-question-2: What is the capital of Greece?
sub-answer-2: Athens
ref. source-2: https://en.wikipedia.org/wiki/Greece

sub-question-3: What peninsula is Athens located on?
sub-answer-3: Balkan peninsula
ref. source-3: https://en.wikipedia.org/wiki/Athens

sub-question-4: Who coined the term "Balkan Peninsula"?
sub-answer-4: August Zeune
ref. source-4: https://en.wikipedia.org/wiki/Balkans

multi-hop-question: Who coined the term for the peninsula where the capital of the country containing the tourist attraction in the video is located?
multi-hop-answer: August Zeune

Figure 2: Four-hop examples in Video SimpleQA including the final multi-hop QA and the decomposed per-fact sub-QAs.

ten conflate external knowledge retrieval with reasoning skills (*e.g.*, numerical calculations). For example, the case in Figure 1(b) bottom from MMVU (Zhao et al. 2025) requires LVLMs to both recognize a chemical reaction in the video and perform numeric computations based on the question context. This *coupling* makes it difficult to pinpoint the error source—whether due to incorrect fact identification (*e.g.*, failing to detect the reaction) or faulty reasoning (*e.g.*, miscalculating). In contrast, Video SimpleQA **exclusively** focuses on fact identification, providing a clearer assessment of LVLMs' fact-grounding ability.[2]

- **Multi-hop fact decomposition**: As shown in Figure 2, Video SimpleQA includes not only the final multi-hop QA pairs but also the decomposed *per-fact sub-QAs*, enabling fine-grained evaluations. While some cases in MMVU (Zhao et al. 2025) also involve knowledge from multiple external sources, they do not provide such explicit per-fact decomposition, making it difficult to assess how each individual fact contributes to the final answer.

We conduct comprehensive evaluations of 33 state-of-the-art LVLMs on `Video SimpleQA`, revealing several critical insights: 1) **Significant performance gap**: Both proprietary and open-source LVLMs substantially underperform compared to human expertise; 2) **Overconfidence bias**: Most LVLMs exhibit systematic overconfidence in their predictions despite output inaccuracies, with notable variations in calibration quality (*c.f.* Figure 7); 3) **Efficiency-performance tradeoff**: Retrieval-Augmented Generation (RAG) yields significant gains at the cost of inference efficiency (*c.f.* Table 3); 4) **Multi-hop performance bottleneck**: Multi-hop QA performance significantly lags behind single-hop sub-tasks, with the initial video-grounded hop acting as the primary bottleneck (*c.f.* Table 4). More experimental findings are available in the supplementary material.

Related Work

Factuality Benchmarks. Factuality is the capability of LLMs to generate content that aligns with factual information, which can be substantiated by authoritative sources such as Wikipedia or textbooks (Akhtar et al. 2023; Wang et al. 2024b). Evaluating LLM factuality presents a non-trivial challenge and various benchmarks are proposed in the text-based (Lin, Hilton, and Evans 2022; Chern et al. 2023; Gou et al. 2023; Wei et al. 2024; He et al. 2024b) and image-based scenarios (Marino et al. 2019; Wang et al. 2017; Jain et al. 2021; Gu et al. 2025). As one of the pioneering works, TruthfulQA (Lin, Hilton, and Evans 2022) specifically targets imitative falsehoods in LLM-generated responses, which stem from erroneous preconceptions or knowledge gaps. Recently, the SimpleQA series of works (Wei et al. 2024; He et al. 2024b; Gu et al. 2025; Cheng et al. 2025) facilitate factuality evaluation by constraining the scope to short, fact-seeking questions with simple answers, making factuality assessment more tractable compared to previous long, open-ended model outputs. Despite of this, the community urgently needs a standard benchmark for trustworthy factuality evaluation *in video contexts*. To address this gap, our Video SimpleQA emerges.

Video Understanding Benchmarks. Recently, video benchmarks have been designed for evaluations in comprehensive tasks, including temporal perception (Li et al. 2024b), reasoning (Cai et al. 2024; Chen et al. 2024), navigation (Yang et al. 2024), long-form comprehension (Song et al. 2024; Chandrasegaran et al. 2024), *etc*. However, current video benchmarks largely overlook factuality evaluation, resulting in a lack of assessment for LVLMs' ability to generate factually accurate responses. Compared to video hallucination benchmarks (Wang et al. 2024c; Guan et al. 2024; Zhang et al. 2024a), which primarily assess models' adherence to video contents, our focused factuality evaluation emphasizes the model's alignment with verifiable external world knowledge (Wang et al. 2023, 2024b).

