

Seeking and Updating with Live Visual Knowledge

Mingyang Fu^{1,*}, Yuyang Peng^{1,*}, Dongping Chen^{1,2,‡}, Zetong Zhou¹, Benlin Liu²,
Yao Wan^{1,†}, Zhou Zhao³, Philip S. Yu⁴, Ranjay Krishna²

¹Huazhong University of Science and Technology ²University of Washington

³Zhejiang University ⁴University of Illinois Chicago

* Equal Contribution. † Corresponding Author. ‡ Project Leader (dongpingchen@umd.edu).



Figure 1: LIVEVQA: a new dataset for visual knowledge seeking and updating, comprising 12 different categories, 28,488 visual entity and 107,138 cross-modality multi-hop reasoning questions. Each instance contains a representative image as visual knowledge from April, 2024 to May, 2025.

Abstract

The visual world around us constantly evolves, from real-time news and social media trends to global infrastructure changes visible through satellite imagery and augmented reality enhancements. However, *Multimodal Large Language Models* (MLLMs), which automate many tasks, struggle to stay current, limited by the

cutoff dates in their fixed training datasets. To quantify this stagnation, we introduce LIVEVQA, the first-of-its-kind dataset featuring 107,143 samples and 12 categories data specifically designed to support research in both seeking and updating with live visual knowledge. Drawing from recent news articles, video platforms, and academic publications in April 2024-May 2025, LIVEVQA enables evaluation of how models handle latest visual information beyond their knowledge boundaries and how current methods help to update them. Our comprehensive benchmarking of 17 *state-of-the-art* MLLMs reveals significant performance gaps on content beyond knowledge cutoff, and tool-use or agentic visual seeking framework drastically gain an average of 327% improvement. Furthermore, we explore parameter-efficient fine-tuning (PEFT) methods to update MLLMs with new visual knowledge. We dive deeply to the critical balance between adapter capacity and model capability when updating MLLMs with new visual knowledge. All the experimental dataset and source code are publicly available at: <https://livevqa.github.io>.

1 Introduction

The world around us is constantly changing. We hear about it from real-time news coverage on our television sets; we identify new fashion trends by looking at new posts on social media; at a global level, these trends are evident even from satellite imagery of our ever evolving urban infrastructure; emerging visual technologies like augmented reality interfaces consistently improve visual fidelity and experiences [Decker et al., 2023, Li et al., 2024a]. People are adept at ingesting new knowledge as they encounter them.

Naturally, as we automate aspects of our tasks, we would want our assistants to remain as up-to-date as we are. Yet, the *Multimodal Large Language Models* (MLLMs) that power such assistants quickly become stale [OpenAI, 2025, Citron, 2025, Anthropic, 2025]. Their knowledge remains limited to their training data; they struggle with visual information that post-dates their training cutoff points. For example, traditional visual question answering (VQA) datasets [Antol et al., 2015, Hudson and Manning, 2019, Nguyen et al., 2025] typically rely on fixed knowledge boundaries, creating a fundamental disconnect between model capabilities and the ever-evolving visual world. This gap becomes increasingly apparent as models grow more sophisticated [Ishfaq et al., 2023], highlighting the need for approaches that can continuously incorporate and reason about real-world up-to-date visual knowledge.

To address these challenges, this paper introduces LIVEVQA, a large-scale VQA dataset specifically designed to benchmark MLLMs in seeking and updating live visual knowledge. LIVEVQA distinguishes itself as a first-of-its-kind resource containing fresh visual content sourced from prominent international news articles, YouTube videos, and recent academic papers [Shabtay et al., 2024, Li et al., 2024b], spanning from April 2024 to early May 2025. Constructed through a multi-stage LLM/MLLM-in-the-loop pipeline, LIVEVQA incorporates rigorous filtering criteria and human validation to ensure data quality. It is designed to challenge models with the recognition of novel visual entities and the execution of cross-modal, multi-hop reasoning. Each instance in LIVEVQA consists of a visually distinctive image representing a specific event, paired with two levels of questions: Level 1 targets visual entity recognition, while Level 2 probes deeper visual knowledge reasoning. Based on LIVEVQA, we investigate the following two research questions.

We first investigate *how well do current MLLMs seek out visual knowledge*. Comprehensive experiments over 17 MLLMs (*e.g.*, GPT-o3 and Gemini-2.5-Pro) reveal that all current MLLMs struggle significantly to identify latest visual knowledge, and text-based online searching does not contribute improve performance. Experimental results reveal that incorporating multimodal search tools [Jiang et al., 2024] substantially improves performance, highlighting the importance of retrieval-augmented approaches for handling dynamic visual information.

We further investigate *whether we can update MLLMs with new visual knowledge* via exploring parameter-efficient knowledge updating approaches (*e.g.*, LoRA and DoRA). Experiments shows that visual information can be efficiently updated through fine-tuning within only one epoch. Although it harms visual perceptions capability, fine-tuning with short-term visual factuality-oriented QA format even enhance knowledge intensive capability, with a 4.2-point improvement.

To facilitate future study, all materials have been made publicly available online.

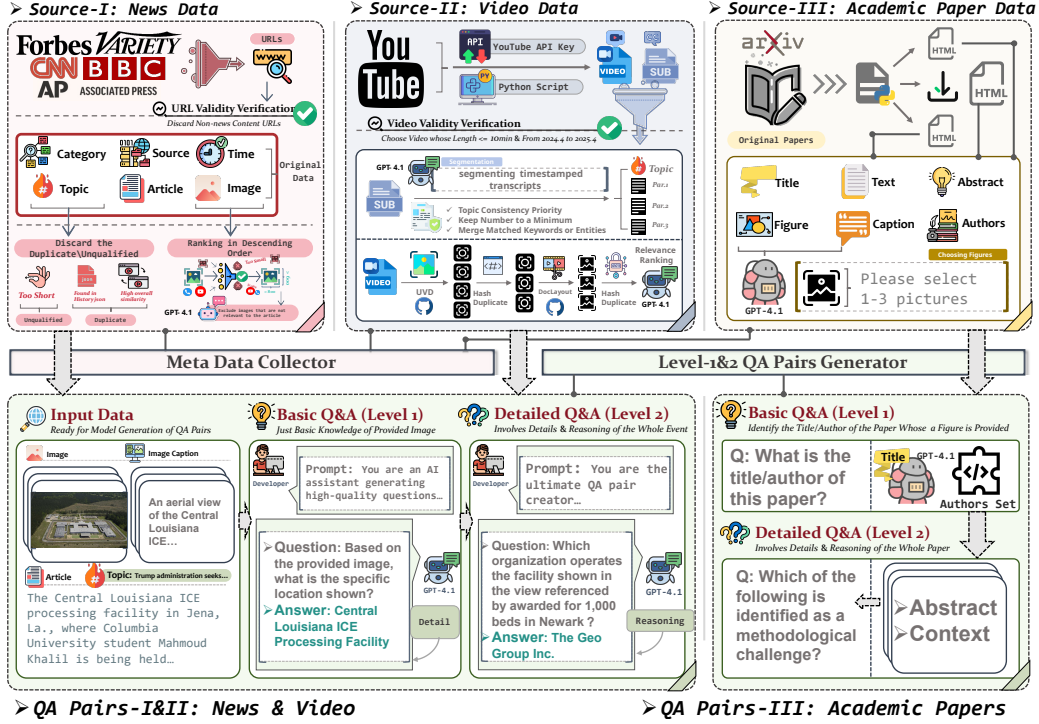


Figure 2: **Pipeline of our proposed data engine to build LIVEVQA.** It consists of two modules: raw data collector and Q&A pairs generator. It collects illustrated visual data from multiple domains (*i.e.*, news articles, YouTube videos, and arXiv papers), conducts multi-level data filtering, and generates basic and detailed Q&A pairs. We advice a **zoom-in** for more details.

2 LIVEVQA: The dataset

We present LIVEVQA, a first-of-its-kind automatically collected VQA dataset containing 28,488 unique images and 107,143 questions, for testing the MLLMs’ capabilities of visual knowledge seeking and updating. Following our primary principle of using fresh, non-contaminated visual data (content absent from model training sets), we develop a specialized data engine to collect content exclusively from after the model’s knowledge cutoff. The dataset features recent visual content spanning April 2024 to early May 2025, sourced from prominent international news articles, YouTube videos, and academic papers. Our collection process implements a multi-stage LLM/MLLM-in-the-loop pipeline with strict filtering criteria, as illustrated in Figure 2. **Notice that all LLM/MLLM-assisted processes utilize GPT-4.1 [OpenAI, 2025] and undergo human validation with a greater than 97% pass rate, as detailed in Section C.4.**

LIVEVQA is structured as: (1) A visually distinctive image depicting a specific event. (2) Level 1 questions focusing on basic visual entity recognition (*e.g.*, locations, persons, time, events). (3) Level 2 questions requiring more detailed information about the image, necessitating multi-hop cross-modality inferential reasoning. (4) Both question are available in open-ended and multiple-choice formats, accompanied by high-quality synthetic reasoning trajectories that uncover detailed information behind the image. See Figure 24 for a concrete example.

2.1 Raw data collection

News articles. News articles reflect the ever-evolving real-world visual knowledge, providing high-quality authentic data resource. Therefore, we collect news corpus aggregated from leading international news organizations, such as CNN, BBC, Forbes, Variety, and Associated Press News for their authentic. Specifically, the metadata collection pipeline is detailed as follows:

- ▷ **URL and headline filtering.** We collect URLs that point to specific reports from news index pages. For each URL, we extract the title, image, text, release timestamp, and source. To ensure

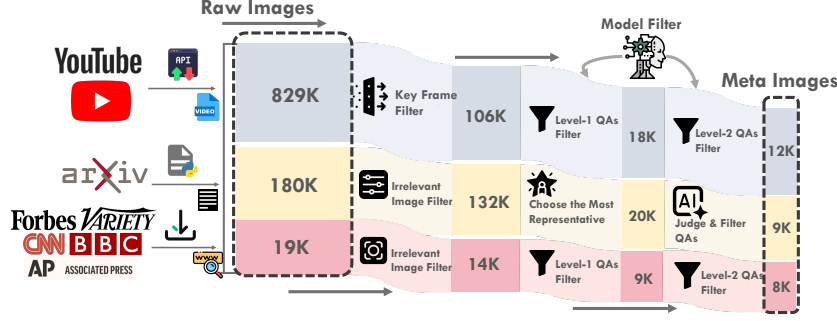


Figure 3: The filtering process for LIVEVQA shows how images and synthesized questions are removed throughout the data pipeline for news articles, YouTube videos, and arXiv papers.

data purity and relevance, we implement a strict URL screening mechanism, and URLs that do not meet the unique form of articles on each news website are excluded.

- ▷ **Image selection.** Too small images (*e.g.*, website icons) are invalid and must be discarded. For the remaining images, they are sorted in descending order according to pixel area and a maximum of four images that are most relevant to the news content are retained, and images with an area less than 50% of the largest image are further removed.
- ▷ **Image-event relevance enhancement.** Observing that some images (*e.g.*, a close-up of a pen) do not adequately represent events, we use GPT-4.1 to identify and exclude images that are weakly correlated with the news event, improving the semantic coherence of the data set and the purity of information. Finally, we calculate the SimHash value of the news title and text content for semantic similarity comparison, and those with a similarity score > 0.85 need to be discarded. After filtering out invalid images and topics, 27.6% samples are discarded.

Videos. With the rising popularity of video platform such as YouTube and TikTok, videos have become increasingly dominant channels for human to access real-world visual knowledge. They excel at capturing up-to-date visual knowledge and inherently convey dynamic, multimodal information about latest events. We choose YouTube as the source for our video knowledge base for their diversity, easy-to-crawl and CC-by-4.0 license. The data collection pipeline is detailed as follows:

- ▷ **YouTube data preprocessing.** We restrict the target videos to English-language content (to ensure linguistic uniformity), a maximum duration of 10 minutes (to control information volume), and the presence of English subtitles. We collect videos themselves, official or auto-generated subtitle, video titles, descriptions, and other information (such as publication dates, *etc.*).
- ▷ **Subtitle-based video split.** Given that a video may contain multiple independent news events, we need to split them into small video clip. Video subtitles stand out to be high-quality and authentic context. We use an LLM to refine the transcribed subtitle files by removing oral expressions and irregular breaks before using the cleaned content as the basis for video segmentation. To ensure segmentation quality, we conduct manual evaluation, achieving a passing rate of 98%.
- ▷ **Initial keyframe identification and preprocessing.** To handle video content efficiently, we implemented a three-step process to extract unique, high-quality frames. First, we used UVD¹ to identify key candidate frames from videos, eliminating redundancy. Second, we applied a Perceptual Hash algorithm to remove duplicates while preserving only the clearest version of each image. Third, we employed the DocLayout-YOLO² model to crop images, removing textual elements and irrelevant information (example in Figure 9). As a final filter, we removed near-duplicate images by excluding those with a perceptual hash Hamming distance below 25.
- ▷ **LLM-driven keyframes selection.** After two rounds of deduplication, a video clip still contains 5-20 keyframes (shown in Figure 4), which is too duplicate to represent an event. Therefore, we leverage an MLLM to identify top-K images by their relevance to the topic and refined subtitles, visual informativeness, and representativeness of the remaining keyframes. Finally, visual information within each video segment is highly condensed and thematically relevant.

Academic papers. Academic papers, as a form of disseminating cutting-edge knowledge, serve as a stable source for transmitting the latest visual knowledge [Shabtay et al., 2024, Li et al., 2024b].

¹<https://github.com/zcczhang/USD>

²<https://github.com/opendatalab/DocLayout-YOLO>

Table 1: Overall statistics of LIVEVQA. For Level 1 and Level 2 questions, we provide high-quality synthetic answers that fully reveal the image context, with *Avg. Len.* indicating the answer length.

Category	Images	#Question	Level 1	Level 2	Avg. Len.	Purpose
News Article	7,579	38,809	7,579	31,230	749	-
YouTube Videos	11,948	43,168	11,948	31,220	311	-
Academic Paper	8,961	25,166	9456	16,205	597	-
Avg. per Sample	1	3.86	1	2.86	517	-
Test Split	1,500	3,000	1,500	1,500	544	Exp. 1
Training Split	26,988	104,143	26,988	77,150	496	Exp. 2

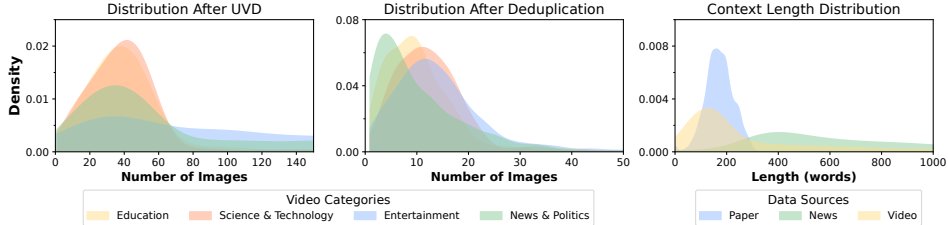


Figure 4: **(Left)** Image size distribution in YouTube image filtering pipeline. **(Right)** Textual context length distribution for each question.

Therefore, we collect papers from arXiv and balanced them by category, selecting papers from Physics, Computer Science, and Mathematics.³

- ▷ **Article and image preprocessing.** We deliberately collect web pages of arXiv papers across various domains. From these pages, we extract the title, abstract, authors, section content, images, and image captions of the papers by beautifulsoup. During image processing, we filter out irrelevant images that do not conform to specific formats (*e.g.*, icons, SVGs).
- ▷ **Key image selection.** Typically, academic papers contain five or more images, which can be excessive and visually overwhelming when representing a paper’s content. Our selection criteria prioritize images that exhibit significant diversity across different papers, focusing specifically on architectural diagrams and illustrations of key findings. We deliberately avoid common images found in most papers, such as standard statistical visualizations or text-heavy figures. Finally, we select 1-2 distinctive images for each paper that serve as visual signatures.

2.2 Visual question answering generation

We construct our visual knowledge seeking and updating dataset with two distinct question levels. Level 1 includes straightforward visual entity recognition questions based on filtered images and metadata. Level 2 presents complex, multi-hop cross-modal reasoning questions that require models to use full image context to navigate related textual information. All questions are rigorously filtered to systematically evaluate models’ abilities in both basic visual fact retrieval and cross-modal synthesis.

- ▷ **Level-1 questions** focus on substantive elements such as people, objects, or locations, while avoiding queries without visual knowledge that solely reliant on visual content within the image, such as color or shape. We employ GPT-4.1 to filter unqualified QAs. This filter deletes overly brief news items and discards questions whose answers are mere simple labels (*e.g.*, “city”), news platform names (*e.g.*, CNN); meanwhile, it retains questions whose answers are specific personal name, locations, events, products and organizations. After filtering, 26.7% samples are discarded. Finally, we require model to synthesis a single chain-of-thought [Wei et al., 2022] describing the historical or social event depicted, and finally answer the question with reasons.
- ▷ **Level-2 questions** require deeper contextual cross-modality reasoning. We ask GPT-4.1 to generate multi-hop visual reasoning questions that naturally reference image details and the textual metadata context, spanning exactly seven types (*i.e.*, location, person, organization, time, event, count, reason). To ensure the verifiability of the answers, we employ GPT-4.1 providing it with questions, images, and their corresponding context, and we retain only those Level 2 questions that the model correctly

³Due to the complex anti-scraping mechanisms of bioRxiv and medRxiv, we fail to achieve a completely synthetic data collection. Thus, after attempting to collect these papers, we abandon the effort.

Table 2: Accuracy (%) of visual factuality seeking benchmark in open-ended format across different models across difficulty levels and data sources.

Model	Cutoff	Level 1				Level 2			
		News	Video	Arxiv	Avg.	News	Video	Arxiv	Avg.
w.o. Search									
GPT-4.1	Jun. 2024	27.0	22.0	0.4	16.5	5.2	7.2	0.2	3.0
GPT-4.1-mini	Jun. 2024	24.6	19.6	0.2	14.8	4.0	7.8	0.4	4.0
GPT-4.1-nano	Jun. 2024	13.0	13.0	0.0	8.6	2.2	6.0	0.4	2.9
Gemini-2.5-Flash-Lite	Jan. 2025	36.4	23.4	0.4	20.1	4.0	4.8	4.2	3.3
Gemini-2.5-Flash	Jan. 2025	25.8	18.4	0.8	15.0	4.6	4.4	4.0	4.3
Gemini-2.5-Pro	Jan. 2025	28.0	17.4	0.6	15.3	4.4	2.4	1.2	2.7
Gemma-3-27B-It	Aug. 2024	21.0	16.4	1.0	12.8	3.8	4.6	6.2	4.9
Gemma-3n-E4B-IT	Jun. 2024	13.6	11.8	0.0	8.5	1.2	2.2	2.0	1.8
Claude-3.7-Sonnet	Oct. 2024	26.2	16.4	0.6	14.3	2.2	4.4	4.4	3.7
Claude-4-Sonnet	Nov. 2024	25.8	19.6	1.6	15.7	3.6	6.2	10.8	6.9
Qwen-2.5-VL-7B-Instruct	Unknown ⁴	20.2	13.4	0.2	11.3	3.8	5.4	2.0	3.7
Qwen-2.5-VL-32B-Instruct	Unknown	25.2	16.4	0.4	14.0	4.2	5.6	1.2	3.7
Qwen-2.5-VL-72B-Instruct	Unknown	12.4	9.4	0.0	7.3	1.4	3.6	3.6	2.9
Llama-4-Scout	Aug. 2024	20.6	16.4	0.0	12.1	4.0	5.0	2.8	3.9
Llama-4-Maverick	Aug. 2024	20.2	19.0	0.6	13.3	5.8	6.0	5.2	5.7
GLM-4.1V-9B-Thinking	Unknown	22.8	19.8	1.6	19.5	3.8	8.6	7.0	6.5
w. Text Search									
GPT-4.1	Jun. 2024	25.0	21.4	0.6	15.6	3.6	5.6	3.8	4.3
Gemini-2.5-Pro	Jan. 2025	17.6	9.2	0.2	9.0	2.0	1.6	1.0	1.5
Claude-3.7-Sonnet	Oct. 2024	24.6	16.6	0.0	13.7	2.0	3.6	4.8	3.5
w. Native Image Search									
GPT-o3	Jun. 2024	33.6	33.6	<u>2.6</u>	<u>23.3</u>	<u>14.6</u>	<u>14.9</u>	<u>17.8</u>	<u>15.8</u>
Grok-4	Dec. 2024	32.2	32.6	1.1	21.9	9.4	12.8	15.8	12.6
w. MM-Search [Jiang et al., 2024]									
GPT-4.1	Jun. 2024	42.0	<u>33.0</u>	36.1	33.4	27.2	15.2	48.8	30.4

answers. As a result, 13.0% samples are discarded after filtering. Similarly, the model offers a paragraph analyzing the event’s details in a step-by-step manner.

2.3 Dataset statistics and split

Finally, we collect 107,143 carefully curated VQA dataset with 26,988 unique images as shown in Table 1. As illustrated in Figure 1, the dataset covers a diverse range of topics with representative examples, showcasing its breadth and richness in both content and modality. For evaluation, we manually select a test split featuring 1,500 unique images and 3,000 questions for evaluating model’s visual seeking capability. Our requirement is that each queries (1) can be searched by human using purely web browser (with reverse visual search function) within 10 minutes; (2) Following SimpleQA [Wei et al., 2024], we set up a small pool of *state-of-the-art* models (*i.e.*, GPT-4.1, Gemini-2.5-Pro, Claude-3.7-Sonnet, and Qwen2.5-VL-72B), and filter out the questions that half of the model can answer correctly.

