LIVEVQA: Assessing Models with Live Visual Knowledge

Anonymous CVPR submission

Paper ID *****

Figure 1. LIVEVQA comprises 14 different News categories, containing 1233 News and 3602 question-answer pairs. Each instance includes a representative image, QA pair for basic image for understanding, and two multimodal multi-hop QA pairs for deeper reasoning.

Abstract

We introduce LIVEVQA, *an automatically collected dataset of latest visual knowledge from the Internet with synthesized VQA problems.* LIVEVQA *consists of 3,602 single- and multi-hop visual questions from 6 news websites across 14 news categories, featuring high-quality image-text coherence and authentic information. Our evaluation across 15 MLLMs (e.g., GPT-4o, Gemma-3, and Qwen-2.5-VL family) demonstrates that stronger models perform better overall, with advanced visual reasoning capabilities proving crucial for complex multi-hop questions. Despite excellent performance on textual problems, models with tools like search engines still show significant gaps when addressing visual questions requiring latest visual knowledge, highlighting important areas for future research.*

1. Introduction

In today's rapidly evolving information landscape, the ability to understand and reason about live content has become increasingly crucial. As news and events, continuously update across the globe, AI systems that can effectively process, comprehend, and respond to this dynamic information flow are essential for applications ranging from personalized experience [24] to real-time decision support [30].

Large language models (LLMs) have made remarkable progress in understanding and reasoning about live textual content when integrated with search engines [9, 15]. However, while live textual knowledge understanding has advanced significantly, a critical question remains unanswered: has other modality knowledge in live contexts—such as visual knowledge—been similarly solved?

To address this research gap, we introduce LIVEVQA, a automatically collected benchmark dataset specifically designed to evaluate current AI system on their ability to answer questions requiring live visual knowledge. LIVEVQA is constructed with three key design principles: *(1)* strict temporal filtering to prevent dataset contamination, ensuring evaluation of true retrieval capabilities rather than memorized knowledge, *(2)* automated ground truth with human-in-the-loop annotations, and *(3)* high-quality and authentic image-question pair to ensure meaningful visual knowledge challenges. Finally, our dataset comprises 1,233 authentic news articles with 3,602 question-answer pairs sourced from six major global news platforms and categorized across 14 domains. Each instance features a representative image paired with three types of questions: a basic visual understanding question and two difficult multi-hop questions requiring deeper reasoning.

Our evaluation encompasses 15 state-of-the-art MLLMs (*e.g.*, Gemini-2.0-Flash [16], Qwen-2.5-VL [18, 19, 29] and Gemma-3 family [17], and visual search engine [8]) reveals that while larger models generally perform better (with Gemini-2.0-Flash achieving the highest overall accuracy of 24.93%), significant challenges remain in addressing complex multi-hop visual questions requiring current knowledge. Models equipped with stronger reasoning capabilities, such as QvQ-72B-Preview, show advantages in handling multi-hop reasoning tasks, achieving 7.41% accuracy on reasoning questions compared to 1.35% for base models. Notably, integrating GUI-based MM-search substantially improves performance, boosting Gemini-2.0-Flash's accuracy to 29.00%, with particularly strong gains on challenging Level 2 and Level 3 questions (reaching 22.75% and 13.66% respectively).

We hope LIVEVQA provides valuable insights into the current state of live visual knowledge and highlights promising directions for future research.

Table 1. The distribution of 1,232 news instances across 14 categories and 6 major sources, containing 3602 VQA.