Differentiation from Knowledge-based and Discipline-base Benchmarks. Existing knowledge-based (Zhang et al. 2024b; Hu et al. 2025) and discipline-based benchmarks (Zhao et al. 2025; He et al. 2024a) either contain *hypothetical/subjective* reasoning (*e.g.*, the categories of `societal norms` and `social interactions` in WorldQA (Zhang et al. 2024b)) or *narrow their scopes* to single TV show (Garcia et al. 2020) or discipline-related

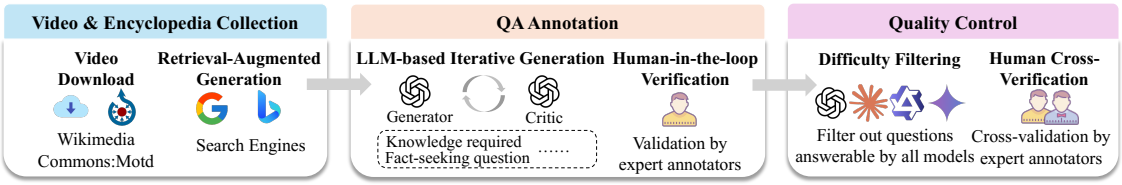


Figure 3: An overview of the construction pipeline of Video SimpleQA .

Associated Descriptions Barbary Ground Squirrel (Atlantoxerus getulus) at Paseo Maritimo in Antigua, Fuerteventura , Canary Islands.

RAG Results

1. Barbary Ground Squirrel: The Barbary ground squirrel (Atlantoxerus getulus) is a species of rodent in the family Sciuridae. It is monotypic within the genus Atlantoxerus. ……
2. Fuerteventura: Fuerteventura (Spanish: [ˌfwerteβenˈtura]) is one of the Canary Islands, in the Atlantic Ocean, geographically part of Macaronesia, and politically part of Spain. ……

Figure 4: The encyclopedia collection process including the raw associated descriptions in Wikimedia and the RAG results[3] for the specialized terms extracted by GPT-4o.

knowledge (Zhao et al. 2025; He et al. 2024a). Our Video SimpleQA addresses these limitations by enforcing *objective* factuality verification and ensuring *diversity* across various categories. In addition, we introduce per-hop fact-grounded QA for *fine-grained* evaluations.

Video SimpleQA

The construction pipeline of Video SimpleQA is illustrated in Figure 3, which includes video & encyclopedia collection, QA annotations, and quality control.

Video & Encyclopedia Collection

Video Collection: To ensure broad coverage, we curate the knowledge-intensive videos from the "Media of the Day" page of Wikimedia Commons[3] together with the accompanied brief descriptions or scientific illustrations. Note that files on the "Media of the Day" page are freely licensed, which avoids introducing any potential copyright concerns.

Encyclopedia Collection: As shown in Figure 4, although the associated textual descriptions on the Wikimedia page provide related descriptions, the explanations for the specialized terms (*e.g.*, `Barbary Ground Squirrel`, `Fuerteventura`) still lack formal definitions. To construct a more comprehensive encyclopedia, we leverage GPT-4o to extract key terms from the original descriptions and then obtain detailed explanations for these terms via RAG. Specifically, we apply LlamaIndex as the RAG method, with search results from Google and Bing as data sources.

[3]https://commons.wikimedia.org/wiki/Commons:Media_of_the_day

QA Annotations

The annotation pipeline for Video SimpleQA follows a two-stage process: (1) automated LLM-based iterative generation and (2) human-in-the-loop verification refinement.

LLM-based Iterative Generation: The iterative generation process involves two LLMs, a *generator* LLM for initial QA pair synthesis and a *critic* LLM for quality assessment. Specifically, the generator receives videos and encyclopedic knowledge to generate candidate QA pairs. Subsequently, the critic LLM systematically evaluates output compliance with predefined quality criteria, providing targeted feedback for refinement. This iterative process continues for up to three refinement cycles, with non-compliant outputs being discarded post-final iteration to ensure rigorous quality control. Both generator and critic are implemented as GPT-4o.