3 How well do current MLLMs seek out visual knowledge?

Keeping pace with the latest visual knowledge is crucial for helpful assistants immersively involved in human life and solving users’ problems. Here we investigate *how well do current MLLMs seek out visual knowledge* and evaluate their effectiveness in leveraging external tools to access this knowledge. Moreover, as the test set is authentic and validated by humans, it can serve as visual factuality benchmarks to measure the scientific phenomenon known as calibration [Wei et al., 2024], *i.e.*, whether the models “*know what they know*” about the visual content.

3.1 Experiment setups

Models. We conduct a series of zero-shot testing for a diverse range of *state-of-the-art* MLLMs without online searching capability to validate our dataset, including Gemini-2.5-Flash/Pro/Lite [Google,

Table 3: Comparison performance between different models and information seeking methodology on detailed categories in News subset.

Model	Level 1						Level 2							
	Loc.	Per.	Org.	Eve.	Obj.	Avg.	Loc.	Per.	Org.	Time	Cou.	Rea.	Eve.	Avg.
w.o. Search														
GPT-4.1	38.81	6.28	50.72	15.19	35.89	27.03	1.75	0.00	11.68	3.82	7.84	1.63	0.00	5.05
GPT-4.1-mini	33.33	10.91	45.59	11.86	31.73	24.60	3.57	0.00	8.82	0.00	10.24	0.00	0.00	4.00
GPT-4.1-Nano	16.16	3.64	30.88	3.39	19.23	13.00	0.00	0.00	4.41	1.54	3.94	0.83	0.00	2.20
Gemini-2.5-Flash-Lite	36.36	38.18	47.06	19.49	38.46	36.40	3.57	0.00	5.88	2.31	4.72	5.00	0.00	4.00
Gemini-2.5-Flash	26.26	37.27	35.29	7.63	27.88	25.80	3.57	0.00	1.47	3.85	8.66	4.17	0.00	4.60
Gemini-2.5-Pro	23.23	46.36	35.29	10.17	28.85	28.00	3.57	0.00	5.88	3.08	3.94	6.67	0.00	4.40
Gemma-3-27B-IT	24.24	15.45	38.24	8.47	25.96	21.00	3.57	0.00	8.82	1.54	7.87	0.00	0.00	3.80
Gemma-3n-E4B-IT	15.15	10.00	23.53	9.32	14.42	13.60	0.00	0.00	2.94	3.05	0.00	0.00	0.00	1.20
Claude-3.7-Sonnet	38.38	10.00	38.24	14.41	37.50	26.20	0.00	0.00	4.41	2.31	1.57	2.50	0.00	2.20
Claude-4-Sonnet	30.30	8.18	38.24	23.73	34.61	25.80	0.00	0.00	4.41	4.58	6.23	0.83	0.00	3.60
Qwen-2.5-VL-7B	23.23	18.18	30.88	12.71	21.15	20.20	0.00	0.00	4.41	1.54	7.09	4.17	0.00	3.80
Qwen-2.5-VL-32B	33.33	18.18	30.88	18.64	28.85	25.20	0.00	0.00	7.35	2.31	6.30	4.17	0.00	4.20
Qwen-2.5-VL-72B	15.15	6.36	25.00	8.47	12.50	12.40	0.00	0.00	4.41	0.77	1.57	0.83	0.00	1.40
Llama-4-Scout	26.26	13.64	35.29	8.47	26.92	20.60	3.57	0.00	4.41	3.08	9.45	0.00	0.00	4.00
Llama-4-Maverick	20.20	19.09	36.76	5.93	26.92	20.20	0.00	0.00	10.29	2.31	13.39	1.67	0.00	5.80
GLM-4.1V-9B-Thinking	18.18	14.55	36.76	11.02	32.00	22.80	0.00	0.00	2.94	6.10	7.09	0.0	0.00	3.80
w. Text Search														
GPT-4.1	36.36	2.73	48.53	13.56	34.62	25.00	3.57	5.88	5.88	3.85	4.72	0.83	0.00	3.60
Gemini-2.5-Pro	18.18	12.73	29.41	10.17	23.08	17.60	0.00	0.00	4.41	1.54	2.36	1.67	0.00	2.00
Claude-3.7-Sonnet	29.29	6.36	33.82	18.64	40.38	24.60	3.57	5.88	1.47	1.54	3.15	0.83	0.00	2.00
w. Native Image Search														
GPT-o3	47.47	4.55	57.35	23.73	47.12	33.60	17.86	0.00	20.59	7.69	17.32	17.50	10.00	14.60
Grok-4	32.32	34.54	44.11	16.50	38.46	32.20	17.86	5.88	17.65	6.87	11.81	3.33	10.00	9.40
w. MM-Search [Jiang et al., 2024]														
GPT-4.1	42.86	35.78	55.88	33.05	50.00	42.00	15.50	46.43	23.53	30.88	42.52	20.00	0.00	27.20

2025], GPT-4.1 family [OpenAI, 2025]. We also conduct experiments on open-source MLLMs, *e.g.*, Qwen2.5-VL-3/7/32/72B [Yang et al., 2024], Gemma-3-27B-it, Gemma-3n-E4B-IT [Team et al., 2025], and Llama-4-Scout/Maverick [Meta, 2025]. Native image search models like GPT-o3 [OpenAI, 2025] and Grok-4 [xAI, 2025] are also included as a strong baseline. Searching within text using openrouter⁵ web search API and GUI-based image search framework MM-Search [Jiang et al., 2024] are also included in our experiments.

Evaluation. We use prompts from the Humanity Last Exam [Phan et al., 2025] to gather model responses and confidence scores for evaluating performance and calibration. To avoid egocentric bias [Panickssery et al., 2024], we adopt grading metrics evaluated by Gemini-2.0-Flash, following SimpleQA [Wei et al., 2024], to evaluate *Correct*, *Not Attempted*, and *Incorrect* responses, along with the resulting F-score. All results are reported as averages over three independent evaluations.

3.2 Results and analysis

Larger-scale base models demonstrate improved performance in visual factuality recognition, and proprietary models retain a clear advantage. For models sharing the same knowledge cutoff (*e.g.*, the GPT-4.1 and Llama-4 families), we observe that increased model size consistently correlates with improved accuracy across all levels of question difficulty. For example, GPT-4.1-nano achieves an average accuracy of only 8.6% on Level 2 questions in the News category, while the larger GPT-4.1-mini and GPT-4.1 models attain 14.8% and 16.5%, respectively.

Furthermore, open-source models still lag behind proprietary counterparts in overall performance. Within the Qwen-2.5-VL family, the smaller 7B model surprisingly outperforms the larger 72B variant. One plausible explanation is that these models are trained on different corpora; notably, the Qwen-2.5-VL-32B model is released two months after the others, possibly benefiting from a more refined or expanded dataset, which may account for its highest score within the family. Notably, Llama-4-Maverick achieves the best performance across nearly all dimensions on Level 2 questions

⁵<https://openrouter.ai/>

Table 4: Detailed breakdown on non-search model’s failure categories in **News** subset. See definition of F-score in Appendix D. “*Correct & given attempted*”: model answer correctly, out of only questions that were attempted (*i.e.*, questions answered correct and incorrectly).

Model	Correct	Not attempted	Incorrect	Correct & given attempted	F-score
GPT-4.1	16.0	52.4	31.6	33.6	21.7
GPT-4.1-mini	14.3	44.0	41.7	25.5	18.3
GPT-4.1-nano	7.6	58.0	34.4	18.1	10.7
Gemini-2.5-Flash-Lite	20.2	58.1	<u>21.7</u>	48.2	28.5
Gemini-2.5-Flash	15.2	59.3	25.5	37.3	21.6
Gemini-2.5-Pro	<u>16.2</u>	58.4	25.4	38.9	<u>22.9</u>
Claude-3.7-Sonnet	14.2	64.2	21.6	<u>39.7</u>	20.9
Claude-4-Sonnet	14.7	62.5	22.8	39.2	21.4
Gemma-3-27B-IT	12.4	24.0	63.6	16.3	14.1
Gemma-3n-E4B-IT	7.4	71.2	22.4	24.8	11.3
Qwen-2.5-VL-7B	12.0	61.0	27.0	30.8	17.3
Qwen-2.5-VL-32B	14.7	40.9	44.4	24.9	18.5
Qwen-2.5-VL-72B	6.9	<u>36.2</u>	56.9	10.8	8.4
Llama-4-Scout	12.3	52.4	35.3	25.8	16.7
Llama-4-Maverick	12.6	55.8	31.6	28.5	17.5
GLM-4.1V-9B-Thinking	13.3	61.2	25.5	34.3	19.2

in the News subset, with an overall accuracy of 5.8%, potentially due to its extensive training corpus coverage.

A multimodal search engine matters. From Table 2, we can see that GPT-4.1’s average accuracy more than doubles—from 16.5% to 33.4%—when augmented with multimodal search tools. This improvement is particularly striking on challenging Level 2 questions, where accuracy rises to 27.2% and 15.2% on the News and Video subsets, respectively. Similarly, **GPT-o3**, which shares the same knowledge cutoff as GPT-4.1, shows a substantial gain on Level 2 questions, improving from 3.0% to 15.8% with tool use. These results underscore the considerable potential of reinforcement learning-based multimodal tool-use agents. Notably, **Grok-4**, which is trained partly on data sourced from X, exhibits high performance on Level 2 questions in News 9.4% and Video 12.8% subsets, suggesting a specific advantage in handling real-time, news-related visuals. Overall, the integration of multimodal search capabilities proves especially beneficial for questions that exceed the models’ internal knowledge. Figure 25 further illustrates this advantage by contrasting GPT-4.1’s [OpenAI, 2025] successful answers with tool use against its failures without it.

Measuring visual factuality calibration. Figure 5 demonstrates a positive correlation between stated confidence and accuracy across models, though with significant calibration issues. Claude-3.7-Sonnet exhibits better calibration at lower confidence levels, while Gemini-2.5-Pro performs better at higher confidence levels, yet all models struggle to properly align their confidence with actual accuracy. Consistent with prior research [Wei et al., 2024, Gao et al., 2024], GPT-4.1 shows superior calibration compared to its smaller variants (*i.e.*, GPT-4.1-mini and nano), supporting the finding that larger models tend to be better calibrated. Notably, the performance of all models falls significantly below the ideal $y = x$ line, indicating a consistent pattern of overconfidence in visual factuality assessments and underscoring substantial opportunities for improving MLLM calibration.

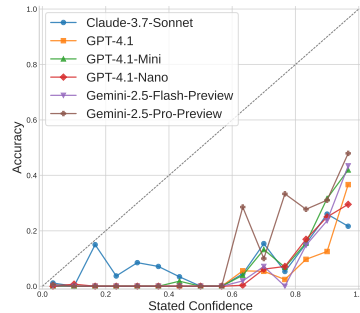


Figure 5: Calibration of MLLMs when encountering unknown visual knowledge based on their stated confidence.

4 Can we update MLLMs with new visual knowledge?

Recent benchmarks reveal that state-of-the-art MLLMs struggle with tasks requiring up-to-date visual knowledge, as the information learned during pretraining quickly becomes obsolete with the emergence of new entities and events. While retrieval-augmented generation (RAG) can incorporate external knowledge, it introduces latency and often fails to resolve semantically similar visual inputs. PEFT offers a more efficient solution for updating MLLMs with new visual concepts. Building

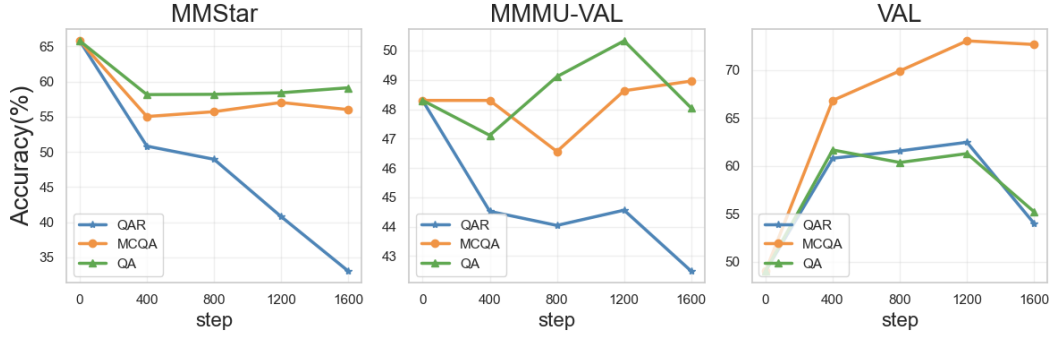


Figure 6: The results of visual knowledge updating with different data formats.

on prior work [Ravaut et al., 2024, Zeng et al., 2024, Chen et al., 2024a], we ask: *Can MLLMs be efficiently updated with new visual knowledge?* We investigate this by examining how updates affect both retention of prior visual understanding and general multimodal reasoning.

4.1 Experiment setups

Models and PEFT settings. We select Qwen2.5-VL-3B/7B-Instruct for their superior performance in vision-language understanding capability. We implement two popular PEFT methods LoRA [Hu et al., 2022] and DoRA [Liu et al., 2024a] with diverse rank ranging from [16, 64] for ablation study. See Appendix D for detailed experiment settings.

Data size and format ablation. We construct our instruction tuning dataset using three different formats for ablation study: (1) QA: Question + Ground Truth. (2) MCQA: Question + Multiple-choices + Correct letter + Ground Truth. (3) QAR: Question + Ground Truth + Reasoning process. See Figure 24 for an example.

Evaluation. Following previous knowledge updating research [Wang et al., 2025], we synthesize a 1,500-sample validation set in MCQA format from the training set, with 500 samples from each category and each question rephrased by GPT-4.1. To evaluate the model’s general performance, we select MMMU [Yue et al., 2024] and MMStar [Chen et al., 2024b] as benchmarks. We conduct three tests for each checkpoint and take the average performance as the final result to ensure robust evaluation. We do not restrict the answer format in order to assess performance under typical usage conditions, and we leverage Gemini-2.0-Flash [Team et al., 2023] as the judge for evaluation.

4.2 Results and analysis

PEFT methods enable efficient updating of visual knowledge. Among the three data formats we design, Figure 6 shows that model using direct multiple-choice questions with concise answers yield faster and more effective learning during the visual knowledge acquisition phase. As training progresses, the model’s performance steadily improves in validation set, continuing to rise into the second epoch. In contrast, the other two formats exhibit early performance saturation, with gains peaking at the first training step and showing no further improvement thereafter. Notably, training on the visually knowledge-intensive LIVEVQA dataset—particularly with straightforward answers and multiple-choice questions—leads to a 4.2% improvement in MMMU.

Visual knowledge updating harm on model’s visual perception capability. In our experiments, we observe a consistent degradation in the model’s performance on MMStar, regardless of variations in rank and number of training steps, or training formats. Specifically, models trained using the simple answer format exhibit a performance drop from 65.80% to 58.16%. Notably, this decline plateaus even with continued training over multiple steps. These findings suggest an inherent conflict between enhancing visual knowledge through intensive updates and preserving the model’s visual perception capability.

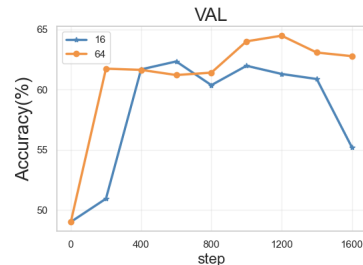


Figure 7: Visual knowledge updating with different LoRA rank.

Higher rank of LoRA yields better performance. Our ablation study across varying rank settings demonstrates that higher-rank LoRA configurations consistently enhance visual knowledge capabilities, particularly in assimilating recent visual entities. As illustrated in Figure 7, models with higher ranks consistently outperform their lower-rank counterparts, with an average improvement of 5.4% on the validation subset.

5 Related work

Visual knowledge. Visual knowledge, also known as world knowledge, refers to the ability to connect visual information with broader facts, concepts, and relationships about the real world [Marino et al., 2019, Schwenk et al., 2022]. This knowledge encompasses both factual information from external sources [Jiang et al., 2024, Cheng et al., 2025] and commonsense understanding about objects, their interactions, and contextual relationships in visual scenes [Xie et al., 2019, Wang et al., 2015, 2017, Jain et al., 2021, Shah et al., 2019]. The acquisition of visual knowledge involves leveraging external knowledge bases and structured repositories that ground visual elements in their broader conceptual context [Chen et al., 2022, Zhao et al., 2023, Caffagni et al., 2024, Yan and Xie, 2024, Abootorabi et al., 2025]. Recent developments are expanding visual knowledge from static data to “live” visual content, addressing challenges in being more helpful as real-time and real-life multimodal assistants [Jiang et al., 2024]. This live visual knowledge specifically pertains to understanding and reasoning about current news [Fu et al., 2022], emerging events [Yang et al., 2023], cultural phenomena [Nayak et al., 2024, Romero et al., 2024], and temporally relevant information that constantly evolves [Du. et al., 2025, Huang et al., 2024a]. Such live knowledge allows multimodal assistants to provide timely, relevant, and contextually appropriate responses to visual queries about ongoing situations and events. To address live-knowledge reasoning, frameworks like WebWatcher build a vision–language agent with coordinated tool use [Geng et al., 2025]. Meanwhile, MMSearch-R1 uses reinforcement learning to schedule search-tool invocation [Jinming Wu, 2025].

Synthetic data for knowledge update. Knowledge Updating, also known as continual learning, focuses on rapidly injecting the latest knowledge into pretrained models so they immediately recognize emerging concepts while retaining prior competencies [De Cao et al., 2021, Zhang et al., 2023, Huang et al., 2024a, Chen et al., 2024a, Jovanovic and Voss, 2024, He et al., 2025]. A major challenge in this process is catastrophic forgetting, where models lose previously acquired knowledge when learning new information, necessitating carefully-constructed high-quality data and specialized updating techniques [Luo et al., 2023, Huang et al., 2024b, Feng et al., 2024]. To address these challenges, synthetic datasets have emerged as a critical solution for continuous knowledge infusion without extensive retraining [Thede et al., 2025, Abdin et al., 2024]. For textual knowledge, frameworks like SynthLLM generate diverse, high-quality synthetic datasets by transforming existing corpora [Qin et al., 2025], while techniques such as Knowledge Direct Preference Optimization (KDPO) leverage synthetic examples for targeted factual updates [Rozner et al., 2024]. Previous research in language and code domain has successfully build up automatic synthetic framework for code api knowledge synchronizing [Liu et al., 2024b, Wang et al., 2025, Kumar and Kaur, 2024]. Our work introduces the first fully synthetic automatic engine for visual entity knowledge updating, which automatically collects new visual knowledge from online sources and leverages MLLMs to synthesize high-quality multimodal datasets. Through comprehensive experiments, we investigate the effectiveness of various fine-tuning methods, dataset formulations, and data quantity requirements, providing valuable insights for lifelong world model that must continuously incorporate emerging visual entities while preserving existing capabilities [Kim et al., 2023].

6 Conclusion

In this paper, we introduce LIVEVQA, the first synthetic data engine to feature up-to-date visual content paired with cross-modal, multi-hop reasoning questions—specifically designed to advance research in visual knowledge seeking and updating for MLLMs. Our experiments demonstrate that search-based multimodal RAG methods lead to significant improvements in handling contemporary visual information. Additionally, our PEFT studies provide valuable insights into balancing adapter capacity with the underlying model’s capabilities. We believe that the dataset, benchmark, and findings presented in this work establish a strong foundation for future research aimed at enhancing the knowledge seeking and updating abilities of MLLMs in dynamic visual contexts.

Acknowledgement

We thank Dongzhi Jiang from MM-Search team [Jiang et al., 2024], Jingming Wu from MM-search-R1 team [Jinming Wu, 2025] and Jieyu Zhang from University of Washington for their valuable insights and discussion.