Category	Overall		By News Source (%)					
	Count	%	VRTY	BBC	CNN	APNWS	FORB	YHO
Sports	305	24.8	1.0	48.8	20.3	7.5	15.5	0.0
Other	219	17.8	1.0	17.3	25.3	28.4	13.6	30.0
Movies	102	8.3	36.7	0.7	1.7	6.0	5.8	0.0
TV	89	7.2	31.0	1.8	2.1	2.5	4.9	5.0
Science	80	6.5	0.0	5.5	7.1	16.9	0.0	20.0
Economy	72	5.8	0.0	4.4	7.9	8.0	14.6	10.0
Health	67	5.4	1.0	6.6	3.3	12.4	1.0	5.0
Media	58	4.7	7.6	3.1	7.5	3.5	1.9	5.0
Music	47	3.8	11.9	2.0	0.8	3.0	4.9	0.0
G.Business	45	3.7	1.9	1.8	7.5	2.5	6.8	15.0
Tech	45	3.7	2.4	2.6	4.2	3.0	10.7	5.0
Opinion	45	3.7	1.0	2.4	8.3	2.5	5.8	5.0
Art/Design	43	3.5	0.0	2.4	4.2	4.0	13.6	0.0
Theater	15	1.2	4.8	0.9	0.0	0.0	1.0	0.0
Total	1,232	100	210	457	241	201	103	20
Source %	100		17.1	37.1	19.6	16.3	8.4	1.6

2. LIVEVQA: The Dataset

We introduce LIVEVQA, a benchmark dataset for live visual knowledge, as illustrated in Figure 1. We incorporates multi-hop question-answer pairs that establish explicit dependencies between textual news facts and visual content, evaluating both the reasoning capabilities and generalization potential of multimodal models. Each instance in LIVEVQA: *(1)* a representative image, *(2)* a basic question-answer pair establishing basic image-text correspondence, and *(3)* Two detailed question-answer pairs requiring deeper reasoning. Based on the classification system of [23], we selected 14 major news categories, such as Sports, Movies, Television, Science, Economy, *etc.* Each news is classified by gpt-4o-mini [14]. The statistics is shown in Table 1.

2.1. Dataset Construction

Data Collection We select six global news platforms: CNN, BBC, Yahoo, Forbes, AP News, and Variety. These sources provide comprehensive geographic coverage and content diversity, with an average of 2–3 images per article to ensure visual richness.

Our collection pipeline consists of three steps:

- **URL Normalization**: Predefined patterns and regular expressions are used to identify actual news pages while filtering out indexes, advertisements, and pure-textual content, ensuring data relevance and authenticity.
- **Structured Content Extraction**: A multi-level parsing strategy is employed, starting with site-specific CSS selectors (*e.g.*, `h1.pg-headline` for CNN, `h1.story-headline` for BBC) to extract key content. Generic selectors serve as a fallback, and metadata extraction is used as a last resort to enhance robustness.
- **Image Filtering**: The pipeline prioritizes content-relevant images over decorative elements, particularly those marked with `og:image` meta tags. All images undergo standardized processing, resized to 1024×680 pix-