The explicit construction criteria are as follows: 1) *Knowledge required*: The questions should necessitate both video content and relevant external factual knowledge. Those that can be answered solely based on either source should be excluded. For example, two questions that should be eliminated are: `What color is the insect in the video?` (which relies solely on video content) and `Which president of the United States was Obama?` (which relies solely on external knowledge); 2) *Fact-seeking question*: The generated question should be factually grounded without any hypothetical or subjective reasoning; 3) *Definitive answer*: To ensure a rigorous evaluation, each question must have a single, unambiguous, and time-invariant answer. To achieve this, we explicitly define the level of granularity in question phrasing. For example, we use "which year" instead of "when" and "which city" instead of "where" to eliminate ambiguity; 4) *Short-form answer*: The answers should be in a short-form format; 5) *Multi-hop facts*: To answer the question, it requires involving multiple factual sources; 6) *Temporal grounded*: The questions are grounded in one or more video segments rather than specific frames.

Human-in-the-loop Verification: Through the iterative generation, we obtain QA annotations of reasonable quality. To further enhance the reliability, we train expert annotators to refine the LLM-generated QA annotations. The expert annotators are first required to watch the complete video and examine the collected encyclopedic knowledge. They then evaluate whether the LLM-generated QA annotations meet the specified criteria and manually revise them if necessary.

To ensure *multi-hop* fact grounding, annotators were additionally instructed to decompose each multi-hop QA into a series of sub-QAs (*c.f.* Figure 2). The decomposition follows two rules: 1) **Single fact per sub-QA**: Each sub-QA

Model	Overall results on 5 metrics					F-score on 4 primary categories			
	CO	IN↓	NA↓	CGA	F-score	ENG	NAT	SCI	SAC
Human Performance									
Human Open-book	87.0	5.0	8.0	94.6	90.6	85.2	83.7	89.1	96.8
Human Closed-book	59.0	18.5	22.5	76.1	66.5	58.4	52.8	54.2	80.6
Proprietary Multi-modal LLMs									
o4-mini	53.7	45.3	0.9	54.2	54.0	44.3	59.4	56.8	54.0
o3	**66.3**	**33.6**	**0.1**	**66.4**	**66.3**	**63.0**	**71.3**	63.5	**68.8**
GPT-4.5	52.9	42.5	4.6	55.4	54.1	49.5	57.5	57.9	51.4
GPT-4o	47.7	45.9	6.4	51.0	49.3	45.1	57.1	52.7	45.4
Claude Sonnet 4	32.8	51.2	16.0	39.0	35.6	33.0	34.3	37.9	35.0
Claude 3.7 Sonnet	32.6	47.3	20.1	40.8	36.2	24.2	40.3	41.9	34.5
Claude 3.5 Sonnet2	33.7	58.2	8.1	36.7	35.2	26.1	37.2	38.5	35.4
Claude 3.5 Sonnet	31.5	53.5	15.0	37.0	34.0	26.8	36.0	38.1	32.5
Gemini 2.5 Pro (Comanici et al. 2025)	61.2	34.3	4.5	64.1	62.6	53.5	65.8	**67.1**	61.5
Gemini 2.5 Flash (Comanici et al. 2025)	53.7	34.9	11.3	60.6	57.0	46.0	61.6	61.4	56.2
Qwen-VL-Max (Bai et al. 2023)	39.2	57.1	3.7	40.7	39.9	27.4	46.2	48.8	35.1
Qwen-VL-Plus (Bai et al. 2023)	21.9	63.3	14.7	25.7	23.7	10.5	25.6	30.5	21.8
Open-source Multi-modal LLMs									
InternVL3-78B (Zhu et al. 2025)	33.7	65.6	**0.7**	33.9	33.8	25.4	41.2	38.6	30.6
InternVL3-38B (Zhu et al. 2025)	31.4	67.7	0.9	31.7	31.5	21.3	33.3	35.7	31.8
InternVL3-14B (Zhu et al. 2025)	24.9	73.3	1.8	25.4	25.2	14.6	32.3	28.4	24.6
InternVL3-9B (Zhu et al. 2025)	22.6	72.9	4.5	23.7	23.1	12.8	33.2	27.9	19.9
InternVL3-8B (Zhu et al. 2025)	23.3	75.2	1.5	23.7	23.5	16.2	30.7	25.6	22.4
Qwen2.5-VL-72B (Bai et al. 2025)	**38.7**	57.3	4.0	**40.3**	**39.5**	**26.1**	**47.0**	**48.2**	**34.7**
Qwen2.5-VL-32B (Bai et al. 2025)	30.3	67.1	2.7	31.1	30.7	18.1	39.3	37.4	27.0
Qwen2.5-VL-7B (Bai et al. 2025)	24.7	71.2	4.1	25.8	25.3	13.8	25.6	30.8	25.1
Qwen2-VL-72B (Wang et al. 2024a)	32.7	59.0	8.3	35.7	34.2	20.2	39.0	40.0	33.2
Qwen2-VL-7B (Wang et al. 2024a)	22.4	69.4	8.2	24.4	23.4	15.9	23.9	25.1	25.0
LLaVA-1.5-13B (Liu et al. 2023)	19.3	76.7	4.1	20.1	19.7	11.6	21.2	21.9	20.8
LLaVA-1.5-7B (Liu et al. 2023)	16.1	78.5	5.4	17.1	16.6	8.9	19.2	19.0	17.1
LLaVa-NeXT-Video-34B (Liu et al. 2024)	11.2	83.8	4.9	11.8	11.5	7.6	11.5	10.3	14.5
LLaVa-NeXT-Video-7B (Liu et al. 2024)	9.3	52.9	37.8	14.9	11.4	7.4	15.1	14.5	8.7
LLaVA-OneVision-72B (Li et al. 2024a)	25.4	73.6	1.0	25.7	25.5	15.9	25.3	28.5	27.3
LLaVA-OneVision-7B (Li et al. 2024a)	18.9	76.6	4.5	19.8	19.3	12.1	26.3	21.3	18.4
DeepSeek-VL2 (Wu et al. 2024)	3.2	49.1	47.7	6.1	4.2	3.0	4.4	3.0	5.9
DeepSeek-VL2-Small (Wu et al. 2024)	5.9	52.1	42.1	10.1	7.4	3.9	10.1	9.5	6.2
DeepSeek-VL2-Tiny (Wu et al. 2024)	16.1	75.6	8.3	17.6	16.8	9.6	27.1	17.5	16.0
Kimi-VL (Team et al. 2025a)	18.3	**44.4**	37.3	29.1	22.4	14.9	19.8	25.2	24.5
Keye-VL (Team et al. 2025b)	25.4	53.9	20.7	32.0	28.3	15.3	23.1	37.2	26.6