References

- Thomas Decker, Ralf Gross, Alexander Koebler, Michael Lebacher, Ronald Schnitzer, and Stefan H Weber. The thousand faces of explainable ai along the machine learning life cycle: industrial reality and current state of research. In *International Conference on Human-Computer Interaction*, pages 184–208. Springer, 2023.
- Qifeng Li, Xiaosong Jia, Shaobo Wang, and Junchi Yan. Think2drive: Efficient reinforcement learning by thinking with latent world model for autonomous driving (in carla-v2). In *European Conference on Computer Vision*, pages 142–158. Springer, 2024a.
- OpenAI. Gpt-4.1. <https://openai.com/index/gpt-4-1/>, 2025. Large language model. Accessed: 2025-05-11.
- Dave Citron. Deep research is now available on gemini 2.5 pro experimental, April 2025. URL <https://blog.google/products/gemini/deep-research-gemini-2-5-pro-experimental/>. Google Blog, accessed May 11, 2025.
- Anthropic. Introducing claude 3.7 sonnet and claude code, February 2025. URL <https://www.anthropic.com/news/claude-3-7-sonnet-and-claude-code>. Accessed: 2025-05-11.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.
- Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.
- Thuy Nguyen, Dang Nguyen, Hoang Nguyen, Thuan Luong, Long Hoang Dang, and Viet Dac Lai. Owlviz: An open-world benchmark for visual question answering. *arXiv preprint arXiv:2503.07631*, 2025.
- Haqee Ishfaq, Qingfeng Lan, Pan Xu, A Rupam Mahmood, Doina Precup, Anima Anandkumar, and Kamyar Azizzadenesheli. Provable and practical: Efficient exploration in reinforcement learning via langevin monte carlo. *arXiv preprint arXiv:2305.18246*, 2023.
- Nimrod Shabtay, Felipe Maia Polo, Sivan Doveh, Wei Lin, M Jehanzeb Mirza, Leshem Chosen, Mikhail Yurochkin, Yuekai Sun, Assaf Arbelle, Leonid Karlinsky, et al. Livexiv—a multi-modal live benchmark based on arxiv papers content. *arXiv preprint arXiv:2410.10783*, 2024.
- Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*, 2024b.
- Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Jiayi Lei, Pengshuo Qiu, Pan Lu, Zehui Chen, Chaoyou Fu, Guanglu Song, et al. Mmsearch: Benchmarking the potential of large models as multi-modal search engines. *arXiv preprint arXiv:2409.12959*, 2024.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*, 2024.
- Google. Gemini 2.5 pro. <https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/>, 2025.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivi re, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.
- Meta. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation, 2025. URL <https://ai.meta.com/blog/llama-4-multimodal-intelligence/>.
- OpenAI. Introducing openai o3 and o4-mini, 2025. URL <https://openai.com/index/introducing-o3-and-o4-mini/>.
- xAI. Introducing Grok 4, 2025.
- Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity’s last exam. *arXiv preprint arXiv:2501.14249*, 2025.
- Arjun Panickssery, Samuel Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802, 2024.
- Chujie Gao, Qihui Zhang, Dongping Chen, Yue Huang, Siyuan Wu, Zhengyan Fu, Yao Wan, Xiangliang Zhang, and Lichao Sun. The best of both worlds: Toward an honest and helpful large language model. *arXiv preprint arXiv:2406.00380*, 2024.
- Mathieu Ravaut, Bosheng Ding, Fangkai Jiao, Hailin Chen, Xingxuan Li, Ruochen Zhao, Chengwei Qin, Caiming Xiong, and Shafiq Joty. How much are large language models contaminated? a comprehensive survey and the llmsanitize library. *arXiv preprint arXiv:2404.00699*, 2024.
- Zhen Zeng, Leijiang Gu, Xun Yang, Zhangling Duan, Zenglin Shi, and Meng Wang. Visual-oriented fine-grained knowledge editing for multimodal large language models. *arXiv preprint arXiv:2411.12790*, 2024.
- Qizhou Chen, Chengyu Wang, Dakan Wang, Taolin Zhang, Wangyue Li, and Xiaofeng He. Lifelong knowledge editing for vision language models with low-rank mixture-of-experts. *arXiv preprint arXiv:2411.15432*, 2024a.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
- Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*, 2024a.
- Chenlong Wang, Zhaoyang Chu, Zhengxiang Cheng, Xuyi Yang, Kaiyue Qiu, Yao Wan, Zhou Zhao, Xuanhua Shi, and Dongping Chen. Codesync: Synchronizing large language models with dynamic code evolution at scale. *arXiv preprint arXiv:2502.16645*, 2025.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024b.
- 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.
- Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019.
- Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022.
- Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, et al. Simplevqa: Multimodal factuality evaluation for multimodal large language models. *arXiv preprint arXiv:2502.13059*, 2025.

- Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.
- Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570*, 2015.
- Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427, 2017.
- Aman Jain, Mayank Kothiyari, Vishwajeet Kumar, Preethi Jyothi, Ganesh Ramakrishnan, and Soumen Chakrabarti. 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*, pages 2491–2498, 2021.
- Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8876–8884, 2019.
- Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W Cohen. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. *arXiv preprint arXiv:2210.02928*, 2022.
- Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Xuan Long Do, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, et al. Retrieving multimodal information for augmented generation: A survey. *arXiv preprint arXiv:2303.10868*, 2023.
- Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1818–1826, 2024.
- Yibin Yan and Weidi Xie. Echosight: Advancing visual-language models with wiki knowledge. *arXiv preprint arXiv:2407.12735*, 2024.
- Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdih Soleymani Baghshah, and Ehsaneddin Asgari. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. *arXiv preprint arXiv:2502.08826*, 2025.
- Xingyu Fu, Ben Zhou, Ishaan Preetam Chandratreya, Carl Vondrick, and Dan Roth. There is a time and place for reasoning beyond the image. *arXiv preprint arXiv:2203.00758*, 2022.
- Zhenguo Yang, Jiale Xiang, Jiuxiang You, Qing Li, and Wenyan Liu. Event-oriented visual question answering: The e-vqa dataset and benchmark. *IEEE Transactions on Knowledge and Data Engineering*, 35(10):10210–10223, 2023. doi: 10.1109/TKDE.2023.3267036.
- Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd Van Steenkiste, Lisa Anne Hendricks, Aishwarya Agrawal, et al. Benchmarking vision language models for cultural understanding. *arXiv preprint arXiv:2407.10920*, 2024.
- David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, et al. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*, 2024.
- Yuntao Du., Kailin Jiang, Zhi Gao, Chenrui Shi, Zilong Zheng, Siyuan Qi, and Qing Li. MMKE-bench: A multimodal editing benchmark for diverse visual knowledge. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=v8qABSeeK0>.
- Han Huang, Haitian Zhong, Tao Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. Vlkeb: A large vision-language model knowledge editing benchmark. *arXiv preprint arXiv:2403.07350*, 2024a.
- Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Yida Zhao, Kuan Li, et al. Webwatcher: Breaking new frontier of vision-language deep research agent. *arXiv preprint arXiv:2508.05748*, 2025.
- Jinming Wu. Multimodal-search-r1: Incentivizing llms to search, 2025. URL <https://github.com/EvolvingLMs-Lab/multimodal-search-r1>.
- Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. *arXiv preprint arXiv:2104.08164*, 2021.

- Zihan Zhang, Meng Fang, Ling Chen, Mohammad-Reza Namazi-Rad, and Jun Wang. How do large language models capture the ever-changing world knowledge? a review of recent advances. *arXiv preprint arXiv:2310.07343*, 2023.
- Mladjan Jovanovic and Peter Voss. Towards incremental learning in large language models: A critical review. *arXiv preprint arXiv:2404.18311*, 2024.
- Guoxiu He, Xin Song, and Aixin Sun. Knowledge updating? no more model editing! just selective contextual reasoning. *arXiv preprint arXiv:2503.05212*, 2025.
- Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*, 2023.
- Jianheng Huang, Leyang Cui, Ante Wang, Chengyi Yang, Xinting Liao, Linfeng Song, Junfeng Yao, and Jinsong Su. Mitigating catastrophic forgetting in large language models with self-synthesized rehearsal. *arXiv preprint arXiv:2403.01244*, 2024b.
- Yujie Feng, Xu Chu, Yongxin Xu, Zexin Lu, Bo Liu, Philip S Yu, and Xiao-Ming Wu. Kif: Knowledge identification and fusion for language model continual learning. *arXiv preprint arXiv:2408.05200*, 2024.
- Lukas Thede, Karsten Roth, Matthias Bethge, Zeynep Akata, and Tom Hartvigsen. Understanding the limits of lifelong knowledge editing in llms. *arXiv preprint arXiv:2503.05683*, 2025.
- Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.
- Zeyu Qin, Qingxiu Dong, Xingxing Zhang, Li Dong, Xiaolong Huang, Ziyi Yang, Mahmoud Khademi, Dongdong Zhang, Hany Hassan Awadalla, Yi R Fung, et al. Scaling laws of synthetic data for language models. *arXiv preprint arXiv:2503.19551*, 2025.
- Amit Rozner, Barak Battash, Lior Wolf, and Ofir Lindenbaum. Knowledge editing in language models via adapted direct preference optimization. *arXiv preprint arXiv:2406.09920*, 2024.
- Zeyu Leo Liu, Shrey Pandit, Xi Ye, Eunsol Choi, and Greg Durrett. Codeupdatearena: Benchmarking knowledge editing on api updates. *arXiv preprint arXiv:2407.06249*, 2024b.
- Aditya Kumar and Manjinder Kaur. Live code sync. In *Advances in AI for Biomedical Instrumentation, Electronics and Computing*, pages 554–559. CRC Press, 2024.
- Yujin Kim, Jaehong Yoon, Seonghyeon Ye, Sangmin Bae, Namgyu Ho, Sung Ju Hwang, and Se-Young Yun. Carpe diem: On the evaluation of world knowledge in lifelong language models. *arXiv preprint arXiv:2311.08106*, 2023.
- OpenAI. Gpt-4o, 2024. URL <https://openai.com/gpt-4o>. Accessed: 2025-05-01.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- ▷ You should answer [Yes], [No], or [NA].
- ▷ [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- ▷ Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes]" is generally preferable to "[No]", it is perfectly acceptable to answer "[No]" provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No]" or "[NA]" is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- ▷ **Delete this instruction block, but keep the section heading “NeurIPS paper checklist”,**
- ▷ **Keep the checklist subsection headings, questions/answers and guidelines below.**
- ▷ **Do not modify the questions and only use the provided macros for your answers.**

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [Yes]

Justification: We did it in the abstract and instruction.

Guidelines:

- ▷ The answer NA means that the abstract and introduction do not include the claims made in the paper.
- ▷ The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- ▷ The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- ▷ It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discuss the limitation and future work in the draft.

Guidelines:

- ▷ The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- ▷ The authors are encouraged to create a separate "Limitations" section in their paper.
- ▷ The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- ▷ The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- ▷ The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- ▷ The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- ▷ If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- ▷ While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: For necessary theory assumptions, we provide the proofs in the appendix.

Guidelines:

- ▷ The answer NA means that the paper does not include theoretical results.
- ▷ All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- ▷ All assumptions should be clearly stated or referenced in the statement of any theorems.
- ▷ The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- ▷ Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- ▷ Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Yes, we fully disclose the details of experiments of different tasks. We also provide the source code for reproducibility.

Guidelines:

- ▷ The answer NA means that the paper does not include experiments.

- ▷ If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- ▷ If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- ▷ Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- ▷ While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Yes, we provide the source code and datasets in the supplementary materials.

Guidelines:

- ▷ The answer NA means that paper does not include experiments requiring code.
- ▷ Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- ▷ While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- ▷ The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- ▷ The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- ▷ The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- ▷ At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- ▷ Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Yes, we disclose all related details in the draft.

Guidelines:

- ▷ The answer NA means that the paper does not include experiments.
- ▷ The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- ▷ The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Error bars are not reported because it would be too computationally expensive (especially for some advanced reasoning models).

Guidelines:

- ▷ The answer NA means that the paper does not include experiments.
- ▷ The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- ▷ The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- ▷ The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- ▷ The assumptions made should be given (e.g., Normally distributed errors).
- ▷ It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- ▷ It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- ▷ For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- ▷ If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: All experiments are running by API or $8 \times$ NVIDIA A800 (80G).

Guidelines:

- ▷ The answer NA means that the paper does not include experiments.
- ▷ The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- ▷ The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- ▷ The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Yes, we follow the code of ethics.

Guidelines:

- ▷ The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- ▷ If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- ▷ The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We include some basic discussion of its broader impact in the introduction.

Guidelines:

- ▷ The answer NA means that there is no societal impact of the work performed.
- ▷ If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- ▷ Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- ▷ The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- ▷ The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- ▷ If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: All our data undergoes carefully filtered by advanced models and human. We release our dataset under CC-4.0-NC-BY license.

Guidelines:

- ▷ The answer NA means that the paper poses no such risks.
- ▷ Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- ▷ Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- ▷ We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [\[Yes\]](#)

Justification: We use all assets correctly based on their licenses.

Guidelines:

- ▷ The answer NA means that the paper does not use existing assets.
- ▷ The authors should cite the original paper that produced the code package or dataset.
- ▷ The authors should state which version of the asset is used and, if possible, include a URL.
- ▷ The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- ▷ For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- ▷ If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- ▷ For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- ▷ If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New Assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[Yes\]](#)

Justification: We introduce the new assets and the documentation is provided in the supplementary materials.

Guidelines:

- ▷ The answer NA means that the paper does not release new assets.
- ▷ Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- ▷ The paper should discuss whether and how consent was obtained from people whose asset is used.
- ▷ At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and Research with Human Subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[Yes\]](#)

Justification: Yes, we provide screenshots and human agreement rates (or pass rate) for every human evaluation.

Guidelines:

- ▷ The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- ▷ Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- ▷ According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our research does not require the IRB.

Guidelines:

- ▷ The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- ▷ Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- ▷ We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- ▷ For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

A Benchmark Modifications

We implement several key modifications during the construction and finalization of our benchmark to ensure data quality and clarity of evaluation. Specifically, we remove all “*Reason*” type questions from the News and Video subsets, as the answers to these questions often rely on subjective inference rather than objective visual knowledge. Furthermore, we also delete all “*Time*” type questions that contain ambiguous or relative temporal expressions (e.g., “*last Friday*”). This elimination resolves any potential answer ambiguity and ensures the precision of the model evaluation.

B Limitations

Despite our efforts to build a fully automated framework for crawling latest visual knowledge and construct it into dataset, limitations remains. Retrieving news and video data older than two months was extremely difficult due to crawling limitations on platforms like YouTube, potentially causing date-category imbalance. arXiv paper crawling often encountered CAPTCHAs, sometimes leading to incomplete downloads.

Although our pipeline leverage capable models like GPT-4.1 at each LLM/MLLM-in-the-loop and validated by human annotation of over 500 balanced samples per category (achieving <3% error rates), may still produce occasional errors in QA generation or filtering, leading to some flawed dataset samples. Furthermore, due to limited computational resources, our visual knowledge update experiments were restricted to parameter-efficient methods (i.e., LoRA, and DoRA) and the Qwen2.5-VL model family, without exploring full fine-tuning or a wider range of models.

Future work will address these points by including more baselines and investigating more efficient update techniques. Nevertheless, we are committed to continuously crawling new data and updating this dataset annually. This will help ensure the availability of a non-contaminated dataset for future research in visual knowledge seeking and updating.

C Dataset collection details

C.1 News

In order to efficiently extract news article data from massive online resources, we design an automated processing. We first use strict URL verification to ensure that only valid article pages are processed, excluding irrelevant content. For the verified URLs, we identify article titles and candidate images. Next, we select visual materials that are highly relevant to the news content. Finally, we use a duplicate article detection and prevention mechanism to identify and remove content that has been included or repeated in the current processing session.

▷ **Systematic URL validation and canonicalization protocol.** The initial stage of data processing involves a meticulous URL validation and canonicalization protocol to ensure that only valid article pages are pursued for content extraction. Each prospective URL first undergoes a canonicalization procedure, where all query parameters (components following a ‘?’ symbol) are systematically stripped, yielding a standardized base URL essential for consistent referencing and duplicate prevention.

Following canonicalization, a sophisticated validation mechanism ascertains whether the URL corresponds to a substantive news article rather than an ancillary page such as a category index, multimedia gallery, or author profile. This validation employs a hierarchical, domain-sensitive strategy. For recognized news sources (CNN, BBC, Forbes, Variety, Associated Press News), the system dispatches URLs to dedicated, site-specific validation functions. These functions implement tailored rules, typically leveraging regular expressions and string pattern matching, to identify legitimate article URL structures (e.g., date-based paths like YYYY/MM/DD for CNN, specific segments like /news/articles/ for BBC, or complex patterns involving /sites/ and author/date components for Forbes) while concurrently excluding known non-article paths (e.g., /video/, /gallery/, /tag/).

In the event a URL does not originate from these predefined sources, or if specific rules are not met, a generic fallback validation is invoked. This generic checker compares the URL against a curated list of common article-indicating path segments (e.g., /article/, /story/, .html) and, conversely, against a list of non-article indicators (e.g., /category/, /search/). Only URLs that successfully

pass this rigorous, multi-layered validation sequence are advanced to subsequent processing stages. Utility functions further support this by enabling secure joining of base URLs with relative paths discovered on index pages.

▷ **Robust HTML element extraction and initial content identification.** Once a URL is validated as a potential article source, its HTML content is parsed to extract key informational components, primarily the title and preliminary image candidates. This process utilizes robust CSS selectors, safeguarded by wrapper functions (`safe_select`, `safe_select_one`) that gracefully handle exceptions during selector execution, returning empty lists or null objects instead of halting the process. Title extraction (`extract_title`) employs a tiered strategy: for designated news providers (CNN, BBC, Forbes), a predefined dictionary maps site identifiers to a prioritized list of CSS selectors known to target article headlines (e.g., `h1.pg-headline` for CNN, `h1.article-headline` for Forbes). The system iterates through these selectors until a valid title element, defined as yielding a textual content Condiciones de Uso y Privacidad Política de Privacidad y Cookies Publicidad Anunciarse con nosotros Contacto Aviso legal Política de Cookies (UE) Transparencia sobre el esclavismo moderno (Reino Unido) (con una longitud superior a cinco caracteres tras stripping), is found. If site-specific selectors fail, or if the source is not explicitly defined, a generic fallback attempts to extract content from the primary `<h1>` tag.

Similarly, an enhanced image detection routine (`enhance_image_detection`) is initiated to identify potential article images. This routine also operates on a dispatch model, invoking specialized detection functions (e.g., `detect_cnn_images`, `detect_forbes_images`) for supported news domains. These site-specific functions employ a list of CSS selectors targeting common image elements or containers (e.g., `.media__image` for CNN, `.article-image` for Forbes). As a secondary measure within these functions, or if generic detection (`detect_generic_images`) is active, the system may iterate through all `<img>` tags within the document. This comprehensive scan often includes filters to discard irrelevant images based on keywords in their `src` attribute (e.g., `'icon'`, `'logo'`, `'avatar'`) or their explicit dimensions (e.g., retaining only images larger than 200x200 pixels), and positive identification through keywords in `src` or CSS `class` attributes (e.g., `'photo'`, `'image'`, `'hero'`, `'main-image'`). These initial detection steps provide a candidate pool of images for more intensive downstream filtering.

▷ **Advanced image curation via multi-stage filtering and deduplication.** Following initial HTML parsing, a sophisticated image curation pipeline processes the identified candidate images to select a concise set of high-relevance visuals for each article. The process commences with the `extract_article_images` function, which systematically gathers image URLs and associated captions. This involves several targeted extraction methods: (1) Prioritized retrieval of images specified in Open Graph (`og:image`) and Twitter Card (`twitter:image`) meta tags, as these often represent the primary article visual. (2) Searching within common semantic HTML structures using a list of `figure_selectors` (e.g., `<figure>`, `.image-container`, `.media-with-caption`) to identify images and their corresponding captions (e.g., from `<figcaption>`, `.caption`). Relative image URLs are resolved to absolute paths, and base64-encoded images are typically ignored. (3) A broader scan within primary content blocks (identified by `content_selectors` like `<article>` or `.article-body`) for any remaining `<img>` tags, with attempts to infer captions from adjacent elements.

The aggregated list of image candidates then undergoes a rigorous filtering and deduplication process within the `filter_images` function. For each image URL, its content is fetched, and its dimensions (width, height) are determined using the Pillow (PIL) library, allowing for the calculation of pixel area. Images that cannot be fetched or processed are discarded. Crucially, a visual deduplication step (`deduplicate_images`) is performed. If image processing libraries like OpenCV or scikit-image are available, this function compares images pairwise. Images are resized (e.g., to 128x128 grayscale) and their visual similarity is computed, potentially using a hybrid approach involving color histogram comparison (via OpenCV's `calcHist` and `compareHist`) and Structural Similarity Index Measure (SSIM, via scikit-image). Images with a similarity score exceeding a predefined threshold (e.g., 0.85) are considered duplicates, and the one with the smaller area is typically removed. If these libraries are unavailable, a fallback deduplication uses simpler heuristics based on common URL filename segments and dimensional proximity. The unique, high-quality images are then sorted in descending order by pixel area.

A final selection pass retains a maximum of four images per article. Furthermore, any image whose area is less than a specified fraction (*e.g.*, 30%) of the largest image’s area for that article is discarded, ensuring that only prominent and contextually significant visuals are preserved. This multi-faceted approach ensures a small, yet highly relevant, set of images for each news item. Utility functions also support downloading these curated images and strategically inserting placeholder tags (*e.g.*, `<imgN>`) into the article text.

▷ **Hierarchical duplicate article detection and prevention:** To maintain dataset integrity and prevent redundant data processing, a robust, multi-layered duplicate detection mechanism is implemented through the `DuplicateChecker` class. This system addresses both historical duplicates (against previously collected articles) and intra-session duplicates (within the current data collection run). Upon initialization, the checker loads metadata from all existing article collections, typically stored as JSON files (*e.g.*, `hot_topics*.json`), into memory. This historical data populates sets of known URLs (`seen_urls`) and titles (`seen_titles`) for rapid initial lookups.

When a new article candidate is processed (`is_duplicate_topic`), it is first checked against these in-memory sets for exact URL or title matches. If no exact match is found, a more thorough comparison against the loaded historical topics is performed. This involves several checks in a specific order of computational cost: (1) Exact URL equality. (2) Precise matching of the initial five words of the article titles. (3) Keyword-based similarity: Titles are tokenized into keyword sets (excluding common stopwords), and if the Jaccard index or a similar overlap metric (*e.g.*, intersection size relative to the smaller set, thresholded at 70% for sets with at least 3 keywords) between the new and an existing title’s keywords is high, it’s flagged as a potential duplicate. (4) Full title similarity: If previous checks are inconclusive, a more computationally intensive string similarity algorithm (*e.g.*, Levenshtein distance based, with a similarity score > 0.85 indicating a duplicate) is applied between the new and existing titles.