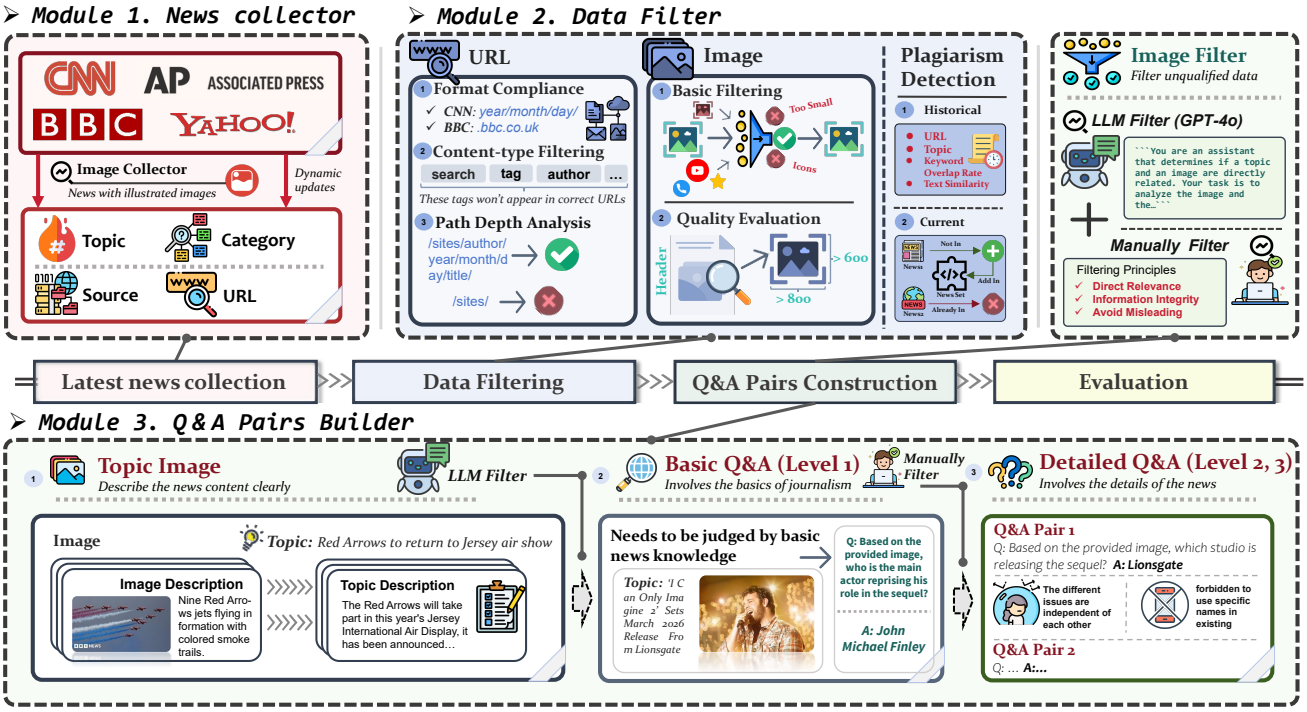

Figure 2. **Pipline of LIVEVQA data engine.** Our pipeline consists of three modules: news collector, data filter, and Q&A pairs builder. It collects illustrated news from mainstream media, performs multi-level data filtering, and generates foundational and detailed Q&A pairs for training multimodal question-answering models.

els while maintaining aspect ratio, assigned unique identifiers, and stored in a dedicated repository to ensure consistency and traceability.

Question-Answer Generation. We employ GPT-4o [13] to generate QA pairs based on raw news document, with each sample comprises three components: *(1)* an image reflecting the news topic accurately, *(2)* a basic question requiring to understand the image content, and *(3)* two muti-hop detailed questions requiring cross-modality reasoning.

To ensure valuable images for generating, we prioritize relevance. Notably, some news articles feature header images that bear little connection to their content. For instance, may some news about harmful foods just show a close-up of the food without conveying substantive information. To address this, we use GPT-4o [13] to evaluate the relevance of images by analyzing the news title, content, and associated visuals. See Appendix 7 for more details.

- **Basic 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 like color or shape. Answers are constrained to factual phrases of 2–7 words. For example, *"Who is the person speaking in the image?"* is valid, whereas *"What color is the person's tie?"* is filtered out.
- **Mutli-hop Questions** require deeper contextual reasoning. The two muti-hop questions that must be answerable only through the news text, covering events, person, time, etc. Questions must remain distinct and must not reveal or reference answers from prior questions.

2.2. Data Statistics

Finally, we collect 1,232 carefully curated news spanning 14 categories and 6 global news platforms, amounting to a total of 3,602 QA pairs. As illustrated in Figure 1, the dataset covers a diverse range of news topics with representative examples, showcasing its breadth and richness in both content and modality. LIVEVQA demonstrates distinct domain specificity. As shown in Table 1, sports news is the most prevalent category, with a significant portion sourced from BBC, highlighting its strength in sports reporting. News sources also exhibit clear domain preferences—Variety primarily covers film and music, Forbes focuses on business, and AP News emphasizes science and health. Additionally, we categorize the final generated questions into 8 distinct types, as shown in Table 2.

3. Experiments and Analysis

3.1. Experiment Setups

Models. We conduct a series of zero-shot testing for a diverse range of *state-of-the-art* MLLMs, including Gemini 2.0 Flash [16], Qwen2.5-VL-3/7/32/72B [29], Gemma-3-4/12/27B-it [17], QVQ-72B-Preview [18], QVQ-Max [19], GPT-4o-mini [14], and GPT-4o [13].