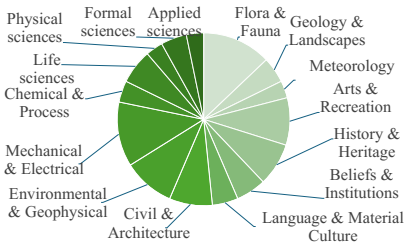
Table 2: Evaluation results (%) of open-source and proprietary multi-modal LLMs on Video SimpleQA . For metrics, CO, NA, IN, and CGA denote "Correct", "Not attempted", "Incorrect", and "Correct given attempted", respectively. For subtopics, ENG, NAT, SCI and SAC represent "Engineering", "Nature", "Science" and "Society and Culture".

targets a single fact that is independently verifiable; 2) **Referential dependency**: Each sub-QA builds upon the answer to the previous one, forming the step-by-step fact chaining.

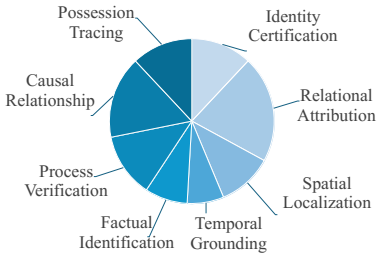
Quality Control

Difficulty Filtering. To ensure an appropriate level of assessment difficulty, we establish filtering rules to exclude questions that are easy to answer. In particular, questions correctly answered by all four state-of-the-art models, including GPT-4o, Claude Sonnet 4, Gemini 2.5 Pro (Comanici et al. 2025), and Qwen-VL-Max (Bai et al. 2023) are deemed insufficiently challenging and consequently excluded from our benchmark. This filtering strategy ensures our dataset maintains a sufficient level of difficulty for meaningful evaluations.

Human Cross-verification. To further enhance the dataset quality, a rigorous human validation process is implemented. Each question is independently evaluated by two annotators for compliance with our predefined criteria. Questions are discarded if either annotator deems them non-compliant. Meanwhile, annotators are required to verify answers against authoritative sources (such as Wikipedia). Finally, the final dataset undergoes security auditing to address potential security issues. All these stringent human verification processes ensure both the accuracy of our dataset and its adherence to established criteria.



(a)



(b)

Statistics of Video SimpleQA	Value
2/3/4-hop QA pairs	928 / 469 / 107
Total QA pairs	1504
Question Length (avg/max)	15.64 / 38
Answer Length (avg/max)	1.28 / 9
Unique Videos	1079
Video Length (Seconds, avg/max)	201 / 8763
Number of primary category	4
Number of secondary category	15
Number of tertiary category	84

(c)

Figure 5: (a) Video distribution at the secondary level; (b) Question type distribution; (c) Key statistics.