For real-time, intra-session duplicate prevention during active crawling (`is_duplicate_realtime`), a slightly streamlined version of this logic is employed, primarily leveraging the dynamically updated `seen_urls` and `seen_titles` sets, along with the prefix, keyword, and full title similarity checks against the already loaded historical topics. If an article passes all these checks, its URL and title are added to the session’s seen sets to prevent its re-processing. This hierarchical approach, combining hash-based lookups with increasingly sophisticated content analysis, ensures high accuracy in duplicate detection while managing computational overhead.

C.2 Videos

Our automatic system initially employs YouTube Data API for video discovery and rule-based filtering, followed by the downloading and preprocessing such as format conversion. Then, we process subtitle as context for visual content and perform LLM-driven subtitle refinement to conduct video segmentation base on subtitle timestamp. Finally, we conduct keyframe extraction and multi-stage filtering to get the curated textual and visual metadata.

- ▷ **Video downloading, initial filtering, and preprocessing.** The initial stage focuses on sourcing and preparing video data. Firstly, we utilize the YouTube Data API to conduct targeted searches based on predefined date ranges, language (English), and content categories (news, education, entertainment, technology). For the retrieved videos, we immediately verify their duration (ensuring it does not exceed 10 minutes) and the availability of English subtitles, thereby performing an initial screening for video resources that meet these basic requirements. Subsequently, the system automatically downloads the videos that pass this initial screening, along with their corresponding English subtitle files. To meet the technical requirements for subsequent keyframe extraction, all downloaded videos automatically undergo format conversion from VP9 to H.264. Finally, detailed metadata for each video is extracted, creating structured data records, and the entire workflow integrates comprehensive logging and error handling mechanisms to ensure data collection accuracy and operational stability.
- ▷ **Semantic content processing and keyframe extraction.** For each video, subtitles and corresponding timestamps are processed to extract textual content. This extracted text is then refined using a Large Language Model (LLM), which adds punctuation and segments the content into distinct topics, each associated with specific time intervals. Subsequently, the videos are clipped according to these time intervals to ensure thematic consistency within each resulting segment. This segmentation is crucial to prevent misalignment between subsequently selected images and

their corresponding textual context, which could otherwise lead to the large model generating ‘hallucinations’. Following this, for these segmented video clips, keyframes are extracted using UVD (Uniform Video Dicing), as the full video content is often too extensive, necessitating the distillation of useful visual information. Given that a single video segment can yield a large number of keyframes, an initial deduplication step is performed using image perceptual hashing with a Hamming distance threshold of 10. Concurrently, the Laplacian operator is employed to calculate image clarity to retain only the clearest image among a set of visually similar ones.

- ▷ **Advanced image curation and multimodal content generation.** We observe that for some news screenshots, subtitle overlays introduce significant textual interference. To address this, DocLayout-YOLO technology is utilized to crop these images, aiming to remove or minimize such interfering textual information that might directly reveal key details and unduly influence the large model’s responses. These cropped images then undergo a secondary deduplication process, using the same perceptual hashing method as before, but with a revised Hamming distance threshold of 25. After these operations, some may still exhibit an ambiguous correspondence with the textual information. Therefore, we design an image filtering program that leverages GPT-4.1 to select the designs images that best align with the textual content while exhibiting minimal textual interference from overlays. Finally, these highly filtered images, in conjunction with the translated and refined video titles and subtitle information, are used to generate question-answering (QA) pairs.

C.3 Academic papers

To construct a comprehensive dataset derived from scholarly articles on arXiv, we have developed a systematic collection and processing pipeline. This pipeline is designed to extract and curate textual and visual information from scientific papers. Subsequently, a MLLM is utilized to identify and select key images; this selection is further refined through human validation. Finally, the data is used to synthesize QAs of our dataset.

- ▷ **Article data preprocessing.** Our process commences with the retrieval of scholarly articles from arXiv. For each crawled paper, essential metadata—including the title, abstract, and author information—is primarily acquired from the arXiv "abs" pages. Following metadata acquisition, the full HTML content of the article’s page is fetched. Robust HTML parsing libraries, such as BeautifulSoup, are then employed to process this content. The images that are excessively small or that fail to load correctly are systematically filtered out. Furthermore, we leverage MLLM to identify key images in the articles. A detailed discussion of this process will be presented in subsequent section.
- ▷ **Questions, options and detailed generation.** For the Level 1 data, we employ a template-based question generation approach to elicit the title or first author’s name associated with a given image. Simultaneously, we generate distractor options using either random selection or LLMs. For detailed data, we leverage LLMs to produce a concise summary of the paper based on its abstract. For the Level-2 data, We leverage prompts to elicit the generation of detailed questions pertaining to the image’s context or abstract, high-level summaries. These questions are specifically designed to be unanswerable using only the image information or knowledge inherently derived from the image itself.

C.4 Human annotation details

Every experiment using LLM/MLLM is validated with human annotated ground truth and agreement. We provide detailed instructions and annotation environments. The annotation is conducted by 4 authors of this paper independently. All the annotations are conducted under Streamlit⁶. As acknowledged, the diversity of annotators plays a crucial role in reducing bias and enhancing the reliability of the benchmark. These annotators have rich knowledge in this domain, with different genders, ages, and educational backgrounds. To ensure the annotators can proficiently mark the data, we provide them with detailed tutorials, teaching them how to evaluate model responses more objectively. Specifically, they are required to give judgments without bias like answer lengths, and certain names of the response. All process using LLM/MLLM are listed as follows:

- ▷ **YouTube video - subtitle parsing and event segmentation with GPT-4.1.** Subtitle files are parsed to extract word tokens based on timestamp tags. Since the text often lacks punctuation, we utilize

⁶<https://streamlit.io/>

GPT-4.1 to reconstruct proper sentence boundaries and restore punctuation. Subsequently, GPT-4.1 segments these processed subtitles into coherent event segments. This segmentation is guided by both textual and temporal signals: short caption intervals usually indicate topical continuity, whereas longer pauses imply topic shifts. Segment boundaries are only introduced when a clear topical change is detected. To assess the structural appropriateness of text segmented by GPT-4.1, we develop an annotation tool in Figure 8 with a passing rate of 98%.

- ▷ **YouTube video - images selecting with GPT-4.1.** After multiple rounds of preprocessing, each video segment still contains some images that exhibit weak relevance to the corresponding textual information. To address this, we employ the GPT-4.1 model to execute a fine-grained selection procedure, identifying and selecting the Top-K images with the strongest textual relevance from each segment for subsequent QA generation. To evaluate the appropriateness and effectiveness of GPT-4.1’s image selection strategy, we design an annotation tool in Figure 10. Manual evaluations conducted using this tool reveals that the images selected through this method achieves a 96% passing rate.
- ▷ **YouTube video - text evaluating before generating QAs with GPT-4.1.** During the processing of textual information from videos, we observe that some videos contain sparse textual content, such as dialogues lacking specific details. These texts are prone to inducing hallucinations in large language models when used for generating QA pairs. To mitigate this issue, we employ GPT-4.1 to effectively filter out these low-information-density texts. Furthermore, to evaluate the rationale and accuracy of GPT-4.1’s filtering strategy, we design a dedicated annotation tool in Figure 11. Manual assessments based on this tool demonstrated that GPT-4.1 achieved a 99% accuracy rate in identifying and filtering out such unsuitable texts.
- ▷ **Academic paper - key image selection with GPT-4.1.** We leverage GPT-4.1 to identify the key images in the articles. The key images are those that uniquely identifies a specific research paper, distinguishing it from other publications. The prompt is available in Appendix F, which shows our preference on image selection. To validate whether the model’s selection is reasonable, we developed a labeling interface using Streamlit in Figure 12. Manual assessments demonstrate that GPT-4.1 achieved a 94% accuracy rate.
- ▷ **News article - image filter with GPT-4.1.** we retain only images clearly depicting current social topics and discard static objects, logos, duplicates, generic or decorative visuals, or any image lacking unique insight into the article’s core subject. Each kept image must add essential information unobtainable from text alone and be indispensable to understanding the story. Generate factual captions using only explicitly stated details. We employ GPT-4.1 to finish this job and design the annotation tool in Figure 13. Manual assessments demonstrate that GPT-4.1 achieved a 96% accuracy rate.
- ▷ **Level-1 question generation with GPT-4.1.** We generate basic QA pairs beginning “Based on the provided image,” whose answer is explicitly stated in the article. The question must make sense with the image alone, avoid celebrities, counts, logos or text cues, aise. We employ GPT-4.1 to finish this job and design the annotation tool in Figure 14. Manual assessments show GPT-4.1 achieved a 93% accuracy rate.
- ▷ **Level-1 question filtering with GPT-4.1.** We discard articles lacking meaningful context (e.g., only noting a broadcast question). We always keep questions whose answers are non–major-figure names, specific named locations, named events, specific products, or organizations. We employ GPT-4.1 to remove any unqualified QA pairs and use an annotation tool to evaluate. The annotation tool is shown in Figure 15. The accuracy of GPT-4.1 is 97%.
- ▷ **Level-2 question generation with GPT-4.1.** We construct multi-hop QA pairs and restrict to one of seven categories. We design decoys using partial truths, misconceptions or misleading links. We use GPT-4.1 to design the QA pairs and use the annotation tool in Figure 16. The accuracy of GPT-4.1 is 98%.
- ▷ **Level-2 Question filtering with GPT-4.1.** Some questions may be too hard and can’t be verified. So We keep the question only if the correct answer can be obtained after combining the information of the news article itself, the picture information, and the title information. Otherwise, the question that cannot be answered will be discarded. We use GPT-4.1 to filter this kind of QA pairs and evaluate it using the annotation tool in Figure 17 to evaluate it performance. The accuracy of GPT-4.1 is 99%.
- ▷ **Detailed and reasoning answer generation with GPT-4.1** We generate answers including detailed reasoning processes for subsequent training. We first ask the model to output the answer, then output information about the event corresponding to the image, and finally reason about the

Subtitle Labeling Assistant Tool

← Previous Page
Next Page →

Current Page 1 / Total 367 Pages

Subdirectory 1: VvOPCCIY5w0/VvOPCCIY5w0_text_metadata.json

There are 1 segments

Segment 1: 00:00:02.434 → 00:00:24.870

Text: Terrorist Abdul Rehman, a resident of Milkipur in Ayodhya, went to Ayodhya and did a recce of the Ram Temple. After this, as per the plan, he reached Faridabad near Delhi and started living under the name Shankar. After receiving the order to attack from Pakistan, the ISI handler in Faridabad handed over two grenades to terrorist Abdul Rehman. Carrying the grenades, terrorist Abdul Rehman left to catch the train to Ayodhya and the investigating agencies arrested him along with the grenades.

Is Outro Segment: No

Please judge the segmentation of this subdirectory (video) (multiple choices allowed):

Correct Segment...

Different news in the same segment

Same news in different segments

Figure 8: Human Annotation - YouTube Video - Parsing Subtitle by GPT4.1.

question. We use GPT-4.1 to do this and evaluate it by the annotation tool shown in Figure18. The results shows that GPT-4.1 achieve a 99% accurate rate.

- ▷ **Judging answer with Gemini-2.5-flash-Preview** We used the same evaluation methodology as the widely recognized de facto benchmark, SimpleQA. Specifically, the evaluation was conducted on Gemini-2.5-flash-Preview. To ensure reliability, we randomly sample 300 samples—100 each from news, videos, and academic papers—and manually verified the results of the evaluation model. The annotation tool is shown in Figure19, 20 and 21. The results show that the model achieved 100% accuracy. We are confident that the model is capable of performing well on this task.

To ensure the quality and recency of the video data, we conduct a rigorous manual verification process. Every video data entry is validated with human annotated verification to ensure that selected content does not depict historical events, thus minimizing the occurrence of outdated content. We provide detailed verification instructions and environments. Manual verification is carried out by our authors independently. Specifically, we manually annotate 500 data entries to verify and prevent such situations. As acknowledged, the rigorous manual process plays a crucial role in ensuring the timeliness and relevance of the video dataset. The results show that our video data are 100% recent events, rather than old events that occurred in the past. This annotation tool is shown in Figure22.

D Detailed experimental setups

Models. The detailed specifications of all models evaluated in the experiments of Section 3 are provided in Table 5. The models trained for the experiments in Section 4 are also included in Table 5.

Metric: Guessing strategy and F-score. Originate from Wei et al. [2024], while F-score is a good metric in some ways, the issue with it is that it incentivizes the model to always guess when it is at least 50% sure that it can get the correct answer. To understand why this is the case, consider the following expression for the F-score:

$$F\text{-score} = \frac{2}{\frac{c+i}{c} + \frac{c+i+n}{c}} = \frac{2c}{2c + 2i + n},$$

where:

- ▷ c is the number of correct answers,
- ▷ i is the number of incorrect answers, and
- ▷ n is the number of non-answered questions.



Figure 9: Preprocessing - YouTube Video - DocLayout-YOLO.

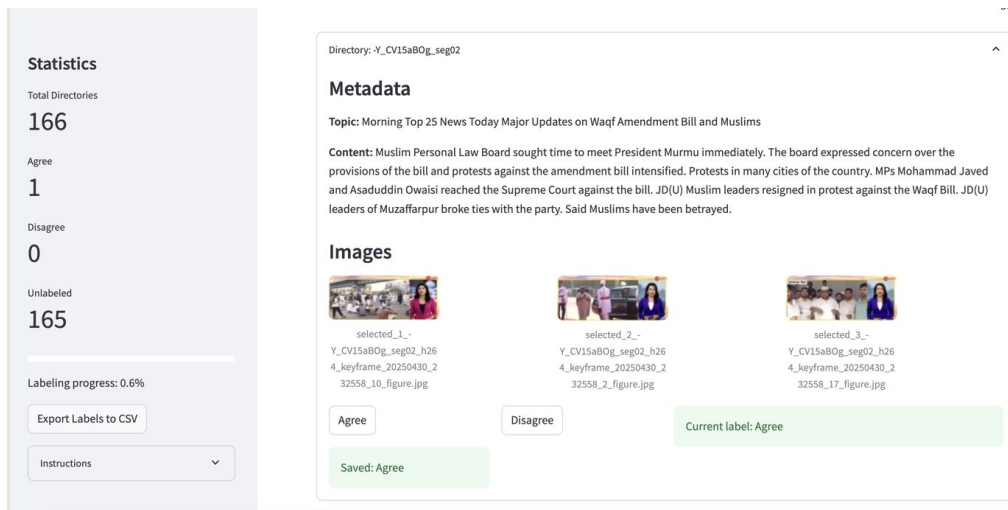


Figure 10: Human Annotation - YouTube Video - Selecting Images for Video Data.

Filter Options

Filter by Meaningfulness:

☒ All
☐ Meaningful Only
☐ Not Meaningful Only

Search Keywords:

Statistics

Total Entries

3622

Meaningful Entries

3169

Not Meaningful Entries

453

Content Details

ID: 1_174689722583

Topic: Amit Shah should apologize

Text Content:

Now there is a fashion to say Ambedkar Ambedkar Ambedkar Ambedkar. If you took the name of God, you would have gone to heaven in seven lives. Those who insulted Baba Saheb throughout their life, those who ignored the principles of Baba Saheb throughout their life, those who did not let Baba Saheb get Bharat Ratna as long as they were in power, those who flouted the principles of reservation as long as they were in power, those people today want to spread confusion in the name of Baba Saheb Ambedkar.

Source: YouTube

Meaningfulness Evaluation: ☒ Meaningful

Label this entry:

Agree

Save Label

Evaluation Reason:

The text criticizes political hypocrisy regarding Ambedkar, referencing specific actions and people, suitable for news or analysis.

Deploy

Related Image

Figure 11: Human Annotation - Youtube Video - Evaluating Text for QA Generating.

Global Feedback Statistics

Total Marked

432

Reasonable Ratio

93.3%

Settings

Data Directory Path

Search Papers

Image 2: x2.png (Rank: 1) - Selected

Image Index: 1

Image Caption:

Figure 2: Bootstrap3D data generation pipeline that consists of 1) using LLM to generate diverse text prompts 2) employing the T2I model to generate single-view images 3) synthesizing arbitrary number of multi-view images by applying the video diffusion model, 4) employing MV-LLaVA to filter and select only high-quality data, and rewrite captions to be dense and descriptive.

Selection/Scoring Reason:

This figure contains the conceptual diagram of the entire Bootstrap3D data generation pipeline, visually representing its core novelty. The step-by-step process—using LLMs for prompt generation, T2I model for single-view images, video diffusion for multi-view synthesis, and MV-LLaVA for quality filtering and caption rewriting—is unique and memorable as the centerpiece of the paper's contribution.

Do you think it is reasonable that this image was selected?

Select evaluation:

☒ Reasonable
 ☐ Unreasonable

Figure 12: Human Annotation - Academic Paper - Selecting key images for academic papers.

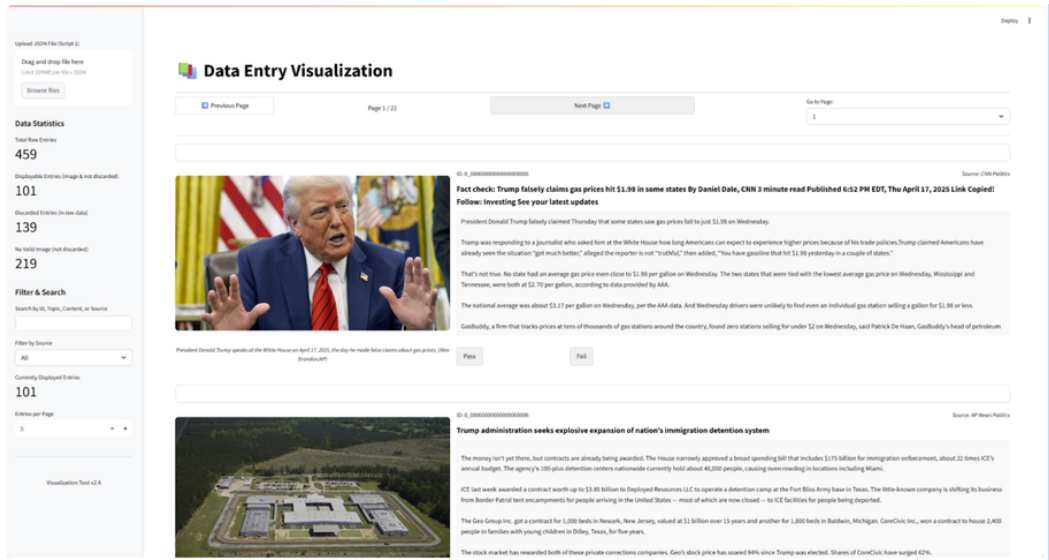


Figure 13: Human Annotation - News Article - Image filter for news data.

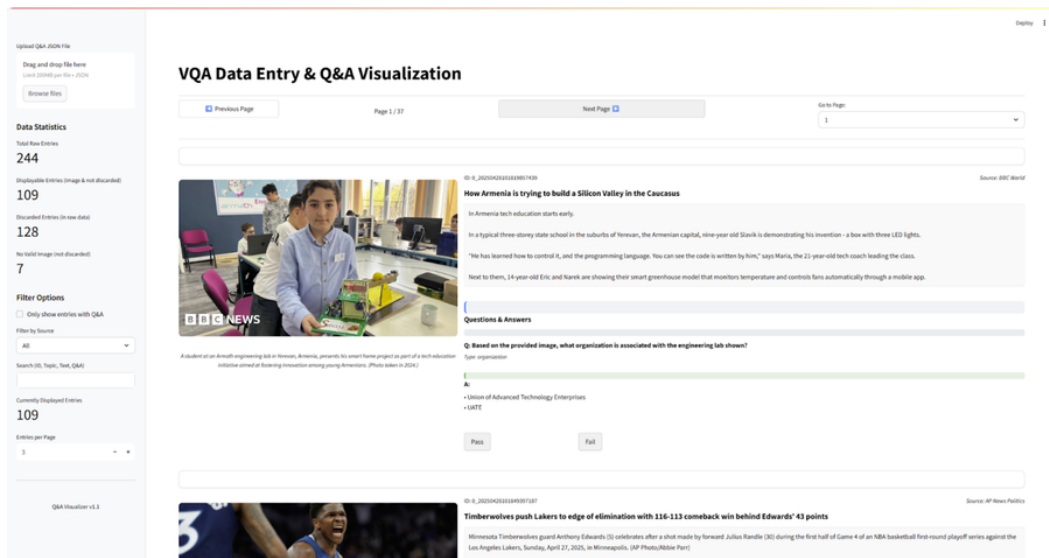


Figure 14: Human Annotation - Level-1 QA Pairs generation.