Table 2. Overall performance on LIVEVQA. See Table 3 for performance on another categorizing taxonomy for live visual knowledge.

Model	Avg.	L1	L2	L3	Per.	Loc.	Tim.	Eve.	Org.	Obj.	Rea.	Oth
w.o. Search												
Gemma-3-4b-it	14.65	38.42	3.10	2.46	19.20	11.96	2.82	14.51	26.75	28.37	2.89	10.26
Gemma-3-12b-it	17.10	44.19	3.47	3.71	23.96	15.78	**5.08**	15.95	29.40	29.58	2.69	12.25
Gemma-3-27b-it	20.43	48.50	7.93	4.92	29.19	17.77	2.82	**20.50**	34.46	35.21	5.17	15.23
Qwen2.5-VL-3B	15.63	39.98	4.58	2.38	25.65	13.29	3.11	12.98	28.67	27.89	2.89	5.30
Qwen2.5-VL-7B	18.74	41.28	7.44	3.63	29.43	17.61	3.07	16.89	30.23	33.82	2.87	10.67
Qwen2.5-VL-32B	18.96	47.93	5.12	3.88	27.19	17.61	2.82	17.54	33.49	35.21	4.75	8.61
Qwen2.5-VL-72B	21.07	55.93	5.94	1.35	32.87	20.60	4.52	19.59	**35.66**	32.96	3.51	12.25
GPT-4o	16.38	41.02	4.54	3.62	2.61	21.43	**5.08**	18.68	28.67	**41.97**	6.20	15.23
GPT-4o-mini	17.30	43.71	4.95	3.19	5.84	**21.93**	3.67	20.05	32.53	41.13	6.20	13.58
Gemini-2.0-Flash	**24.93**	**58.81**	8.75	5.86	**43.01**	20.93	4.24	19.36	**35.66**	43.10	6.61	**19.54**
QVQ-72B-Preview	19.94	39.90	**11.62**	**7.41**	21.81	19.44	2.25	19.95	34.46	36.52	**10.33**	13.58
QVQ-Max	17.80	38.10	9.50	4.91	24.88	17.94	3.67	15.95	33.01	26.76	4.34	11.59
w. Search												
GPT-4o	13.38	28.43	5.78	5.34	2.46	13.79	3.39	18.45	22.17	34.93	6.61	13.91
GPT-4o-mini	22.27	32.58	**19.49**	**14.22**	12.14	21.26	**11.58**	**26.42**	34.22	42.25	**15.50**	**23.51**
Gemini-2.0-Flash	**29.46**	**59.63**	16.43	11.03	**44.85**	**25.91**	**11.58**	24.60	**44.10**	**45.63**	9.92	**23.51**
w. MM Search [8]												
GPT-4o	20.20	34.88	15.57	9.32	8.18	20.75	16.98	**25.97**	**34.48**	**52.38**	8.33	10.64
GPT-4o-mini	21.80	41.28	14.97	8.07	24.55	24.53	7.55	16.88	31.03	47.62	10.00	17.02
Gemini-2.0-flash	**29.00**	**49.42**	**22.75**	**13.66**	**44.55**	**26.42**	**20.75**	20.78	29.31	42.86	**11.67**	**27.66**

Models with Search Engine. We enable the built-in search functionalities and MM-Search [8] with Gemini-2.0-Flash [16], GPT-4o-mini [14], and GPT-4o [13].

Metric. We instruct GPT-4o-mini [14] as an impartial judge, which strictly answer only *"yes"* or *"no"* and marks the final answer with <*answer*> tags.

3.2. Experiment Results

Larger-scale models demonstrate improved performance across difficulty levels, though proprietary models retain a clear advantage. For models within the same family (*e.g.*, Gemma or Qwen), we observe that increasing model size leads to consistently better accuracy across all question difficulty levels. For instance, Gemma-3-4b-it achieves only 2.46% on L3-level questions, whereas Gemma-3-27b-it reaches 4.92%. Despite these improvements, open-source models still lag behind proprietary models in overall performance. Notably, **Gemini-2.0-flash** achieves the best results across nearly all dimensions with an overall accuracy of **24.93%**, which may be attributed to its more recent data coverage, extending up to 2024 June.