Dataset Statistics. The key statistics of Video SimpleQA are demonstrated in Table 5c. As shown, it consists of 1079 videos with 1504 expert-annotated QA pairs. The videos span 4 primary categories, 15 secondary categories and 84 tertiary categories. The average lengths of questions and answers are 15.64 and 1.28 words, respectively, aligning with our intended short-form design. The video distribution at the secondary level is demonstrated in Figure 5a. The question type distribution is visualized in Figure 5b.

Experiments

Experimental Setup

Evaluated Models. We benchmark comprehensive state-of-the-art LLMs, including **12 proprietary models**, including o4-mini, o3, GPT-4.5, GPT-4o, Claude Sonnet 4, Claude 3.7 Sonnet, Claude 3.5 Sonnet series, Gemini 2.5 series, and Qwen-VL series, and **21 open-source models**, including InternVL3 series, Qwen2.5-VL series, Qwen2-VL series, LLaVA-1.5 series, LLaVA-NeXT-Video series, LLaVA-OneVision series, DeepSeek-VL2 series, Kimi-VL, and Keye-VL. Following Video-MME (Fu et al. 2024), we maximize frame utilization of each model by inputting the maximum frames that fit within its context window.

Evaluation Metrics. Following SimpleQA (Wei et al. 2024), we set up five evaluation metrics: (1) **Correct**: The predicted answer comprehensively contains all key information from the reference answer while containing no contradictory elements. (2) **Incorrect**: The predicted answer contradicts the reference answer. The indirect or equivocal responses (*e.g.*, "possibly", "I think, although I'm not sure") are also considered incorrect. (3) **Not attempted**: The reference answer is not fully given in the predicted answer, and no statements in the answer contradict the gold target. (4) **Correct given attempted**: The ratio of correctly answered questions among attempted ones. (5) **F-score**: The harmonic mean values between *correct* and *correct given attempted* metrics. We follow the paradigm of *LLM-as-a-Judge* (Gu et al. 2024) and employ `o3` as the judge model.

Experimental Findings[4]

The evaluation results on Video SimpleQA are presented in Table 2, and key findings are summarized as follows:

[4]More experiments are available in the supplementary material.

Model	F-score		Inference Time	
	vanilla	*w/* RAG	vanilla	*w/* RAG
o3	66.3	**69.0**	**27.8**	54.7
GPT-4o	49.3	**61.3**	**30.1**	53.9
Gemini 2.5 Pro	62.6	**66.2**	**33.2**	60.1
Claude Sonnet 4	35.6	**58.8**	**29.9**	56.4
Qwen-VL-Max	39.9	**57.0**	**24.2**	61.2

Table 3: Comparisons between vanilla models and models with RAG by F-score (%) and the inference time (min).

Video SimpleQA is challenging: To assess human performance on Video SimpleQA , we sample 200 instances and recruit five participants to independently complete the tasks under two distinct conditions: with access to external resources (*e.g.*, Internet, textbooks) and without such access. These configurations correspond to the *human open-book* and *human closed-book* settings documented in Table 2.

Compared to the human open-book performance, both open-source and proprietary models demonstrate suboptimal performance. Specifically, the top-performing proprietary model, o3, achieves an F-score of 66.3%. Open-source models exhibit even poorer results, with the best-performing one, Qwen2.5-VL-72B (Wang et al. 2024a) attaining only 39.5% F-score. This demonstrates that LVLMs still exhibit limited capability in factuality adherence within video contexts, while also highlighting the necessity of establishing Video SimpleQA .

LVLMs are overconfident in what they generate: All models exhibit higher IN values (incorrect predictions) than NA values (non-attempted responses), indicating a prevalent tendency to generate answers despite insufficient factual knowledge. To further investigate this overconfidence phenomenon, we conduct *calibration* experiments (Guo et al. 2017) to examine whether language models "know what they know", *i.e.*, whether the model's assessed confidence scores align with the actual likelihood of its responses being correct. Specifically, we instruct LVLMs to self-assess confidence scores (0-100) for their predictions. Responses are grouped into confidence intervals (10-point bins), and we calculate *interval accuracy* (correct predictions per bin). As shown in Figure 7, except for o3, which demonstrates superior calibration, all other models *mostly* fall below the perfect calibration line, indicating systematic overconfidence.