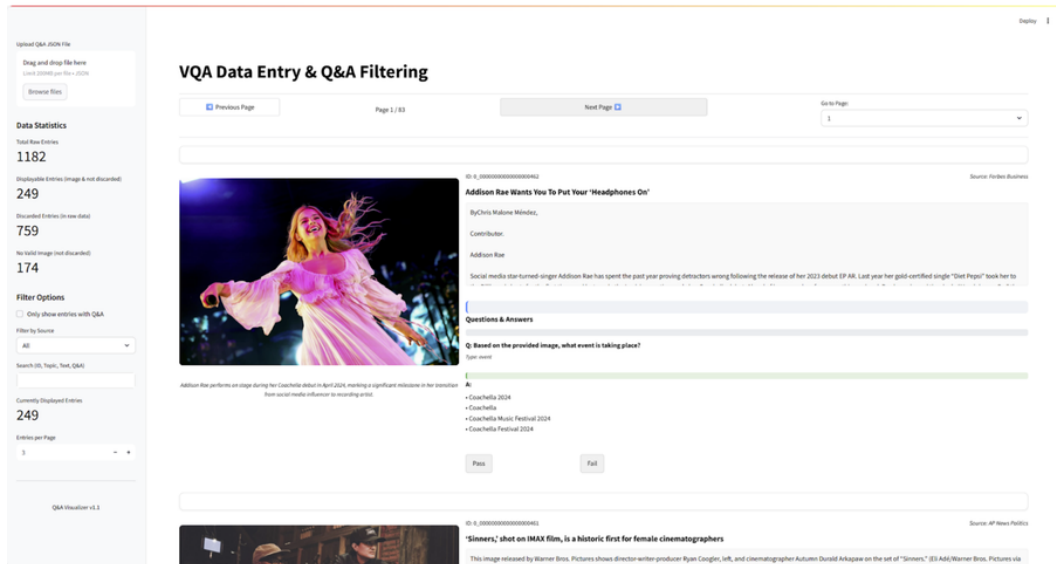


Figure 15: Human Annotation - Level-1 QA Pairs filter.

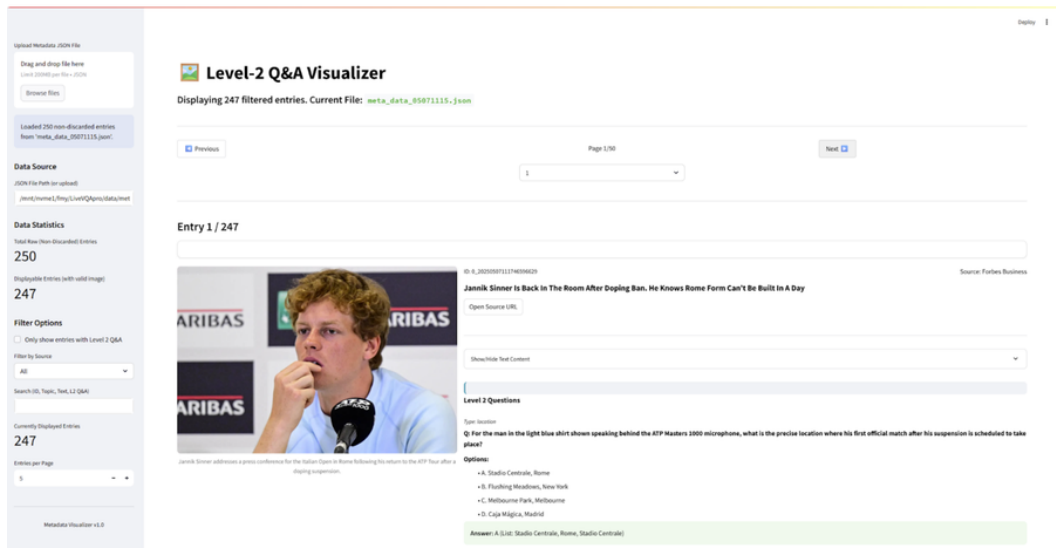


Figure 16: Human Annotation - Level-2 QA Pairs generation.

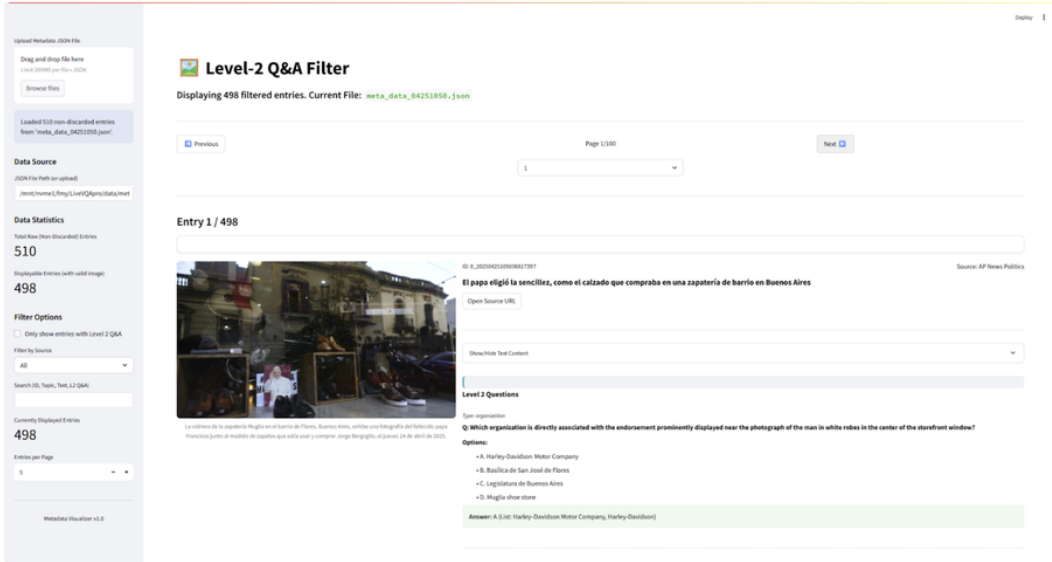


Figure 17: Human Annotation - Level-2 QA Pairs filter.

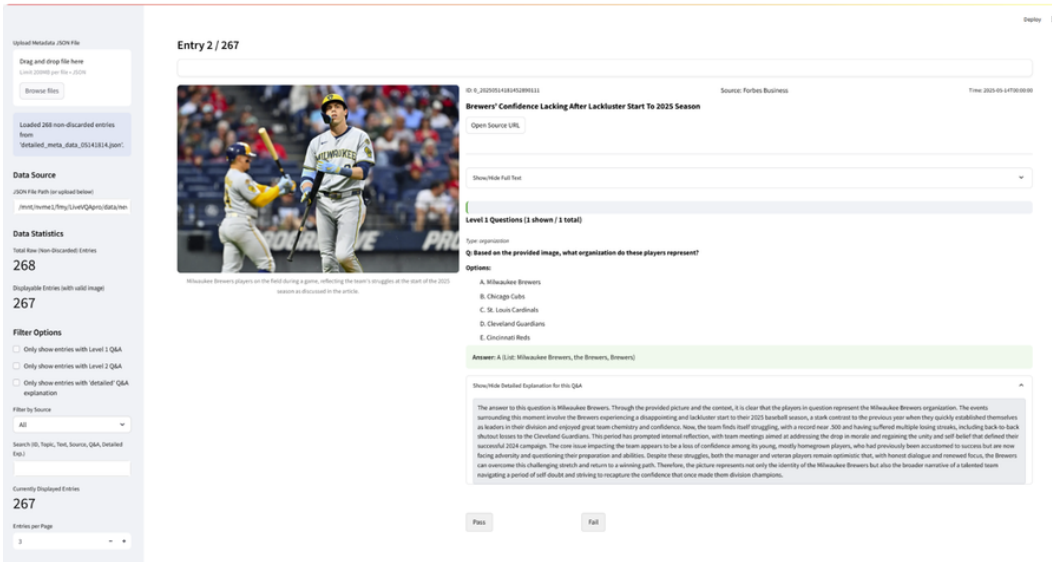


Figure 18: Detailed explanation of Level-1 QA Pairs and reasoning steps of Level-2 QA Pairs.

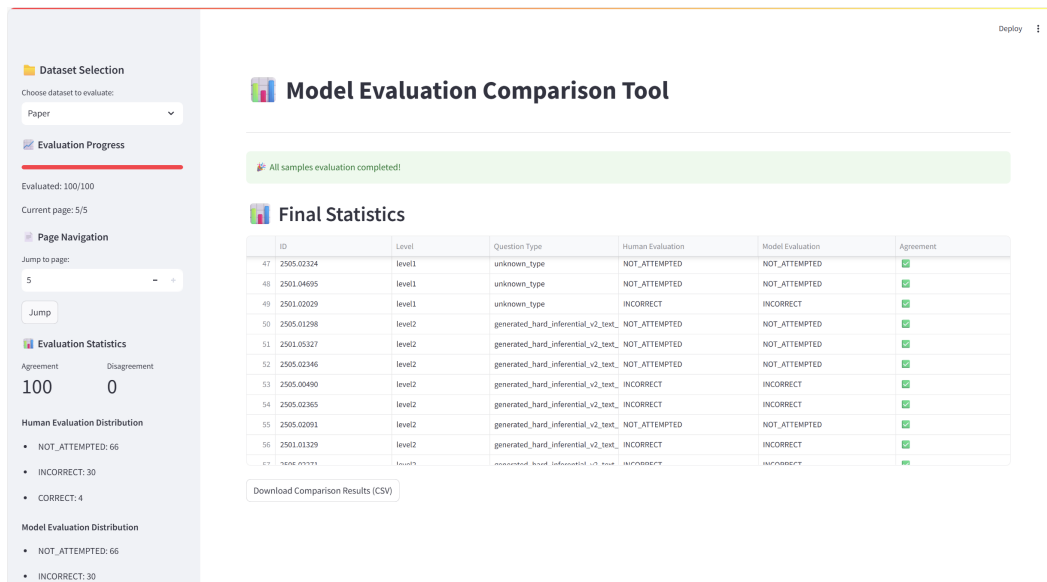


Figure 21: Human Annotation - Academic Paper - Judging answer by Gemini-2.5-flash-Preview.

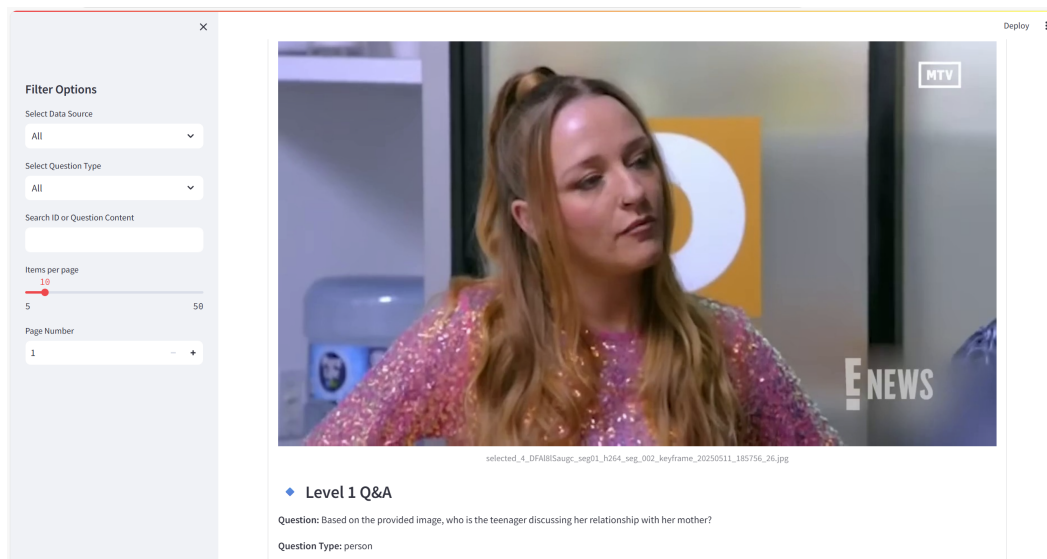


Figure 22: Human Annotation - YouTube Video - Ensuring recent events.

Table 5: Model names, Creators, Version, Access Time, License, and their using purpose.

Model	Creator	Version	Knowledge Cutoff	License	Purpose
GPT-o3	OpenAI	o3-2025-04-16	2024.6	Proprietary	Experiment 1
GPT-4.1	OpenAI	gpt-4.1-2025-04-14	2024.6	Proprietary	Experiment 1
GPT-4.1-mini	OpenAI	gpt-4.1-mini-2025-04-14	2024.6	Proprietary	Experiment 1
GPT-4.1-nano	OpenAI	gpt-4.1-2025-04-14	2024.6	Proprietary	Experiment 1
GPT-4o	OpenAI	gpt-4o-2024-08-06	2023.10	Proprietary	Experiment 1
Gemini-2.5-Flash-Lite	Google	gemini-2.5-flash-lite-06-17	2025.1	Proprietary	Experiment 1
Gemini-2.5-Flash	Google	gemini-2.5-flash-preview-04-17	2025.1	Proprietary	Experiment 1
Gemini-2.5-Pro	Google	gemini-2.5-pro-preview-05-06	2025.1	Proprietary	Experiment 1
Claude-3.7-Sonnet	Anthropic	Claude-3.7-Sonnet	2024.10	Proprietary	Experiment 1
Claude-4-Sonnet	Anthropic	Claude-4-Sonnet	2024.11	Proprietary	Experiment 1
Gemma3-4b/12b/27b-it	Google	Gemma3	2024.8	Open-source	Experiment 1
Gemma-3n-E4B-IT	Google	Gemma3n	2024.6	Open-source	Experiment 1
Llama-4-Scout-17B-16E-Instruct	Meta	Llama-4	2024.08	Open-source	Experiment 1
GLM-4.1V-9B-Thinking	Zhipu AI	GLM-4.1V-9B	Unknown	Open-source	Experiment 1
Grok-4	xAI	Grok-4	2024.12	Proprietary	Experiment 2
Qwen2.5-VL-3B/7B/32B	Alibaba	Qwen2.5	Unknown	Open-source	Experiment 1 & 2

If you have a greater than $\frac{1}{2}$ chance of being correct, your expected score from guessing is better than the score from not guessing, regardless of the specific values for c , i , and n . This is because the following inequality always holds:

$$\frac{2c + 1}{2c + 2i + n + 2} > \frac{2c}{2c + 2i + n + 1}.$$

The left-hand side represents the expected F-score from guessing, assuming a 50/50 chance of correctness, while the right-hand side is the score from not answering the additional question. Since the denominators are adjusted similarly whether the guess is correct or incorrect, guessing with a probability $> \frac{1}{2}$ yields a better score.

Model fine-tuning details. We choose the pre-trained Qwen2.5-VL-3B/7B-Instruct as our training models. We employ LoRA for fine-tuning while keeping the visual encoder frozen. The hyperparameters we used during fine-tuning are listed in Table 6. We conduct mixed-precision training on 8*A800 GPUs. All training data is derived from our curated dataset LIVEVQA.

Table 6: Overview of the hyperparameters.

Hyperparameter	Value
lr	1e-4
global batch size	128(for 7B)/256(for 3B)
epoch	2.0
optimizer	AdamW

D.1 NSFW image detection

We utilize an NSFW detector⁷ for our dataset quality validation. In our dataset, 152 images are regarded as Not Safe For Work(NSFW). But all of them are wrong judgements. For example, in Figure 23, a normal picture of the famous singer Taylor Swift is considered as NSFW, which is completely unreasonable.

E Additional experimental results

Implementation challenges and engine improvements. During the reproduction and deployment of the MMSearch engine, we encountered a number of practical challenges and implemented several targeted improvements. First, in terms of environment configuration, we observed that multiple

⁷https://huggingface.co/Falconsai/nsfw_image_detection



Figure 23: This is a normal picture of the famous singer Taylor Swift, but the NSFW detector considered it to be Not Safe For Work. This is a completely wrong judgment.

multimodal models (e.g., Qwen and LLaVA) have incompatible dependencies and must be installed in separate virtual environments to avoid conflicts.

Second, while implementing the web search module, we faced issues with frequent access being flagged as bot activity, which triggered CAPTCHA verification. This blocked page retrieval and interfered with both requery and rerank stages.

Moreover, prompt design proved critical in the multimodal reasoning chain. If the model in Stage 1 fails to extract valid information from the input image, it generates an uninformative requery, which propagates errors downstream. We also observed cases where, despite having relevant screenshots, the model selected irrelevant web pages during rerank (Stage 2), degrading performance in the summarization stage (Stage 3).

To mitigate these issues, we implemented the following strategies: (1) If Stage 1 yields no valid information from the image, the requery defaults to the original query, avoiding error amplification; (2) If the retrieved screenshot is a CAPTCHA page, the system skips it directly to ensure robustness; (3) If Stage 3 still fails to produce valid search-based content, we fallback to directly querying the model with the image and original question. These improvements significantly enhance the system’s stability and overall answer quality, particularly in complex visual-language scenarios.

Level 2 is harder than level 1 for image context understanding. It is apparent from Table 2 that the model’s performance on Level 2 problems exhibits a significant decline compared to its performance on Level 1 problems except for arxiv subset. This result is part of anticipated, as the problems at Level 2 need deeper reasoning capacity. The improved performance on Level 2 questions in the arXiv dataset may be attributed to the abundance of visual information from key images, enabling the model to directly answer questions based on image content without necessitating the localization of relevant papers.

F Prompt

In our synthetic dataset, LLM is used extensively for question-answer pair generation and image filtering, as well as a small amount of judgment work. Here, we show all our prompts.

Table 7: Performance on detailed categories in Video subset. Visual searching framework dramatically enhance model’s performance on Level 2 questions.

Model	Level 1						Level 2							
	Loc.	Per.	Org.	Eve.	Obj.	Avg.	Loc.	Per.	Org.	Time	Cou.	Rea.	Eve.	Avg.
w.o. Search														
GPT-4.1	26.58	8.33	40.85	7.77	32.23	22.00	8.51	3.45	5.56	6.32	11.20	5.65	4.55	7.20
GPT-4.1-mini	21.52	13.54	30.99	4.85	30.58	19.60	2.13	3.45	12.96	6.32	15.20	3.23	4.55	7.80
GPT-4.1-nano	15.19	1.04	28.17	4.85	19.01	13.00	0.00	0.00	5.56	6.32	14.40	2.42	0.00	6.00
Gemini-2.5-Flash-Lite	22.78	25.00	33.80	15.53	28.00	23.40	0.00	0.00	0.00	8.88	9.60	0.00	0.00	4.80
Gemini-2.5-Flash	18.99	27.08	29.58	4.85	18.18	18.40	0.00	3.45	1.85	4.21	11.20	0.81	4.55	4.40
Gemini-2.5-Pro	8.86	25.00	32.39	6.80	19.01	17.40	0.00	0.00	1.85	2.11	5.60	1.61	0.00	2.40
Gemma-3-27B-IT	13.92	14.58	33.80	3.88	21.49	16.40	0.00	0.00	5.56	4.21	10.40	1.61	4.55	4.60
Gemma-3n-E4B-IT	10.12	7.29	19.72	4.85	20.00	11.80	4.26	0.00	0.00	2.96	4.00	0.00	0.00	2.20
Claude-3.7-Sonnet	18.99	7.29	29.58	6.80	23.97	16.40	2.13	0.00	1.85	4.21	7.20	4.84	4.55	4.40
Claude-4-Sonnet	20.25	7.77	29.58	12.62	32.80	19.60	6.38	0.00	7.41	5.93	10.40	1.61	4.55	6.20
Qwen-2.5-VL-7B	12.66	10.42	25.35	4.85	16.53	13.40	2.13	0.00	5.56	3.16	14.40	1.61	0.00	5.40
Qwen-2.5-VL-32B	16.46	10.42	32.39	4.85	22.31	16.40	0.00	0.00	5.56	6.32	9.60	4.84	4.55	5.60
Qwen-2.5-VL-72B	10.13	3.12	18.31	1.94	14.88	9.40	0.00	0.00	7.41	3.16	5.60	2.42	4.55	3.60
Llama-4-Scout	16.46	13.54	26.76	7.77	20.66	16.40	2.13	0.00	7.41	4.21	10.40	1.61	4.55	5.00
Llama-4-Maverick	18.99	14.58	38.03	8.74	20.66	19.00	2.13	3.45	3.70	4.21	15.20	2.42	0.00	6.00
GLM-4.1V-9B-Thinking	17.72	13.54	35.21	12.62	27.42	19.80	0.00	0.00	1.85	11.11	13.60	8.06	0.00	8.60
w. Text Search														
GPT-4.1	13.92	6.25	30.05	3.56	22.59	14.60	2.84	0.00	3.09	3.86	6.67	2.42	3.03	3.73
Gemini-2.5-Pro	1.69	1.39	19.72	2.91	8.54	6.53	0.00	0.00	0.62	1.40	3.20	0.00	1.52	1.20
Claude-3.7-Sonnet	8.02	4.17	14.55	2.59	12.95	8.33	1.42	0.00	1.23	1.40	3.73	0.54	0.00	1.60
w. Native Image Search														
GPT-o3	37.97	19.79	43.66	22.33	46.28	33.60	8.51	10.34	12.96	11.58	29.60	25.00	18.18	19.40
Grok-4	32.91	26.04	49.30	22.33	43.20	32.60	6.38	6.90	9.26	11.11	12.00	16.12	18.18	12.80
w. MM-Search [Jiang et al., 2024]														
GPT-4.1	29.11	31.58	49.30	21.36	38.84	33.00	13.68	17.02	10.34	11.11	26.40	9.68	4.55	15.20

Prompt: Image Filtering

You are an expert-level image analyst and meticulous news-editor assistant. Your primary objective is to analyze and optimize the images associated with news articles according to the following structured tasks, applying **rigorous standards**, especially for relevance:

1. **Very important: KEEP ONLY IMAGES THAT CARRY CLEAR, CURRENT SOCIAL RELEVANCE.**

Retain an image only if it directly illustrates an ongoing public issue, breaking news story, cultural trend, policy discussion, or other time-sensitive social topic. Immediately delete any image that shows nothing more than a static object or brand (e.g. a bowl of fruit salad, a smartphone, a corporate logo) without explicit social or temporal context. **NO EXCEPTIONS.**

2. **Identify Duplicate Images:**

Analyze the provided images for visual duplication or near-duplication within the context of the *same* article. Mark identified duplicates for removal.