Strong visual reasoning ability plays a critical role in boosting cross-modality multi-hop question. Models equipped with stronger reasoning capabilities, such as QvQ-72B-Preview and QvQ-Max, outperform their base model Qwen2.5-VL-72B, highlighting the effectiveness of enhancing visual reasoning abilities in live knowledge.

Current models excel in entity-centric categories but struggle with abstract reasoning. Across different question categories, models perform better on concrete entity recognition tasks such as **Person**, **Organization**, and **Object**. In contrast, their performance drops significantly on more abstract tasks like **Time** and **Reason**, indicating that current models are still limited in their ability to conduct causal reasoning and temporal understanding. As shown in Table 3, models achieve higher performance in domains with rich visual and textual cues are present. However, performance is notably lower in more ambiguous or opinionated domains such as **Opinion** and **Other**, reflecting the difficulty of handling multi-intent or subjective content in current multimodal models.

Incorporating the MMSearch and integrated search engine significantly improves the performance. Gemini-2.0-Flash sees its average accuracy rise from 24.93% (Table 2) to 29.00%, with substantial gains on harder questions—achieving 22.75% and 13.66% on L2 and L3 respectively. These results demonstrate that integrating retrieval-based evidence is particularly helpful for addressing questions that go beyond the internal knowledge scope of the models. Notably, GPT-4o-mini exhibits a more pronounced improvement than GPT-4o, highlighting its strong synergy with retrieval pipelines and its potential as a lightweight yet effective reasoning agent.

4. Conclusion

This paper presents LIVEVQA, a comprehensive benchmark for evaluating MLLMs on live visual knowledge across 15 models (3B-72B parameters). Equipping models with online search tools or GUI-based image search [8] significantly enhances performance on these queries, indicating promising approaches for future study.

References

[1] Sandipan Basu, Aravind Gaddala, Pooja Chetan, Garima Tiwari, Narayana Darapaneni, Sadwik Parvathaneni, and Anwesh Reddy Paduri. Building a question and answer system for news domain. *arXiv preprint arXiv:2105.05744*, 2021. 1

[2] Alessandro Castelnovo, Riccardo Crupi, Nicole Inverardi, Daniele Regoli, and Andrea Cosentini. Investigating bias with a synthetic data generator: Empirical evidence and philosophical interpretation. *arXiv preprint arXiv:2209.05889*, 2022. 1

[3] Simin Chen, Yiming Chen, Zexin Li, Yifan Jiang, Zhongwei Wan, Yixin He, Dezhi Ran, Tianle Gu, Haizhou Li, Tao Xie, et al. Recent advances in large langauge model benchmarks against data contamination: From static to dynamic evaluation. *arXiv preprint arXiv:2502.17521*, 2025. 1

[4] Pranay Gupta and Manish Gupta. Newskvqa: Knowledge-aware news video question answering. In *Pacific-asia conference on knowledge discovery and data mining*, pages 3–15. Springer, 2022. 1

[5] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020. 1

[6] Soumya Jahagirdar, Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Watching the news: Towards videoqa models that can read. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4441–4450, 2023. 1

[7] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024. 1

[8] 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. 2, 4, 1, 3, 6

[9] Bowen Jin, Hansi Zeng, Zhenrui Yue, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025. 2, 1

[10] Kausik Lakkaraju, Aniket Gupta, Biplav Srivastava, Marco Valtorta, and Dezhi Wu. The effect of human v/s synthetic test data and round-tripping on assessment of sentiment analysis systems for bias. In *2023 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, pages 380–389. IEEE, 2023. 1

[11] Xiang Lisa Li, Evan Zheran Liu, Percy Liang, and Tatsunori Hashimoto. Autobencher: Creating salient, novel, difficult datasets for language models. *arXiv preprint arXiv:2407.08351*, 2024. 1