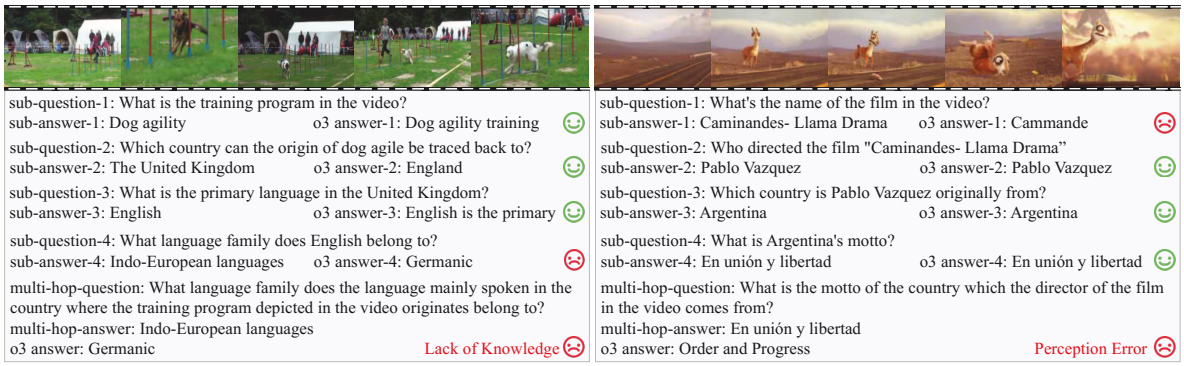


Figure 6: Visualizations of per-hop evaluation results of o3.

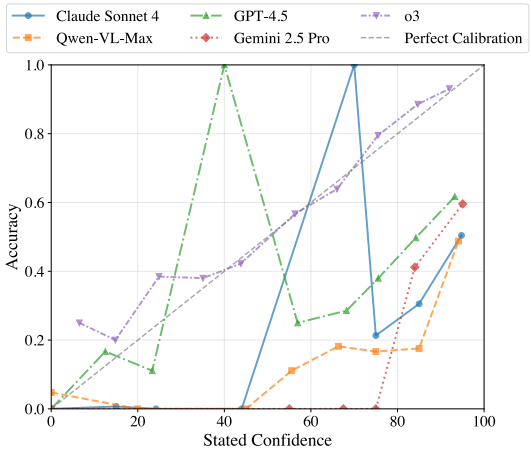


Figure 7: Calibration curves based on the self-stated confidence scores and interval-level accuracy.

RAG yields significant gains at the cost of inference efficiency: We explore RAG to facilitate Video SimpleQA benchmark comprehension in a three-step approach: 1) Prompting GPT-4o with video and questions to extract key textual entities; 2) Applying LlamaIndex with Google and Wikipedia as sources to retrieve relevant information based on these extracted key entities; 3) Augmenting the input question with the retrieved information.

As shown in Table 3, RAG achieves consistent and significant F-score improvements over vanilla models. For instance, when integrated with Claude Sonnet 4, RAG delivers an absolute improvement of 23.2% (35.6% *vs*. 58.8%). However, this performance gain comes with substantial computational overhead. Table 3 also quantifies the total inference time, demonstrating that RAG significantly impairs inference efficiency. Our findings highlight the critical trade-off between performance gains and computational practicality.

Per-hop factual evaluation: In addition to multi-hop factual QA, our Video SimpleQA benchmark also incorporates decomposed per-fact sub-QAs to facilitate fine-grained evaluations. As shown in Table 4, we present F-scores for both the final multi-hop QA and the sub-QAs across all 4-hop questions in Video SimpleQA . Our analysis reveals: 1) **Multi-hop challenge**: The final multi-hop QAs achieve substantially lower F-scores than most single-hop sub-QAs, underscoring the difficulty of multi-hop fact grounding; 2) **First-hop bottleneck**: The F-score for the first hop is markedly lower than those of later hops, likely due to its reliance on accurate object or event recognition, which poses a key challenge. In contrast, LVLMs perform better on subsequent hops (Q2–Q4) given clearer contextual grounding.

Model	QA1	QA2	QA3	QA4	Multi-hop
o3	**74.9**	**89.7**	**95.3**	**88.8**	**78.5**
GPT-4o	59.1	85.4	92.5	80.9	47.0
Claude Sonnet 4	43.5	85.7	80.8	61.6	47.7
Gemini 2.5 Pro	65.1	78.7	60.4	28.3	69.8
Qwen-VL-Max	37.9	69.2	83.6	72.3	40.7

Table 4: Per-hop factual evaluations for 4-hop questions in terms of F-score (%). Q1-Q4 denote the decomposed per-hop questions. Refer to Figure 6 for an illustrative case.