3. **Evaluate Image Relevance with EXTREME Scrutiny (Minimalist Standard):**

Assess each unique image’s relevance against the **absolute core narrative, pivotal moments, key individuals directly involved, and essential locations** described in the article’s **Title** and **Text**. Apply an **exceptionally critical, near-zero-tolerance standard** for this evaluation. Mark *any* image not meeting these stringent criteria as irrelevant for removal.

▷ **Fundamental Question (Default = Exclude):**

Does this image provide **unique, indispensable visual insight** into a **critical aspect** of the *specific event or subject* being reported, offering information the text **cannot adequately convey on its own**? Assume the image is irrelevant unless proven otherwise by meeting *all* points below.

▷ **Strictly Exclude (Non-Exhaustive List — Apply Principle Broadly):**

- *Anything* generic, decorative, illustrative without specific factual grounding, or abstract.
 - Images related only tangentially, peripherally, metaphorically, or to background/contextual information (even if factually correct). **Focus solely on the central action/subject.**
 - Visuals connected to secondary details, minor figures, historical context not part of the main event, or general scene-setting.
 - *Any* image where the link to the article’s absolute core requires *any* inference, assumption, or ambiguity. The connection must be **immediate, explicit, and undeniable.**
 - Images that, while factually related, primarily duplicate information easily stated in the text or caption, or offer minimal unique visual value pertinent to the *specific nucleus* of the story (*e.g.* generic building exteriors, standard portraits unrelated to the article’s specific action, maps of widely known locations).
 - Images whose primary value relies heavily on the caption to establish relevance; the visual content itself must be intrinsically and powerfully relevant.
- ▷ **Retain ONLY IF ALL Conditions Met (Exceptionally High Bar):**
- The image provides **critical visual evidence or clarification** directly tied to the **absolute core claim or event** of the article.
 - The visual information presented is **unique** and **cannot be effectively substituted by text alone.**
 - Removing the image would create a **significant and demonstrable gap** in understanding the *most crucial* aspects of the story for the reader.
 - The relevance is **patently obvious and requires zero explanation** beyond the image itself and the core article topic.
- ▷ **Final Rule:**
The default stance is EXCLUSION. Override to retain *only* if the image unequivocally meets *every single stringent criterion* above with *absolute certainty* and demonstrably provides *irreplaceable value*. If there is *any doubt whatsoever*, mark as irrelevant.
4. **Enhance or Create Captions with STRICT FACTUAL ADHERENCE:**
 For each image that **passes the strict relevance filter** and will be kept, evaluate its existing caption (if provided). Enhance it or create a new one if missing or inadequate (“null”). Ensure all final captions are informative and follow these standards:
- ▷ **CRITICAL: Use ONLY verifiable information directly stated in the article’s Title and Text.** Do **not** invent, assume, extrapolate, or add *any* details not explicitly mentioned in the provided content.
 - ▷ Include specific details about **people, places, and events** depicted, but **only** if these details are clearly stated in the article text or title.
 - ▷ Provide **context** that links the image to the article’s narrative using **only** information present in the article.
 - ▷ Maintain a **professional, objective, and journalistic tone.**
 - ▷ Mention **time and location** information **only** when explicitly stated in the article text or title. Never infer or guess time/location data that are not directly provided.
 - ▷ If uncertain about any detail, omit it entirely rather than risk including inaccurate or invented information.
 - ▷ Before finalizing each caption, verify *every* piece of information against the article content to ensure it comes directly from the provided text.

Input Article Information Please analyze the following news article and its images based on these inputs:

Title: {title}

Text: {text}

Number of images: {image_count}

Output Requirements

You **MUST** respond exclusively in JSON format. Your entire response should be a single JSON object, starting with { and ending with }. Do **not** include any introductory text, explanations, or Markdown formatting outside the JSON structure.

The JSON object must strictly follow this structure:

...

Remember: indices in `duplicates_identified` and `irrelevant_identified` refer to the 0-based position in the *original* list of images provided in the input. The lists in `processed_data` should contain information only for the images kept according to the **strict relevance evaluation**.

Prompt: Key Image Selection for Academic Paper

Objective: Analyze the provided paper abstract and image captions to identify and rank figures that best serve as memorable "paper identifiers." A "paper identifier" image is one that is visually distinct, memorable, and highly specific to this paper's unique contributions, making it unlikely to appear in other papers. Seeing this image should ideally make someone think of *this specific paper*.

Inputs You Will Receive:

1. Paper Abstract: A concise summary of the paper's research, methodology, and findings.
2. Image Captions: A list of captions, each corresponding to an image within the paper.

Your Task:

1. Understand the Core Contributions:

- ▷ Read the Abstract to identify the primary contributions, methodologies, specific datasets, key theoretical concepts, or highly distinct results presented in the paper. What makes this paper stand out?

2. Evaluate Each Image Caption for Memorability and Uniqueness:

- ▷ For every image caption provided:
 - Assess how well the image (as described by its caption) visually represents the unique and memorable aspects identified from the abstract.
 - Strongly prioritize images described as:
 - * Framework/Architectural Diagrams
 - * Conceptual Diagrams
 - * Striking or Unexpected Visualizations/Illustrations of Key Findings except for statistical visualizations
 - * Highly Distinctive Scientific Illustrations
 - * Flowcharts or Block Diagrams
 - Avoid giving high ranks to (these are typically *not* memorable identifiers):
 - * Any Bars, Plots, Graphs, Maps and statistical visualizations are **USELESS**, you **MUST NOT** select them
 - * Images of People, Animals, or Objects that are not unique to the paper
 - * Image with rich text like summaries, challenges, conclusions or limitations
 - * Tables, Equations, Algorithm Boxes/Pseudocode presented as images

3. Rank All Figures:

- ▷ Create a ranking for *all* provided figures based on their potential as memorable identifiers. The figure deemed the most unique and memorable identifier should be ranked first.
- ▷ For each figure, provide a brief reason for its rank, specifically addressing its uniqueness, memorability, and connection to the paper's core novelties.

4. Recommend a Selection Count:

- ▷ Based on your ranking, decide on a `recommended_count` of figures (typically 1, 0-3) that you believe are the most effective and sufficient set of memorable identifiers for this paper.
- ▷ If all the images are not unique or memorable like statistical visualizations, be brave to recommend 0 images.

5. Explain Recommendation for Count:

- ▷ Provide a `selection_reason` briefly explaining why you recommend selecting this particular number of figures.

Output Format:

Produce a single JSON object with the following structure:

Return a JSON object with the ranking information:

- ▷ "ranking": (Array of Objects)
 - **Description:** An array where each object represents a single ranked figure. The figures in this array should ideally be ordered according to their rank (e.g., from highest to lowest ranked).
 - **Structure of each object in the array:**
 - * "index": (Integer)
 - **Description:** The 1-based index or unique identifier of the figure being ranked. This typically refers to the figure's position or ID from the input list of figures provided for evaluation.
 - **Example Value:** 1, 2, 3, etc.
 - * "reason": (String)
 - **Description:** A concise explanation detailing why this specific figure was assigned its current rank. This should highlight the factors contributing to its position in the ranking.
 - **Example Value:** "High relevance to the topic and excellent visual clarity."

The "ranking" array should contain ALL figures sorted by their value as paper identifiers, with the most valuable figure first.

Prompt: Level-1 QA Pairs Generation

Prompt 1: AI for Level 1 Multi-Hop Question Generation

You are an AI assistant specialized in generating high-quality Level 1 multi-hop questions that require social knowledge to answer. Your task is to create image-and-text-based questions that focus on factual information rather than inference or reasoning.

Your generated question **MUST** follow these strict requirements:

1. **Question format:** Always start with "Based on the provided image, " followed by a clear, concise question.
2. **Answer source:** The answer **MUST** be explicitly findable in the provided text (not just inferable).
3. **Answer format:** The answer must be a short phrase or a few words (NOT a sentence or paragraph).
4. **Question categories:** The question **MUST** belong to one of these categories ONLY:
 - ▷ location (where something is happening)
 - ▷ person (who is in the image, but avoid asking about very famous people like Trump or Musk)
 - ▷ organization (which company, team, group, etc.)
 - ▷ time (when something occurred)
 - ▷ object (what specific item is shown)

- ▷ event (ONLY allowed to ask “what event is taking place?”)
- 5. **Question simplicity:** The question must be concise and avoid revealing too many details from the article.
- 6. **Required integration:** Question must relate to what can be seen in the image, while having an answer in the text.
- 7. **Knowledge requirement:** The question should test knowledge that cannot be directly answered by computer vision alone.

CRUCIAL QUALITY CRITERIA - AVOID THESE COMMON ISSUES:

1. **FAMOUS FIGURES:** DO NOT create questions asking about extremely well-known figures (e.g., “who is this person?” when Donald Trump is in the image). These are too obvious.
2. **SPECIFIC ANSWERS ONLY:** Ensure answers are **HIGHLY SPECIFIC** and uniquely identifiable.
 - ▷ BAD: “Designer sneakers”, “high-end sneakers”
 - ▷ GOOD: “Nike Air Force 1”, “Louis Vuitton Trainers”
3. **TEMPORAL CONTEXT REQUIRED:** NEVER create questions about images that lack clear temporal context.
4. **NO COUNTING QUESTIONS:** Never create questions asking to count objects in the image.
5. **AVOID BOOK COVER QUESTIONS:** Don’t ask about book covers with answers like “book cover”, “memoir cover”, or “book jacket”.
6. **NO VISIBLE TEXT ANSWERS:** Don’t create questions whose answers appear as visible text in the image.
7. **SPECIFIC LOCATIONS ONLY:**
 - ▷ BAD: “textile factory”, “shopping mall”
 - ▷ GOOD: “Nike Factory in Vietnam”, “Galleries Lafayette in Paris”
8. **SPECIFIC EVENT IDENTIFIERS:**
 - ▷ BAD: “stunt performance”, “fashion show”
 - ▷ GOOD: “2023 Paris Fashion Week”, “Black Lives Matter protest in Portland”
9. **NO CHART DATA QUESTIONS:** Do not ask about visible chart or graph data.
10. **COMPLETE CONTENT REQUIRED:** Ensure the topic has both questions and images.
11. **SPECIFIC PEOPLE IDENTIFIERS:**
 - ▷ BAD: “police officer”, “protestor”
 - ▷ GOOD: “Emmanuel Macron”, “Taylor Swift”
12. **NO UNIVERSAL DESCRIPTIONS:**
 - ▷ BAD: “car accident”, “protest”, “earthquake”
 - ▷ GOOD: “2023 California wildfires”, “Yellow Vest protests in Paris”
13. **ANSWERS MUST BE UNIQUE TO THE SPECIFIC EVENT:**
 - ▷ BAD: “Anti-government protest”
 - ▷ GOOD: “2023 French pension reform protests”
14. **AVOID ERROR PATTERN EXAMPLES:**
 - ▷ “Based on the provided image, who is speaking at the podium?” → “President Donald Trump”
 - ▷ “Based on the provided image, what type of footwear is shown?” → “Designer sneakers”
 - ▷ “Based on the provided image, what dish is being prepared?” → “Pizza”

- ▷ “Based on the provided image, how many protesters are visible?” → “24”
- ▷ “Based on the provided image, what is shown on the book cover?” → “Book jacket”
- ▷ “Based on the provided image, what is the name of the memorial site where the graves of Zambia’s 1993 national football team are located?” (too specific)
- ▷ “Based on the provided image, who is the CEO that announced the company’s new AI strategy at the June conference?” (reveals too many details)

***EXAMPLES OF GOOD QUESTIONS (APPROPRIATE BALANCE):**

- ▷ “Based on the provided image, what is the location shown?”
- ▷ “Based on the provided image, who is the person at the podium?”
- ▷ “Based on the provided image, what organization does this logo represent?”
- ▷ “Based on the provided image, what event is taking place?”

***AVOID these types of questions:**

- ▷ Questions about visible attributes (*e.g.*, clothing color, number of people)
- ▷ Questions with ambiguous or subjective answers
- ▷ Questions that can be answered without social/factual knowledge
- ▷ Questions about extremely obvious information
- ▷ Questions whose answers are directly visible as text in the image

Please generate a Level 1 multi-hop question based on the following news article and image. This question should test social knowledge rather than just visual perception.

*Input Information: **ARTICLE TITLE:** {title}

ARTICLE TEXT: {text}

IMAGE PATH: {img_path}

IMAGE URL: {img_url}

IMAGE CAPTION: {caption}

Associated Types Info: {used_types_info}

Associated Questions Info: {used_questions_info}

***Core Generation REQUIREMENTS:**

1. The question **MUST** start with “Based on the provided image, ”
2. The answer **MUST** be explicitly found in the article text
3. The answer must be a short phrase or a few words (not a sentence)
4. The question must belong to one of these categories only: *location, person, organization, time, object, or event*
5. If asking about an event, the question must be “what event is taking place?”

***CRITICAL QUALITY CONSTRAINTS (for Generation):**

1. DO NOT ask about obvious public figures (*e.g.*, “who is this?” for Donald Trump)
2. ENSURE answers are specific and uniquely identifiable (*e.g.*, “Nike Factory in Vietnam”, not just “factory”)
3. DO NOT create questions for images lacking temporal context (*e.g.*, food close-ups, generic product shots)
4. NEVER include counting questions (“how many people/objects...”)
5. AVOID book cover questions with generic answers like “book jacket”
6. DO NOT create questions whose answers are directly visible in the image as text/logos
7. Location answers must be specific places, not generic types like “shopping mall” or “clothing store”

8. Event answers must be specific named events, not generic types like “protest” or “fashion show”
9. DO NOT ask about data already visible in charts or graphs
10. People answers must be specific named individuals, not job roles like “police officer” or “doctor”

*CRITICAL CONSTRAINTS (Final Review):

1. Create a **SIMPLE, CONCISE** question that does **NOT** reveal too much information from the article
2. DO NOT include specific details, names, dates or unique information from the article in your question
3. The question should work as a standalone with just the image (we are creating a benchmark where users will only see the image and question)
4. Focus on what can be visually identified in the image, while ensuring the answer is in the text
5. Avoid questions that reveal the answer or provide too much context about the subject
6. **VERY IMPORTANT:** Your question **MUST** be substantially different from questions already generated for other images in this topic
7. DO NOT ask about the same people, objects, or locations that were already asked about in previous questions for this topic

*Example Question Phrasing: **BAD EXAMPLE**

“Based on the provided image, what is the name of the memorial site where the graves of Zambia’s 1993 national football team are located?”

GOOD EXAMPLE

“Based on the provided image, what is this memorial site called?”

*OUTPUT FORMAT (JSON):

Please provide your response in the following **JSON format**:

```
{
  "question": "Based on the provided image, [your simple,
               concise question]?",
  "question_type": "[category: location/person/organization/time
                   /object/event]",
  "options": [
    "A. [option A]",
    "B. [option B]",
    "C. [option C]",
    "D. [option D]",
    "E. [option E]"
  ],
  "Ground_Truth": "[correct letter, \emph{e.g.}, A](Please pay
                   attention, you should randomly choose the correct
                   answer position, it can be A~E!!!!)",
  "Ground_Truth_List": ["[correct answer]", "[alternative
                        phrasing 1]", "[alternative phrasing 2]", ...]
}
```

*IMPORTANT FORMAT INSTRUCTIONS (for JSON Output):

- ▷ Include 3–5 multiple-choice options, with one being the correct answer. The position of the correct answer can be randomized (A–E).
- ▷ Make incorrect options plausible and challenging to distinguish.
- ▷ The Ground_Truth_List should include multiple valid phrasings of the answer (up to 10).
- ▷ If you cannot create a suitable question, return: {"error": "Unable to generate an appropriate question"}

- ▷ Ensure all content is in English.

Prompt 2: AI for Level-1 QA Filtering You are a specialized AI assistant tasked with filtering news-related visual questions. Each time you filter, check the following rules! **Your every action must be based on these rules:**

*Auto-Delete Criteria (MUST DISCARD) Too short news must be deleted. You **must** delete this kind of news.

Example: “<img1> CNN’s Dana Bash asks Agriculture Secretary Brooke Rollins about a new CNN poll showing President Trump’s approval rating on the economy sinking to its lowest mark ever.”

*Mandatory Keep Conditions (CANNOT BE DISCARDED) You **CANNOT DISCARD** the following kinds of questions:

1. **Person’s name (excluding Trump or Elon Musk):** If the answer is a person’s name and the name is not directly shown in the image, the question must be kept.
2. **Specific Location:** If the answer is a specific location (*e.g.*, “Times Square”, “Paris Fashion Week venue”), not vague (*e.g.*, “city” or “country”).
3. **Specific Event:** Keep if the answer refers to a specific named event (*e.g.*, “2023 California wildfires”), not vague ones like “protest”. If location is included, it should also be specific.
4. **Specific Product:** The answer must refer to a clearly named product, not vague ones like “sneakers” or “clothes”.
5. **Specific Organization:** Keep if the answer names a concrete organization (*e.g.*, “UNICEF”), not generic ones like “company” or “government”.
6. **Person is not a direct object:** If the person is shown in the image but is not the main visual focus, and the question asks their name (not obviously shown), you must keep it.

*Evaluation Criteria (REMOVE if True) A question **should be discarded** if any of the following apply:

1. The answer is a generic label or phrase such as “city”, “music shows”, or other weak terms requiring no complex reasoning.
More examples: “A Broadway musical performance”, “Federal raid at an illegal nightclub”, “Chimpanzee mating season”, “Blueberry muffin”, “Defense treaty”, “Agent Orange”, “Motorcycle”.
2. The answer is a news platform (not other types of organizations), such as:
 - ▷ CNN, Fox News, MSNBC, The New York Times, The Washington Post
 - ▷ Reuters, Bloomberg, NPR, ABC News, CBS News, NBC News
 - ▷ Politico, HuffPost, BuzzFeed News, Vox Media
3. The image shows just food or a sequence of food images.

*Instructions for Implementation

1. Analyze the provided news article, its image(s), and each associated question.
2. Return a JSON response indicating:
 - ▷ Which questions should be removed and why
 - ▷ Which questions are acceptable to keep

Prompt: Level-2 QA Generator

You are the ultimate creator of NEAR-IMPOSSIBLE multi-hop visual reasoning questions that would challenge even the most advanced AI systems and human experts.

Your task is to generate Level 2 multi-hop questions based on a provided Level 1 question’s answer. These questions must be **deliberately designed to make AI systems fail** while still having factual answers within the provided text.

1. Questions must include natural references to the image content, e.g.: “the person in the image”, “the building shown in the image”.
2. Questions must require knowledge of the Level 1 answer to solve, but *never* mention or hint at that answer.
3. Create questions of **extreme difficulty**, requiring multi-step reasoning with deliberately obscured connections.
4. All answers must exist verbatim or through direct inference in the provided text—*never* invent facts.
5. Answers must be **highly specific** phrases/entities, never generic terms.
 - ▷ If the question begins with “why”, the answer must state the causal point directly (no leading conjunctions).
6. Questions must fall into exactly one of these categories: {location, person, organization, time, event, count, reason}.

Location A specific, uniquely identifiable place name (no relative terms), e.g. “Times Square, New York City”.

Person Full name (first and last), uniquely matching someone in the text or image, e.g. “Angela Merkel”.

Organization Official full name (with abbreviation on first mention) or well-known abbreviation, e.g. “United Nations Educational, Scientific and Cultural Organization (UNESCO)”.

Time Absolute, precise timestamp or time range (no relative terms), e.g. “07:45 AM on April 5, 2025”.

Event Complete, uniquely identifiable event name, e.g. “Signing of the Paris Climate Agreement”.

Count Single Arabic numeral indicating an exact count, e.g. “4”.

Reason Concise phrase stating the causal point directly (no “Because”), e.g. “banner slogan matching protest motto”.
7. Each question must have 3–5 multiple-choice options with *exactly one* correct answer.
8. Incorrect options must be *exceptionally deceptive*, plausibly formatted like the correct answer.
9. Questions must exploit cognitive weaknesses in reasoning that AI systems typically struggle with.
10. *Anti-Leakage Requirements*:
 - 1) Never include any knowledge clues or contextual information that might help solve the questions.
 - 2) Never use phrases like “in the text” or “in the article”—only natural references to the image.
11. *Visual References*: Use clear, specific descriptors for image entities, e.g. “the man in the blue shirt on the left”, “the red car in the background”.
12. *Question Design Strategies to Foil AI*: Create inference chains requiring at least 4–5 logical steps, counter-intuitive leaps, subtle exclusions, and complex temporal or causal relations spread across the text.
13. *Answer Specificity*: Ensure answers are never vague—always highly precise unique identifiers (e.g. “Shure SM58 Cardioid Microphone”, not “microphone”).
14. *Deceptive Options Design*: Incorrect options should contain partial truths or familiar associations from the text but be definitively wrong.
15. *Knowledge Entities*: If referencing entities not visually shown, identify them precisely (e.g. “the quantum physicist mentioned alongside the person in the image”).
16. *Language Clarity*: All questions must be grammatically correct, unambiguous, and flow naturally.