[12] Yucheng Li, Frank Guerin, and Chenghua Lin. Latesteval: Addressing data contamination in language model evaluation through dynamic and time-sensitive test construction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18600–18607, 2024. 1

[13] OpenAI. Gpt-4o, 2024. Accessed: 2024-06-01. 3, 4, 2

[14] OpenAI. Gpt-4o mini: Advancing cost-efficient intelligence. https://openai.com, 2024. 2, 3, 4

[15] OpenAI. Search gpt. OpenAI, 2024. Accessed: 2025-03-29. 2, 1

[16] 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. 2, 3, 4, 6

[17] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. 2, 3

[18] Qwen Team. Qvq-72b-preview: A large multimodal model by qwen. https://qwenlm.github.io/zh/blog/qvq-72b-preview/, 2024. 2, 3, 6

[19] Qwen Team. Qvq-max: Think with evidence. https://qwenlm.github.io/zh/blog/qvq-max/, 2024. 2, 3

[20] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*, 2016. 1

[21] Jiexin Wang, Adam Jatowt, and Masatoshi Yoshikawa. Archivalqa: A large-scale benchmark dataset for open-domain question answering over historical news collections. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3025–3035, 2022. 1

[22] Siyuan Wang, Zhuohan Long, Zhihao Fan, Zhongyu Wei, and Xuanjing Huang. Benchmark self-evolving: A multi-agent framework for dynamic llm evaluation. *arXiv preprint arXiv:2402.11443*, 2024. 1

[23] Zhen Wang, Xu Shan, Xiangxie Zhang, and Jie Yang. N24news: A new dataset for multimodal news classification. *arXiv preprint arXiv:2108.13327*, 2021. 2

[24] Jiale Wei, Xiang Ying, Tao Gao, Felix Tao, and Jingbo Shang. Ai-native memory 2.0: Second me. *arXiv preprint arXiv:2503.08102*, 2025. 2

[25] Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, et al. Livebench: A challenging, contamination-free llm benchmark. *arXiv preprint arXiv:2406.19314*, 2024. 1

[26] Cheng Xu, Shuhao Guan, Derek Greene, M Kechadi, et al. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244*, 2024. 1

[27] Shicheng Xu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. Search-in-the-chain: Interactively enhancing large language models with search for knowledge-intensive tasks. In *Proceedings of the ACM Web Conference 2024*, pages 1362–1373, 2024. 1

[28] Junxiao Xue, Quan Deng, Fei Yu, Yanhao Wang, Jun Wang, and Yuehua Li. Enhanced multimodal rag-llm for accurate visual question answering. *arXiv preprint arXiv:2412.20927*, 2024. 1

[29] 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. 2, 3

[30] Bhada Yun, Dana Feng, Ace S Chen, Afshin Nikzad, and Niloufar Salehi. Generative ai in knowledge work: Design implications for data navigation and decision-making. *arXiv preprint arXiv:2503.18419*, 2025. 2

LIVEVQA: Assessing Models with Live Visual Knowledge

Supplementary Material

5. Related Works

Dynamic Testing via Synthetic Data Engine. Traditional static benchmarks are increasingly susceptible to data contamination and may fail to adequately capture the evolving capabilities of models. Consequently, dynamic testing and automatically-updated benchmarks have emerged as promising approaches to address these limitations and provide more rigorous assessments [7, 25, 26]. Building on these developments, recent advancements in dynamic testing and automatically-updated benchmarks [11, 22, 25] further mitigate the data contamination issues inherent in static benchmarks, bolstering the reliability of evaluation results [3]. These dynamic, time-sensitive testing methods are crucial for avoiding data contamination and enhancing evaluation accuracy [12]. Furthermore, the diversification of evaluation methodologies is being explored through the use of synthetic data and round-tripping techniques, which have demonstrated varying impacts in sentiment analysis evaluations [10]. Synthetic data generation frameworks also offer valuable insights into analyzing model biases and exploring their broader ethical and philosophical implications [2].