In Figure 6, we visualize the per-hop evaluation results of the o3 model. This allows us to clearly identify which specific piece of factual knowledge the model lacks. For instance, in the left case, o3 lacks knowledge about the language family of English, while in the right case, it fails to recognize the film Caminandes: Llama Drama.

Conclusions

We present `Video SimpleQA`, the first benchmark explicitly designed for evaluating factual grounding in video contexts. Distinct from prior works, our framework introduces the following diagnostic dimensions: knowledge integration, multi-hop fact-seeking questioning, short-form definitive answering, and temporal grounded demands. Through the extensive evaluation of 33 state-of-the-art LVLMs, we reveal notable deficiencies in factual adherence, uncover prevalent model overconfidence, trade-offs associated with RAG, and the critical performance bottleneck.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Akhtar, M.; Schlichtkrull, M.; Guo, Z.; Cocarascu, O.; Simperl, E.; and Vlachos, A. 2023. Multimodal Automated Fact-Checking: A Survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 5430–5448.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Cai, M.; Tan, R.; Zhang, J.; Zou, B.; Zhang, K.; Yao, F.; Zhu, F.; Gu, J.; Zhong, Y.; Shang, Y.; et al. 2024. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*.

Chandrasegaran, K.; Gupta, A.; Hadzic, L. M.; Kota, T.; He, J.; Eyzaguirre, C.; Durante, Z.; Li, M.; Wu, J.; and Fei-Fei, L. 2024. HourVideo: 1-Hour Video-Language Understanding. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Chen, X.; Lin, Y.; Zhang, Y.; and Huang, W. 2024. Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering. In *European Conference on Computer Vision*, 179–195. Springer.

Cheng, X.; Zhang, W.; Zhang, S.; Yang, J.; Guan, X.; Wu, X.; Li, X.; Zhang, G.; Liu, J.; Mai, Y.; et al. 2025. SimpleVQA: Multimodal Factuality Evaluation for Multimodal Large Language Models. *arXiv preprint arXiv:2502.13059*.

Chern, I.; Chern, S.; Chen, S.; Yuan, W.; Feng, K.; Zhou, C.; He, J.; Neubig, G.; Liu, P.; et al. 2023. FacTool: Factuality Detection in Generative AI–A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios. *arXiv preprint arXiv:2307.13528*.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Fang, X.; Mao, K.; Duan, H.; Zhao, X.; Li, Y.; Lin, D.; and Chen, K. 2024. MMBench-Video: A Long-Form Multi-Shot Benchmark for Holistic Video Understanding. *arXiv preprint arXiv:2406.14515*.

Fu, C.; Dai, Y.; Luo, Y.; Li, L.; Ren, S.; Zhang, R.; Wang, Z.; Zhou, C.; Shen, Y.; Zhang, M.; et al. 2024. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. *arXiv preprint arXiv:2405.21075*.

Garcia, N.; Otani, M.; Chu, C.; and Nakashima, Y. 2020. KnowIT VQA: Answering knowledge-based questions about videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 10826–10834.

Gou, Z.; Shao, Z.; Gong, Y.; Shen, Y.; Yang, Y.; Duan, N.; and Chen, W. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*.

Gu, J.; Jiang, X.; Shi, Z.; Tan, H.; Zhai, X.; Xu, C.; Li, W.; Shen, Y.; Ma, S.; Liu, H.; et al. 2024. A Survey on LLM-as-a-Judge. *arXiv preprint arXiv:2411.15594*.

Gu, J.; Wang, Y.; Bu, P.; Wang, C.; Wang, Z.; Song, T.; Wei, D.; Yuan, J.; Zhao, Y.; He, Y.; et al. 2025. "See the World, Discover Knowledge": A Chinese Factuality Evaluation for Large Vision Language Models. *arXiv preprint arXiv:2502.11718*.

Guan, T.; Liu, F.; Wu, X.; Xian, R.; Li, Z.; Liu, X.; Wang, X.; Chen, L.; Huang, F.; Yacoob, Y.; et al. 2024. Hallusion-Bench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14375–14385.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International conference on machine learning*, 1321–1330. PMLR.

He, X.; Feng, W.; Zheng, K.; Lu, Y.; Zhu, W.; Li, J.; Fan, Y.; Wang, J.; Li, L.; Yang, Z.; et al. 2024a. MMWorld: Towards Multi-discipline Multi-faceted World Model Evaluation in Videos. *arXiv preprint arXiv:2406.08407*.

He, Y.; Li, S.; Liu, J.; Tan, Y.; Wang, W.; Huang, H.; Bu, X.; Guo, H.; Hu, C.; Zheng, B.; et al. 2024b. Chinese simpleqa: A chinese factuality evaluation for large language models. *arXiv preprint arXiv:2411.07140*.