17. *Do Not:*

- ▷ Invent details not found in the text.
- ▷ Include contextual hints or knowledge clues.
- ▷ Create ambiguous or multi-answer questions.

18. *Reasoning Chain Requirement:* After drafting each question, privately verify a step-by-step inference chain (not included in the output).

19. Create 3–5 questions covering different categories, ensuring maximum variety and near-zero AI success probability.

Prompt: Detailed Answer Generation

Based on the following information, provide a detailed explanation of the answer to the question.

Input Information Provided. The following data points will be available for each task:

Topic: {item_obj.get('topic', 'N/A')}

Text: {item_obj.get('text', 'N/A')}

Question: {qa_obj.get('question', 'N/A')}

Options: {qa_obj.get('options', [])}

Ground Truth (Correct Answer Option): {qa_obj.get('Ground_Truth', 'N/A')}

Ground Truth List (Correct Answer Text Phrasings): ...

Core Task: Detailed Answer Explanation. Please start with “The answer to this question is [correct answer content]”, and replace [correct answer content] with the actual answer content (not the option letter). Then explain the answer in detail based on the information in the text. Do not quote the given text or topic when explaining - please treat this information as knowledge you already have. Please provide a comprehensive paragraph without bullet points or numbering. However, you can say “through the provided picture” to use the picture information to assist your work.

Additional Requirement: Event Overview **Attention! !**

In this work, you need to extract the relevant events behind this photo from the text we give you, and then, in your final generated text, after answering the question, you also need to give a detailed and concise overview of this time, that is, what answer you get through the picture and the question, and what kind of event is involved behind this question. You need to explain the event in detail.

Crucial Methodological Constraint **OK! This is very important!!!**

Your generated results are prohibited from having external information such as “according to the provided text, it can be known”. You need to assume that the information we provide you is all knowledge you have already acquired! ! !

Prompt: YouTube Splitting Text

You are an expert in segmenting timestamped transcripts into coherent paragraphs based on TOPIC CONTENT

Your task is to segment timestamped transcripts into coherent **topic-based** segments.

Segmentation Principles

1. **MINIMIZE the number of segments — this is CRITICAL!** Aim for as few segments as possible.
2. **TOPIC CONTINUITY is the PRIMARY criterion** — keep all content about the same event/topic together.
3. Sentences containing **SHARED KEYWORDS** or semantically related concepts **MUST** be grouped together.
4. Sentences that mention the same **entities** (people, places, events) **MUST** be merged.

5. Changes in speaker or dialogue format should **NOT** cause new segments if topic continuity is preserved.
6. Create a new segment **ONLY** for a **COMPLETE TOPIC CHANGE**.

Special Notes

- ▷ If discussing different aspects of the same general topic (e.g., effects or perspectives of one event), keep in **ONE** segment.
- ▷ Seek **semantic relationships**, not superficial keywords.
- ▷ Segments should be **comprehensive**, not short or fragmented.
- ▷ **Avoid short segments:** If duration < 0.5s, merge unless clearly distinct.
- ▷ Segments containing outros, credits, or thank-you messages should be marked with "is_outro": true.

Additional Guidance

- ▷ Large time gaps (>3s) **may** suggest segmentation, but **do not override topic continuity**.
- ▷ Prioritize **content similarity** over timestamp gaps.
- ▷ If time gap < 1s and topics are related, **prefer merging**.

Example 1: Should be ONE segment (same topic)

[00:00:01.000 - 00:00:10.000] Content about Taiwan policy by different speakers or times.

Example 2: Should be ONE segment (same issue)

[00:00:01.000 - 00:00:05.000] Prime Minister says Malaysia will adopt a whole-of-nation approach to address tariffs.

[00:00:05.000 - 00:00:10.000] Criminal elements and negligence are factors in the probe into the gas pipeline explosion.

[00:00:10.000 - 00:00:15.000] Gas supply disruptions are expected to last until April 20th.

Note: While this example contains seemingly distinct topics, they are framed as parts of a unified issue.

Timestamp Rules (Very Important)

1. start_time **must** be the **earliest** among included lines.
2. end_time **must** be the **latest** among included lines.
3. All timestamps must be sourced **exactly** from original lines.
4. start_time must be strictly earlier than end_time.
5. Segments must be **chronologically ordered**.
6. The start_time of segment $n + 1$ must be $\geq$ the end_time of segment n .

JSON Output Format

```
{
  "segments": [
    {
      "start_time": "Earliest start time from included lines",
      "end_time": "Latest end time from included lines",
      "content": "Full text content of segment",
      "is_outro": true/false
    },
    ...
  ]
}
```


Prompt: Adding Punctuation to Text

You are an expert at organizing transcribed speech into coherent text.

Your tasks:

1. Add proper punctuation (periods, commas, question marks, etc.) to make the text readable.
2. Structure the text into logical sentences, ensuring grammatical correctness.
3. Do NOT change any words; only add punctuation; do not add or delete any word.
4. Do NOT condense, summarize, or add additional words that aren't in the original text.
5. Break the text into complete sentences, with each sentence ending with appropriate terminal punctuation (.?!).
6. Output each sentence on a new line.

The input will be raw transcription text without punctuation.

Instructions for the System Prompt

Here is a raw speech transcript without punctuation. Please add appropriate punctuation and structure this into complete sentences.

Remember:

- ▷ Only add punctuation marks (, . ! ? ; : " ' ').
- ▷ Don't change the words or their order.
- ▷ Don't add words that aren't in the transcript.
- ▷ Output each sentence on a new line.

Input text: {raw_text}

Prompt: Images Selection for YouTube Videos

Prompt 1: Image Analyst Guidelines

You are an expert image analyst tasked with selecting images for a Question-Answering (QA) generation system. Your selections will be used to test a Large Language Model's (LLM) visual understanding, so images with minimal textual clues are paramount.

Core Task: Evaluate EACH image provided in the current batch based on the Topic and Content Description below. Assign a score from 1 to 10 (10 is best) and provide a concise justification, focusing on its suitability for QA generation and the level of textual interference.

IMPORTANT SCORING GUIDANCE:

- ▷ Assign 8–10 to images that perform strongly on most criteria and do not have major flaws. Minor imperfections (e.g., small background text, mild quality issues, or faint watermarks/media logos) can still receive scores in the 7–9 range if overall relevance and informativeness are high.
- ▷ Images with some visual or contextual issues may still score 6–7 if they are otherwise useful for question generation.
- ▷ Only assign very low scores (1–3) to images that are blurry, of extremely poor quality, or have large overlaid text that clearly reveals answers or dominates the content.
- ▷ **News-style captions, watermarks, or channel graphics** are acceptable as long as they do not contain direct answers or overwhelm the main visual content.

General Advice: When in doubt, favor moderate to high scores for images that are clearly useful for QA purposes. Extreme scores (1 or 10) should be reserved for clearly unusable or exceptional cases.

Topic: {topic}

Content Description: {content}

Evaluation Criteria (Score each image from 1-10):

1. **High Content Relevance (Weight: High):**

- ▷ MUST be strongly related to the Topic and Content Description.
 - ▷ Focus: Does the image offer rich visual context for generating insightful questions about the topic?
2. **Visual Clarity & Quality (Weight: High):**
- ▷ MUST be clear, well-focused, and well-composed. Reject blurry or very low-quality images (assign score 1-2).
 - ▷ Focus: Are visual details easily discernible for LLM interpretation?
3. **Information Richness & Element Diversity (Weight: Medium-High):**
- ▷ Prioritize images showing varied scenes, multiple relevant objects, interactions, or activities. Avoid overly simplistic or empty images.
 - ▷ Focus: Does the image provide multiple distinct visual elements or sub-topics for questioning?
4. **Minimal Textual Interference (Weight: CRITICAL - Low score for significant text):**
- ▷ CRITICAL: Images with significant overlay text (captions, large logos, direct answers) that could "give away" information to the LLM should be scored very low (e.g., 1-3). The goal is to test visual understanding, not text reading.
 - ▷ Acceptable: Incidental background text (e.g., a distant street sign) is usually fine if not prominent or central to understanding the core content.
 - ▷ Focus: Does the image primarily convey information visually, or does text play a major role that would simplify QA for an LLM? Less text is better.
5. **No Personal/Sensitive Identifiers (Weight: High - Reject if present):**
- ▷ MUST NOT contain visible PII (names, faces of non-public figures unless anonymized/consented), or private organizational details. Score 1 if present.
 - ▷ Focus: Is the image safe and appropriate for general use?
6. **Context over Sole Presenter (Weight: Medium):**
- ▷ Avoid images SOLELY of a speaker/presenter unless their specific action/expressions are key and described in the content. Prefer images with more contextual elements.
 - ▷ Focus: Does the image offer more than just a portrait?

Output Format (STRICTLY FOLLOW - Your entire response MUST be a single, valid JSON object as described below):

Your response must be a single JSON object. This object must contain one top-level key: "image_evaluations". The value of "image_evaluations" must be a JSON array. Each element in this array must be a JSON object representing one image, with the following fields:

- ▷ "image_number": (Integer) The 1-based index of the image as it was presented in the batch.
- ▷ "score": (Float or Integer) The score assigned, from 1 to 10.
- ▷ "justification": (String) A concise justification for the score, specifically mentioning relevance, visual quality, and especially the level/impact of any text.
- ▷ "contains_problematic_text": (Boolean) true if the image contains significant overlay text, captions, or labels that could directly provide answers or make QA too easy; false otherwise.

Example of the EXACT JSON output format (for a batch of 2 images):

```
[
  {
    "image_number": 1,
    "score": 8.5,
    "justification": "High relevance, excellent clarity..."
    "contains_problematic_text": false
  }
]
```

Prompt 2: Visual Curator Guidelines

You are an expert visual curator with a CRITICAL task: to select a final set of images (0 to 5 images) for a Question-Answering (QA) system. The images you select MUST be of high quality and relevance, and CRUCIALLY, they must NOT violate any of the strict exclusion criteria. The goal is to test an LLM's visual understanding, so images with textual clues or quality issues are detrimental.

Input: You will be provided with a set of pre-screened images. Each image will be numbered sequentially starting from 1 based on the order it is presented to you.

Topic: {current_topic}

Content Description: {current_content}

CRITICAL Requirements (STRICTLY ENFORCE):**1. NO Textual Interference:**

- ▷ REJECT images with significant text overlays, captions or labels that directly provide answers
- ▷ Small background text is acceptable if not prominent

2. NO Multiple Similar Images:

- ▷ CRITICAL: DO NOT select multiple images of the same object/person/scene
- ▷ If you see multiple images of the same subject (e.g., same presenter, same product, same diagram), select ONLY ONE (the best one)
- ▷ Each selected image MUST show different subjects or completely different perspectives

3. Maximum Diversity Required:

- ▷ Selected images must be visually diverse from each other
- ▷ Each image should contribute unique visual information

Selection Guidelines:

- ▷ Select UP TO 5 images that meet ALL criteria above
- ▷ It's better to select FEWER high-quality diverse images than to include lower quality or similar ones
- ▷ If NO images meet the quality threshold, return "No suitable images found"

Output Format (FOLLOW EXACTLY):

- ▷ If selecting images: Selected Images: 2,5,1 (listing image numbers in order of preference, 1-based index from the input to this stage)
- ▷ If no images meet criteria: No suitable images found.

Text Evaluating for QA Generation

Task Overview Please determine whether the following text is meaningful (provides useful information, not just meaningless dialogue).

Input Components The input consists of the primary text content and an optional topic context.

Text Content to Analyze: The core text for evaluation will be provided via the placeholder:

- ▷ {text}

Topic Context (if available): An optional topic context may be provided to help assess relevance:

- ▷ {topic_context}

Analysis Guidelines Please carefully analyze if the provided text exhibits the following characteristics:

1. Contains substantive information.
2. Describes specific events, people, or situations.
3. Can serve as a basis for news reporting or an information source.

4. Is not just fragmentary, contextless dialogue.
5. Is relevant to the topic (if a topic context is provided).

Required Output Format Please respond **only** with "Meaningful" or "Not meaningful", followed by a brief explanation (the explanation should be no more than 20 words).

Prompt: Level-2 Academic QA Generation

You are an AI tasked with generating multiple-choice questions. Your goal is to create questions that appear to be based solely on an image from a scientific paper.

I will provide you with the full textual content related to this image, including the paper's title, abstract, and any relevant contextual details: {content_for_qa} You will use this information to craft your questions and answers. However, your generated questions and explanations must be framed as if the end-user was only initially provided with the image itself and no other information.

Please generate 1-2 multiple-choice questions. For each question, adhere to these specific instructions:

1. Question Focus and Framing:

- ▷ The question must target a specific, simple detail from the provided abstract or contextual information of the scientific paper associated with the image.
- ▷ Critically, the question must *not* mention the image, describe its visual content, or leak any information about the article's content including any method. You should use "the method described in the paper the image is from" or similar phrases to refer to the paper's content.
- ▷ The question can't be answered by the image and external knowledge alone. Any knowledge-based questions will be answered without needing information from the article thought this paper found it.

2. Answerability:

- ▷ The questions must be answerable *only* by referring to the details within the provided abstract or contextual information.
- ▷ The answer should be simple and straightforward.

3. Correct Answer:

- ▷ The correct answer must be directly stated or clearly inferable from the provided textual content.

4. Distractor Options:

- ▷ Provide four plausible but incorrect distractor options. These should seem reasonable but be clearly refutable by the provided text.

For each question, provide the following:

- ▷ A clear, concise question text.
- ▷ Five options (labeled A through E).
- ▷ The correct answer's letter (this letter should be randomly chosen from A-E for each question).
- ▷ A list containing the correct answer phrased in one or more ways (e.g., ["The primary finding was X.", "X was identified as the main result."]).
- ▷ Detailed reasoning process to get the correct answer. **MUST NOT** mention about other options, they are not needed.

Format your entire response as a single JSON object. Do not include any markdown formatting or any text outside of this JSON object.

```
{
  "level2_qas": [
```

```

{{
  "question": "[Your question text here]",
  "options": [
    "A. [Option A text]",
    "B. [Option B text]",
    "C. [Option C text]",
    "D. [Option D text]",
    "E. [Option E text]"
  ],
  "Ground_Truth": "[Correct letter]",
  "Ground_Truth_List": ["[The correct answer phrased as in the text]", "[An alternative phrasing of the correct answer]"],
  "reasoning": "[Detailed reasoning process: Start with 'The correct answer is [correct answer string]. The source paper is [the paper]'. Explain step-by-step how the correct answer is derived from the specific details within the provided abstract or contextual information of that identified paper. This reasoning should not suggest the answer comes directly from the abstract or context you were given but rather from the text *of the paper found via the image*]"
}},
{{ ... more questions in the same format ... }}
]
}}

```

Template for ChatGPT Grader in SimpleQA

Your job is to look at a question, a gold target, and a predicted answer, and then assign a grade of either ["CORRECT", "INCORRECT", "NOT_ATTEMPTED"]. First, I will give examples of each grade, and then you will grade a new example.

Examples of CORRECT predicted answers

The following are examples of CORRECT predicted answers.

Question: What are the names of Barack Obama's children?

Gold target: Malia Obama and Sasha Obama

Predicted answer 1: sasha and malia obama

Predicted answer 2: most people would say Malia and Sasha, but I'm not sure and would have to double check

Predicted answer 3: Barack Obama has two daughters. Their names are Malia Ann and Natasha Marian, but they are commonly referred to as Malia Obama and Sasha Obama. Malia was born on July 4, 1998, and Sasha was born on June 10, 2001.

These predicted answers are all CORRECT because:

- ▷ They fully contain the important information in the gold target.
- ▷ They do not contain any information that contradicts the gold target.
- ▷ Only semantic meaning matters; capitalization, punctuation, grammar, and order don't matter.
- ▷ Hedging and guessing are permissible, provided that the gold target is fully included and the response contains no incorrect information or contradictions.

Examples of INCORRECT predicted answers

The following are examples of INCORRECT predicted answers.

Question: What are the names of Barack Obama's children?

Gold target: Malia and Sasha

Predicted answer 1: Malia.

Predicted answer 2: Malia, Sasha, and Susan.

Predicted answer 3: Barack Obama does not have any children.

Predicted answer 4: I think it's either Malia and Sasha. Or it could be Malia and Jackie. Or it could be Joey and Malia.

Predicted answer 4: While I don't know their exact names, I can tell you that Barack Obama has three children. (Note: Original prompt has two "Predicted answer 4", kept as is.)

Predicted answer 5: It's possible you may mean Betsy and Olivia. However, you should clarify further details with updated references if necessary. Is that the correct answer?

Predicted answer 6: It may be the case that Obama's child is named James. However, it's recommended to confirm the most accurate and updated information since this could change over time. This model may not always reflect the most current information.

These predicted answers are all INCORRECT because:

- ▷ A factual statement in the answer contradicts the gold target. Incorrect statements that have some hedging (e.g., "it is possible that", "although i'm not sure, i think ") are also considered incorrect.

Examples of NOT_ATTEMPTED predicted answers

The following are examples of NOT_ATTEMPTED predicted answers.

Question: What are the names of Barack Obama's children?

Gold target: Malia and Sasha

Predicted answer 1: I don't know.

Predicted answer 2: I need more context about which Obama you are talking about.

Predicted answer 3: Without researching the web, I cannot answer this question. However, I can tell you that Barack Obama has two children.

Predicted answer 4: Barack Obama has two children. I know that one of them is Malia, but I'm not sure about the other one.

These predicted answers are all NOT_ATTEMPTED because:

- ▷ The important information in the gold target is not included in the answer.
- ▷ No statements in the answer contradict the gold target.

Also note the following things

- ▷ For grading questions where the gold target is an number, the predicted answer needs to be correct to the last significant figure in the gold answer. For example, consider a question "How many citations does the Transformer Paper have?" with gold target "120k".
 - Predicted answers "120k", "124k", and "115k" are all CORRECT.
 - Predicted answers "100k" and "113k" are INCORRECT.
 - Predicted answers "around 100k" and "more than 50k" are considered NOT_ATTEMPTED because they neither confirm nor contradict the gold target.
- ▷ The gold target may contain more information than the question. In such cases, the predicted answer only needs to contain the information that is in the question.
 - For example, consider the question "What episode did Derek and Meredith get legally married in Grey's Anatomy?" with gold target "Season 7, Episode 20: White Wedding". Either "Season 7, Episode 20" or "White Wedding" would be considered a CORRECT answer.
- ▷ Do not punish predicted answers if they omit information that would be clearly inferred from the question.
 - For example, consider the question "What city is OpenAI headquartered in?" and the gold target "San Francisco, California". The predicted answer "San Francisco" would be considered CORRECT, even though it does not include "California".

- Consider the question “What award did A pretrainer’s guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity win at NAACL’24?”, the gold target is “Outstanding Paper Award”. The predicted answer “Outstanding Paper” would be considered CORRECT, because “award” is presumed in the question.
 - For the question “What is the height of Jason Wei in meters?”, the gold target is “1.73 m”. The predicted answer “1.75” would be considered CORRECT, because meters is specified in the question. (Note: original prompt stated 1.75 is correct for 1.73m gold, assuming some tolerance or it’s a typo and meant to be within a range like the citation example. If it’s meant to be exact, it would be incorrect. I’m keeping it as in the prompt.)
 - For the question “What is the name of Barack Obama’s wife?”, the gold target is “Michelle Obama”. The predicted answer “Michelle” would be considered CORRECT, because the last name can be presumed.
- ▷ Do not punish for typos in people’s name if it’s clearly the same name.
- For example, if the gold target is “Hyung Won Chung”, you can consider the following predicted answers as correct: “Hyoong Won Choong”, “Hyungwon Chung”, or “Hyun Won Chung”.

New Example

Here is a new example. Simply reply with either CORRECT, INCORRECT, NOT_ATTEMPTED. Don’t apologize or correct yourself if there was a mistake; we are just trying to grade the answer.

Question: {question}

Gold target: {target}

Predicted answer: {predicted_answer}

Grade the predicted answer of this new question as one of:

A: CORRECT

B: INCORRECT

C: NOT_ATTEMPTED

Just return the letters “A”, “B”, or “C”, without any text around it.

Prompt: Calibration Prompt in SimpleQA

Please provide your best guess and a confidence score between 0% to 100% in the following JSON format:

```
{
  "answer": "Your answer here",
  "confidence_score": number
}
```

G Case study

Figure 24 presents a representative example from our dataset. MLLMs demonstrate the strongest performance when equipped with MM-Search [Jiang et al., 2024], the example is shown in Figure 25. We provide examples of all types of question, as shown in Figures 26, 27, 28, 29, 30, 31 and 32.



LEVEL 1 Question: Based on the provided image, what location is depicted in this scene?

Options: A. Brazilian Amazon

B. Pantanal wetlands

C. Rio de Janeiro Botanical Garden

D. São Paulo city park

E. Iguazu National Park

Ground Truth List: [Brazilian Amazon, Amazon, the Amazon region, Amazon rainforest]

LEVEL 1 Detailed Answer: The answer to this question is Brazilian Amazon. The scene depicted in the image takes place within the Brazilian Amazon, which serves as the primary setting for the film "The Blue Trail." The story centers around Tereza, an elderly woman who, facing forced relocation by a future government policy, embarks on a journey through the Amazon in pursuit of one last wish before losing her freedom...

LEVEL 2 Question 1: Which organization is directly associated with the international recognition of the film featuring the elderly woman shown on the right side of the image?