Large Models as Search Engines. Recent advances in LLMs as search engines have demonstrated significant progress [9, 15, 27].Traditional search engines rely on keyword matching, while LLMs use natural language understanding for precise, context-aware answers. Retrieval-augmented generation (RAG) enhances LLMs' knowledge retrieval capabilities, making them more suitable for search engines. Early RAG methods [5] integrated retrieval and generation, and later work [28] optimized interactive search. The latest work, Search GPT [15], combines RAG with online search to create a more efficient search engine architecture. Multimodal search engines have also shown great potential, integrating text, images or other types of information to provide a richer search experience. Recent research [8] improves both search accuracy and interactivity. Additionally, an enhanced version of multimodal RAG-LLM [28] has been proposed for accurate visual question answering, showcasing the application of multimodal RAG-LLM in cross-modal information retrieval.

News QA. News Question Answering (News QA) aims to enable systems to comprehend and respond to news-related questions, requiring efficient information retrieval capabilities to handle rapidly updating news data. Early research [1, 20] in News QA primarily focused on news text, where answers were composed of multiple textual fragments extracted from original articles. In recent years, News Visual Question Answering (NewsVQA) has emerged as a novel field that extends news information retrieval by enabling models to answer questions related to news images or videos through the integration of textual and visual information. However, NewsVQA faces several challenges, including multimodal fusion, temporal information processing, and scene text understanding. Current datasets [4, 6, 21] are used to evaluate model performance, with research efforts focusing on optimizing visual question answering models, integrating OCR for video text recognition, and developing more accurate evaluation methods.

6. Limitations

Our work is still in progress. While our study provides a comprehensive evaluation of *state-of-the-art* MLLMs on latest visual questions, several limitations remain: *(1)* Currently, our dataset primarily structures latest information into visual question answering formats. Additional synthetic data approaches such as image captioning or Chain-of-Thought reasoning could further enhance MLLMs' understanding and reasoning capabilities on *Live Visual Content*. *(2)* Our research predominantly derives *Live Visual Content* from mainstream news websites such as CNN and BBC, which may lead to imbalance and incomplete representation of current Internet content. Incorporating data from social media platforms such as X (formerly Twitter) and Reddit[1] could provide a more diverse and comprehensive dataset. *(3)* More robust visual search engines need to be developed to enhance model performance on latest visual knowledge queries.

7. Details in Constructing LIVEVQA

Raw Data Filtering. To ensure dataset quality, we design a multi-level filtering mechanism covering URL validation, image screening, and duplicate removal.

- **For URL filtering**, our pipeline judge the current URL according to corresponding news platforms (*e.g.* Format compliance ensures CNN URLs contain *"/year/month/day/"*, while BBC domains end with *".bbc.co.uk"*). Content-type filtering excludes non-news pages using regular expressions. Path depth analysis enforces platform-specific URL structures (*e.g.* Forbes' five-level paths).
- **For image selection**, website-specific CSS selectors (*e.g.*, *".image_container"* for CNN) extract candidate images. Basic filtering removes images under 100×100 pixels and excludes *"icon"/"logo"* URLs. Quality prioriti-

[1] reddit.com

zation selects lead images above 800×600 pixels, ensuring editorial relevance.
- **To eliminate duplicates**, we implement hierarchical deduplication for both historical and intra-session duplicates. Historical duplicates are identified by matching URLs, title fragments, or textual similarity (Levenshtein-based score>0.8). Intra-session duplicates are blocked through real-time title hashing, ensuring each article remains unique within a session.

Image Filtering. We provide GPT-4o [13] with comprehensive contextual information, including the news topic, original article URL, cleaned news text and summary, Base64-encoded image, and article classification labels, and instruct model to generate questions prefixed with *"Based on the provided image."*. Additionally, previously generated question-answer pairs are retained to ensure consistency.

8. Additional Experiment 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 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.