Hu, K.; Wu, P.; Pu, F.; Xiao, W.; Zhang, Y.; Yue, X.; Li, B.; and Liu, Z. 2025. Video-MMMU: Evaluating Knowledge Acquisition from Multi-Discipline Professional Videos. *arXiv preprint arXiv:2501.13826*.

Jain, A.; Kothyari, M.; Kumar, V.; Jyothi, P.; Ramakrishnan, G.; and Chakrabarti, S. 2021. Select, substitute, search: A new benchmark for knowledge-augmented visual question answering. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2491–2498.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Li, Y.; Liu, Z.; and Li, C. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; et al. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22195–22206.

Lin, S.; Hilton, J.; and Evans, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3214–3252.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Marino, K.; Rastegari, M.; Farhadi, A.; and Mottaghi, R. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 3195–3204.

Pan, S.; Luo, L.; Wang, Y.; Chen, C.; Wang, J.; and Wu, X. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*.

Reid, M.; Savinov, N.; Teplyashin, D.; Lepikhin, D.; Lillicrap, T.; Alayrac, J.-b.; Soricut, R.; Lazaridou, A.; Firat, O.; Schrittwieser, J.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Song, E.; Chai, W.; Wang, G.; Zhang, Y.; Zhou, H.; Wu, F.; Chi, H.; Guo, X.; Ye, T.; Zhang, Y.; et al. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18221–18232.

Team, K.; Du, A.; Yin, B.; Xing, B.; Qu, B.; Wang, B.; Chen, C.; Zhang, C.; Du, C.; Wei, C.; et al. 2025a. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

Team, K. K.; Yang, B.; Wen, B.; Liu, C.; Chu, C.; Song, C.; Rao, C.; Yi, C.; Li, D.; Zang, D.; et al. 2025b. Kwai Keye-VL Technical Report. *arXiv preprint arXiv:2507.01949*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, C.; Liu, X.; Yue, Y.; Tang, X.; Zhang, T.; Jiayang, C.; Yao, Y.; Gao, W.; Hu, X.; Qi, Z.; et al. 2023. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, P.; Wu, Q.; Shen, C.; Dick, A.; and Van Den Hengel, A. 2017. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10): 2413–2427.

Wang, P.; Wu, Q.; Shen, C.; Hengel, A. v. d.; and Dick, A. 2015. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570*.

Wang, Y.; Wang, M.; Manzoor, M. A.; Liu, F.; Georgiev, G.; Das, R.; and Nakov, P. 2024b. Factuality of large language models: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 19519–19529.

Wang, Y.; Wang, Y.; Zhao, D.; Xie, C.; and Zheng, Z. 2024c. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*.

Wei, J.; Karina, N.; Chung, H. W.; Jiao, Y. J.; Papay, S.; Glaese, A.; Schulman, J.; and Fedus, W. 2024. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*.

Wu, Z.; Chen, X.; Pan, Z.; Liu, X.; Liu, W.; Dai, D.; Gao, H.; Ma, Y.; Wu, C.; Wang, B.; et al. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Yang, J.; Yang, S.; Gupta, A. W.; Han, R.; Fei-Fei, L.; and Xie, S. 2024. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv preprint arXiv:2412.14171*.

Yu, W.; Iter, D.; Wang, S.; Xu, Y.; Ju, M.; Sanyal, S.; Zhu, C.; Zeng, M.; and Jiang, M. 2022. Generate rather than retrieve: Large language models are strong context generators. *arXiv preprint arXiv:2209.10063*.

Zellers, R.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6720–6731.

Zhang, J.; Jiao, Y.; Chen, S.; Chen, J.; and Jiang, Y.-G. 2024a. Eventhallusion: Diagnosing event hallucinations in video llms. *arXiv preprint arXiv:2409.16597*.

Zhang, Y.; Zhang, K.; Li, B.; Pu, F.; Setiadharma, C. A.; Yang, J.; and Liu, Z. 2024b. WorldQA: Multimodal World Knowledge in Videos through Long-Chain Reasoning. *arXiv preprint arXiv:2405.03272*.

Zhao, Y.; Xie, L.; Zhang, H.; Gan, G.; Long, Y.; Hu, Z.; Hu, T.; Chen, W.; Li, C.; Song, J.; et al. 2025. MMVU: Measuring Expert-Level Multi-Discipline Video Understanding. *arXiv preprint arXiv:2501.12380*.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.