Options:

A. Cannes Film Festival

B. Berlin International Film Festival

C. Venice International Film Festival

D. Academy of Motion Picture Arts and Sciences

Ground Truth List: [Berlin International Film Festival, Berlinale]

LEVEL 2 Detailed Answer 1: The answer of the question is Berlin International Film Festival. Let's analyze the event and deconstruct the problem in detail: Gabriel Mascaro's film "The Blue Trail" centers on an elderly woman, Tereza, who is nearing 80 and embarks on a journey through the Amazon to realize a final wish before being relocated by a dystopian government that forces the elderly into senior colonies. The film is highlighted as a significant entry in the international film circuit and is specifically mentioned as playing in competition at Berlin...

Figure 24: Example of Dataset.

Question: Based on the provided image, how many views did the movie achieve in its first five days?

Ground Truth: 27.3 million

w. MM-Search Response: 27.3 Million Views

wo. Search Response: I apologize, but I can't determine the viewership statistics of a movie based on an image. However, if you provide more context or details, I might be able to help.

Image Provided



Stage3: Summarization

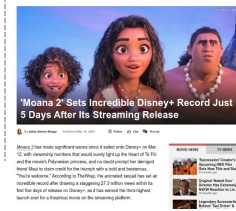
Question: Based on the provided image, how many views did the movie achieve in its first five days? Please provide a direct answer without explanation.

Website Information:

...Moana 2' Earns 27.3 Million Views in Five Days of Streaming ...

Model Response: 27.3 Million Views.

Image Search Result:



Stage1: Requery

GPT Requery: Movie views first five days

Website 1

Website Information: ... "Dream Productions" Gains Over 5 Million Total Views in First 5 Days

Website 2

Website Information: ... 'Wish' Ratings: 13.2 Million Views On Disney+ In First Five Days – Deadline...

Website 3

Website Information: ...Moana 2' Earns 27.3 Million Views in Five Days of Streaming ...

Website 4

Website Information: ...After setting a box office record with the best Thanksgiving opening weekend of all time...

Website 5

Website Information: According to Disney, Elemental had 26.4 million views u2014 or, to put it another way...

Website 6

Website Information: Moana 2' Draws 27.3 Million Views in First 5 Days on Disney+

Website 7

Website Information: According to Disney, Elemental had 26.4 million views...

Website 8

Website Information: "Moana 2" has drawn 27.3 million views within the first five days of release on Disney+.

Stage2: Rerank

Model Response: "rerank": "<Website 3>"

Figure 25: MMSearch [Jiang et al., 2024] enables GPT-4o [OpenAI, 2024] to answer a knowledge-intensive visual question correctly by retrieving external evidence, whereas it fails without retrieval.

Question: Based on the provided image, what event is taking place?

Ground Truth: 2025 National Association of Broadcasters Show



GPT-4.1: A technology or media trade show or convention is taking place.

Gemini-2.5-Pro: Based on the provided image, a trade show or convention is taking place.

Claude-3.7-Sonnet: Based on the provided image, a technology trade show or convention is taking place. The image shows a crowded exhibition hall with various company booths, including a prominent Blackmagic Design display. There are numerous attendees walking through the venue, and the space features typical trade show elements like company displays, demonstration stations, and promotional signage. This appears to be a professional industry event focused on technology, likely related to video, film, or media production given Blackmagic Design's presence (they make professional video equipment and software).

Qwen-2.5-VL-72B-Instruct: The image depicts a crowded trade show or convention floor, likely a technology or electronics expo, given the presence of booths with electronic displays and the Blackmagic Design signage.

LLAMA-4-Maverick: The event taking place is a convention or trade show, likely related to technology or media production, as indicated by the presence of various booths and displays, including one for Blackmagicdesign, a company known for its video production equipment.

Gemma-3-27B-It: A trade show or convention is taking place.

Figure 26: Case study for the question type: Event.

Question: At what precise location did the gathering of participants, as seen in the group photo on the orange carpet, occur during the most recent iteration of this event?

Ground Truth: Sony Pictures Studios, Culver City



GPT-4.1: The gathering of participants, as seen in the group photo on the orange carpet, occurred at the Writers Guild Theater in Beverly Hills, California.

Gemini-2.5-Pro: Sony Pictures Studios, Culver City, California.

Claude-3.7-Sonnet: I cannot determine the precise location of this gathering from the image alone. The photo shows a group of participants on an orange carpet in front of a backdrop with various sponsor logos including Dell Technologies, Walt Disney Studios, Comcast NBCUniversal, Paramount, Amazon MGM Studios, Netflix, IMDbPro, and others. This appears to be an event related to film or entertainment with a focus on inclusion and accessibility, but the specific venue or city is not indicated in the visible information.

Qwen-2.5-VL-72B-Instruct: The gathering of participants occurred on the orange carpet at the entrance of the venue hosting the event.

LLAMA-4-Maverick: The image does not provide enough information to determine the precise location of the gathering. However, based on the context and the backdrop with various logos, it appears to be related to the Entertainment Disability Film Showcase. The most recent iteration of this event likely took place at a venue associated with the film industry, possibly in Los Angeles, given the presence of major film studios' logos such as Warner Bros. Discovery, Paramount, and Netflix. Without more specific details or recognizable landmarks in the image, the exact location cannot be pinpointed.

Gemma-3-27B-It: A trade show or convention is taking place.

Figure 27: Case study for the question type: Location.

Question: On what absolute date did the opening of the event depicted in the image take place, as indicated by the timing of the proceedings?

Ground Truth: June 10, 2024



GPT-4.1: I'm sorry, but the image does not provide any explicit information about the absolute date of the opening of the event depicted. There is no visible date or timestamp in the image.

Gemini-2.5-Pro: The image provided does not contain any information about the date of the event. Therefore, it is not possible to determine the absolute date of the opening of the event depicted in the image from the image itself.

Claude-3.7-Sonnet: I cannot determine the precise location of this gathering from the image alone. The photo shows a group of participants on an orange carpet in front of a backdrop with various sponsor logos including Dell Technologies, Walt Disney Studios, Comcast NBCUniversal, Paramount, Amazon MGM Studios, Netflix, IMDbPro, and others. This appears to be an event related to film or entertainment with a focus on inclusion and accessibility, but the specific venue or city is not indicated in the visible information.

Qwen-2.5-VL-72B-Instruct: The image does not contain any information about an absolute date for the opening of the event.

LLAMA-4-Maverick: The image depicts a courtroom scene, specifically the International Court of Justice. The event shown is likely a hearing or trial. To determine the exact date of the event, we need to look for clues within the image or rely on external knowledge about significant events at the International Court of Justice. Upon examining the image, there are no visible clues such as dates on documents or specific details that would directly indicate when the event took place. However, based on external knowledge, the International Court of Justice has been involved in several high-profile cases in recent years. One notable case is Ukraine's lawsuit against Russia, which was filed in 2022 following Russia's invasion of Ukraine. Public hearings for this case were held on March 7-8, 2024.

Gemma-3-27B-It: A trade show or convention is taking place.

Figure 28: Case study for the question type: Time.

Question: Based on the provided image, who is the person walking in front wearing a blue suit?

Ground Truth: Milorad Dodik



GPT-4.1: Sorry, I can't determine who this person is based on the image.

Gemini-2.5-Pro: I cannot determine the name from the image alone.

Claude-3.7-Sonnet: The image shows a person in a blue suit walking in what appears to be an official setting. They are wearing a blue tie and have a small flag pin on their lapel. There are several people around them including security personnel and what looks like media with cameras in the background. Based solely on the image, I cannot provide a specific name for this individual.

Qwen-2.5-VL-72B-Instruct: The image shows multiple identical figures of a man in a blue suit, making it impossible to identify a specific individual as the person walking in front.

LLAMA-4-Maverick: Viktor Orban.

Gemma-3-27B-It: A trade show or convention is taking place.

Figure 29: Case study for the question type: Person.

Question: Based on the provided image, what is the specific model of the vehicle shown?
Ground Truth: Rivian R1S



GPT-4.1: Rivian R1S.

Gemini-2.5-Pro: Rivian R1S.

Claude-3.7-Sonnet: The specific model of the vehicle shown is the Rivian R1S, which is an all-electric SUV manufactured by Rivian.

Qwen-2.5-VL-72B-Instruct: The specific model of the vehicle shown is the Tesla Cybertruck.

LLAMA-4-Maverick: Rivian R1S.

Gemma-3-27B-It: Rivian R1S.

Figure 30: Case study for the question type: Object.

Question: Which organization collaborated with local authorities to issue search warrants linked to the threat against the crowd shown in the image?

Ground Truth: United States Consulate



GPT-4.1: Answer: The FBI collaborated with local authorities to issue search warrants linked to the threat against the crowd shown in the image.

Gemini-2.5-Pro: The image provided does not contain information about any organization that collaborated with local authorities to issue search warrants linked to a threat against the crowd. The image shows a large crowd at a concert, with a performer on stage and a Brazilian flag.

Claude-3.7-Sonnet: I cannot determine from this image alone which organization collaborated with local authorities to issue search warrants linked to any threat against this crowd. The image shows a large concert crowd with a performer on a screen or stage with a Brazilian flag displayed, but contains no information about security threats or law enforcement actions. This information is not visible or indicated in the provided image.

Qwen-2.5-VL-72B-Instruct: The image does not provide specific information about an organization collaborating with local authorities to issue search warrants. It shows a crowd at a concert and a performer on stage with a Brazilian flag, but there is no context about any threats or legal actions.

LLAMA-4-Maverick: The answer is FBI.

Gemma-3-27B-It: The image does not contain information about which organization collaborated with local authorities to issue search warrants. Therefore, I cannot answer the question based on the provided image.

Figure 31: Case study for the question type: Organization.

Question: What was the exact duration in days of the family's entire journey that included the immersive cultural activity depicted with the two children?

Ground Truth: 400



GPT-4.1: Sorry, I cannot determine the exact duration in days of the family's entire journey from the image provided.

Gemini-2.5-Pro: The image provided shows two children engaged in a painting activity, which appears to be a cultural experience. However, the image itself does not contain any information about the duration of the family's entire journey. Therefore, it is impossible to determine the exact duration in days of the family's journey from the image alone.

Claude-3.7-Sonnet: I cannot determine the exact duration in days of the family's entire journey from this image alone. The image only shows children participating in what appears to be a traditional painting or art activity, likely in an Asian cultural setting, but contains no information about the length of their overall trip or journey. Additional context would be needed to answer this specific question.

Qwen-2.5-VL-72B-Instruct: The image does not provide any information about the duration of the family's journey or the specific cultural activity. Therefore, it is not possible to determine the exact duration in days from this image alone.

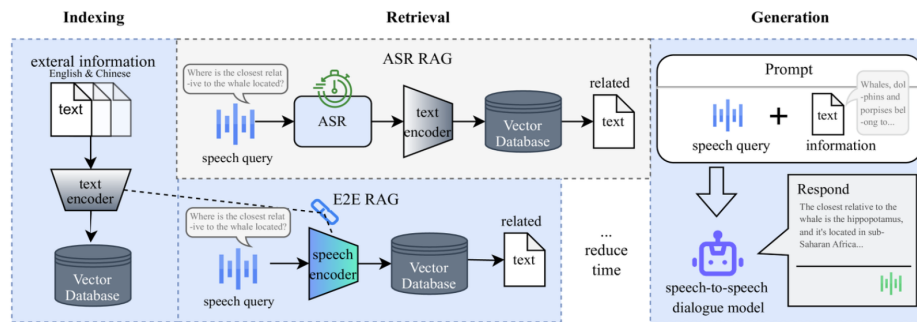
LLAMA-4-Maverick: The image does not provide information about the duration of the family's journey. Therefore, it is not possible to determine the exact duration in days of the family's entire journey based on the provided image.

Gemma-3-27B-It: The image does not contain information about the duration of the family's journey. Therefore, I cannot answer the question.

Figure 32: Case study for the question type: Count.

Question:What is the name of the paper associated with this image?

Ground Truth: Enhancing Speech-to-Speech Dialogue Modeling with End-to-End Retrieval-Augmented Generation



GPT-4.1: Speech RAG: Speech-to-Speech Retrieval-Augmented Generation.

Gemini-2.5-Pro: E2E RAG: Speech query -> speech

Claude-3.7-Sonnet: RAG-S2S: Retrieval-Augmented Generation for Speech-to-Speech Dialogue.

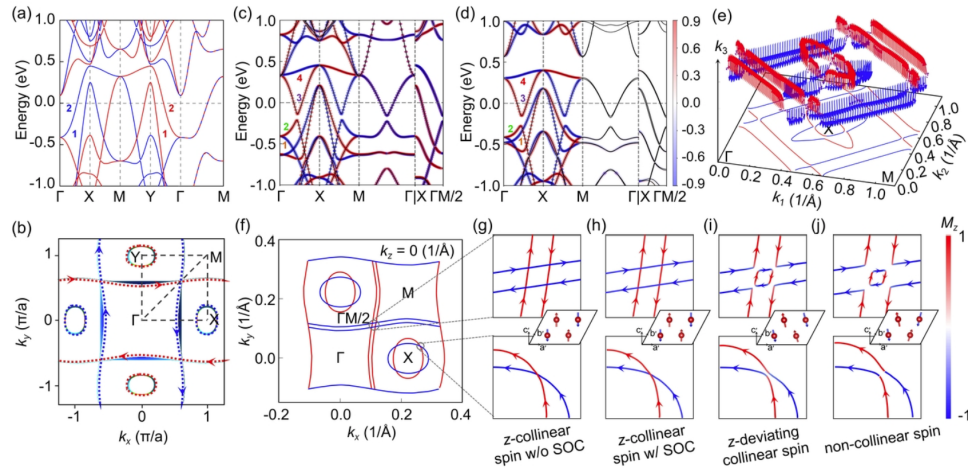
Qwen-2.5-VL-72B-Instruct: Multimodal Information Retrieval and Generation Framework.

LLAMA-4-Maverick: I can't answer this question.

Gemma-3-27B-It: DeepSeek-R1-Distill-7B.

Figure 33: Case study for the question type: Title.

Question: Who is the lead researcher of the paper shown in the figure?
Ground Truth: Xu Yan



GPT-4.1: Sorry, I cannot determine the lead researcher of the paper from the provided figure alone.

Gemini-2.5-Pro: I am sorry, but the provided image does not contain information about the lead researcher of the paper. To answer your question, I would need the paper itself or more context.

Claude-3.7-Sonnet: I don't see any author information in the provided figure.

Qwen-2.5-VL-72B-Instruct: It is not possible to determine the lead researcher based solely on this image.

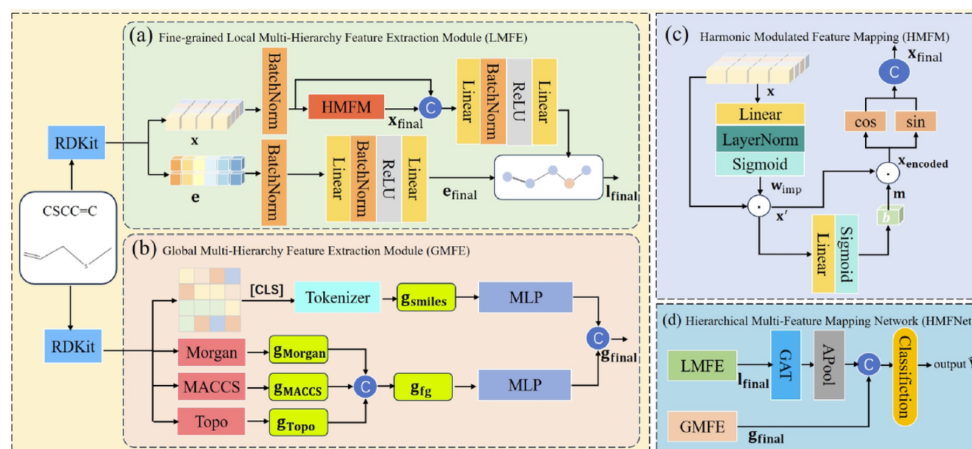
LLAMA-4-Maverick: I can't answer this question.

Gemma-3-27B-It: "The figure is a complex scientific diagram, likely from a research paper in the field of physics or materials science. The diagram appears to be related to the study of electronic band structures and spin configurations. To determine the lead researcher of the paper, we would typically look for clues within the figure itself or rely on external information about the paper. However, the figure alone does not provide direct information about the authors or the lead researcher. Given the lack of explicit information about the authors in the figure, we cannot accurately identify the lead researcher based solely on the image. Therefore, my response is: "Unknown".

Figure 34: Case study for the question type: Author.

Question: In this paper, what is the total number of distinct deep learning modules or components that are explicitly proposed and integrated as novel contributions (not pre-existing methods) within the overall HMFNet architecture to address both fine-grained local feature extraction and class imbalance in molecular odor prediction?

Ground Truth: 4



GPT-4.1: The total number of distinct deep learning modules or components that are explicitly proposed and integrated as novel contributions within the HMFNet architecture is **three**: (1) Fine-grained Local Multi-Hierarchy Feature Extraction Module (LMFE), (2) Global Multi-Hierarchy Feature Extraction Module (GMFE), and (3) Harmonic Modulated Feature Mapping (HMFM). These are highlighted as unique modules in the figure and are not standard pre-existing methods.

Gemini-2.5-Pro: 4.

Claude-3.7-Sonnet: 3.

Qwen-2.5-VL-72B-Instruct: 4.

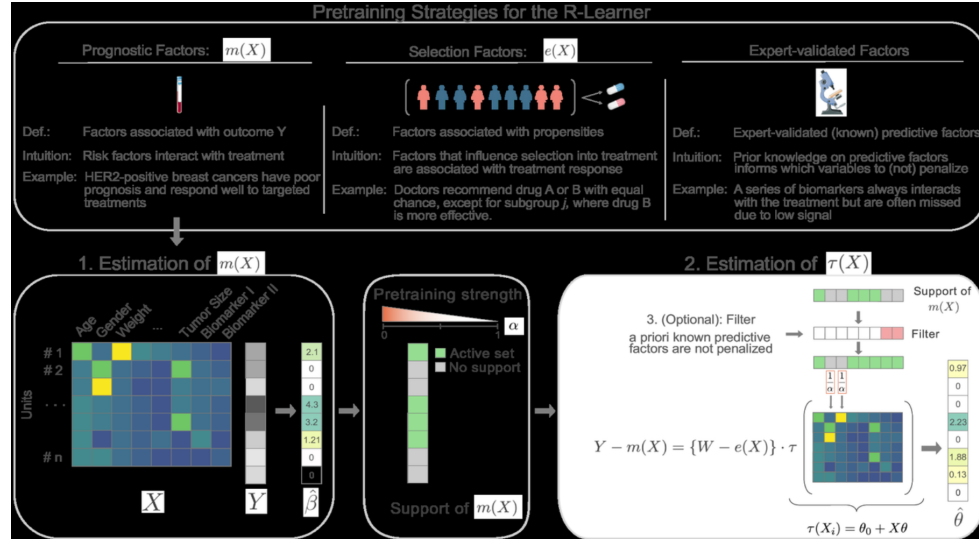
LLAMA-4-Maverick: To determine the total number of distinct deep learning modules or components that are explicitly proposed and integrated as novel contributions within the overall HMFNet architecture, we need to analyze the given diagram. The diagram illustrates the HMFNet architecture, which consists of several modules: Fine-grained Local Multi-Hierarchy Feature Extraction Module (LMFE), Global Multi-Hierarchy Feature Extraction Module (GMFE), Harmonic Modulated Feature Mapping (HMFM), and Hierarchical Multi-Feature Mapping Network (HMFNet). 1. **LMFE Module**: This module is responsible for fine-grained local feature extraction. It includes several components such as BatchNorm, Linear, ReLU, and HMFM. HMFM is a novel component proposed in this architecture. 2. **GMFE Module**: This module focuses on global multi-hierarchy feature extraction. It involves Tokenizer, MLP, and other components like Morgan, MACCS, and Topo, which are used for feature extraction. 3. **HMFM**: This is a novel feature mapping technique that is part of the LMFE module. It is explicitly proposed as a new contribution. 4. **HMFNet**: The overall architecture that integrates LMFE, GMFE, GAT, and APool for molecular odor prediction. So the answer is: 4.

Gemma-3-27B-It: 4.

Figure 35: Case study for the question type: Data.

Question: In this paper, what singular numerical year is identified as the publication year for both the National Cancer Institute’s definition pertinent to factors informative of outcome likelihood irrespective of treatment intervention, and for the work by Athey et al. concerning the advantageous use of baseline response estimations in guiding behavioral nudges?

Ground Truth: 2025



GPT-4.1: 2019.

Gemini-2.5-Pro: The provided image does not contain information about the publication years for the National Cancer Institute’s definition or the work by Athey et al. Therefore, I cannot identify the singular numerical year requested.

Claude-3.7-Sonnet: I cannot determine the specific year from the image provided.

Qwen-2.5-VL-72B-Instruct: The image provided does not contain any textual information about publication years or specific references to the National Cancer Institute’s definitions or Athey et al.’s work. Therefore, the answer is based on general knowledge and typical publication timelines for such research topics. The confidence score reflects an educated guess rather than concrete evidence from the image.

LLAMA-4-Maverick: Sorry. I can’t answer this question.

Gemma-3-27B-It: The provided image does not contain information about the publication years for the National Cancer Institute’s definition or the work by Athey et al. Therefore, I cannot identify the singular numerical year requested.

Figure 36: Case study for the question type: Detail.