Accuracy variate across different category. We show the accuracy percentages of different models across various question types, difficulty levels, and domains, evaluated both with and without search functionality. In Table 3, Models like Gemini-2.0-Flash [16] stand out across several domains, achieving the highest overall accuracy of 24.93%. It also performs exceptionally well in domains such as Music, Sports, Global, and Science, with accuracy reaching 24.63%, 27.59%, 29.85%, and 25.59%, respectively. Other models like Qwen2.5-VL-72B and Qwen2.5-VL-32B [29] also perform well, but Gemini-2.0-Flash [16] generally outperforms them, particularly in more challenging domains. Gemini-2.0-Flash [16] again demonstrates superior performance with an overall accuracy of 29.00%, excelling in domains like Movies, Technology, and Other, where it achieves 33.80%, 33.33%, and 40.91%, respectively. The GPT-4o-mini [14] model also shows competitive results, especially in the domains of Music and Art, with scores of 17.65% and 46.15%, respectively. Models with search functionality, especially Gemini-2.0-Flash [16], consistently outperform those without, indicating the positive impact of integrating search capabilities into these models. These results suggest that while all models have strengths in specific domains, Gemini-2.0-Flash [16] is particularly effective across a broader range of topics.

Unexpected performance gap between GPT-4o-mini and GPT-4o. Notably, we observe that GPT-4o-mini outperforms its base version GPT-4o, which was initially unexpected. Through experimental analysis, we find that 4o-mini tends to consume more tokens during inference, which may account for its superior visual reasoning capabilities. Furthermore, under the *w.o. Search* setting, 4o actually performs better than its counterpart with search enabled. This could be attributed to the model's self-assessment mechanism, where it believes it can answer the query without invoking external resources, even though search functionality is permitted. In addition, both 4o and 4o-mini exhibit poor performance on tasks involving **face recognition**, likely due to platform-level privacy protections and policy constraints that restrict the handling of biometric information.

9. Prompt

We design a series of prompts to help the model better handle the following key process of our pipeline: (1) content classification for domain-specific organization, as shown in Table 4 (2) basic image-based QA generation with relevance verification, as shown in Table 5 (3) contextual QA generation requiring news background knowledge, as shown in Table 6 (4) question type diversification to ensure coverage across different information categories, as shown in Table 7 (5) binary correctness evaluation for answer validation, Table 8.

Table 3. Accuracy (%) of different models across question types and difficulty levels

Model	Avg.	Mus.	Mov.	The.	Tel.	Med.	Spo.	Glo.	Sci.	Eco.	Tec.	Hel.	Opi.	Art.	Oth.
w.o. Search															
Gemma-3-4b-it	14.65	8.82	14.77	13.64	11.41	19.16	16.02	20.15	14.66	16.59	16.15	11.73	12.98	13.11	13.55
Gemma-3-12b-it	17.10	8.21	13.13	18.18	14.07	22.75	18.46	21.64	15.52	19.43	15.38	17.35	17.56	13.93	18.22
Gemma-3-27b-it	20.43	16.42	17.51	18.18	15.97	26.35	23.14	23.88	18.53	21.80	23.85	18.88	19.08	14.75	19.94
Qwen2.5-VL-3B	15.63	10.45	12.79	9.09	9.89	22.75	18.69	18.66	12.93	19.91	15.38	14.80	12.98	13.93	14.80
Qwen2.5-VL-7B	18.74	13.46	18.52	16.00	14.51	21.53	21.52	19.76	17.32	**24.26**	20.78	16.10	16.78	16.18	16.78
Qwen2.5-VL-32B	18.96	11.94	14.48	11.36	13.69	23.35	21.80	23.13	18.53	23.70	21.54	18.37	17.56	13.93	18.69
Qwen2.5-VL-72B	21.07	15.67	17.85	**20.45**	16.73	23.95	25.25	26.87	19.40	23.70	21.54	16.33	19.85	15.57	20.09
GPT-4o	16.38	5.97	14.81	**20.45**	11.03	11.98	18.91	22.39	21.12	14.69	23.08	15.31	11.45	15.57	16.51
GPT-4o-mini	17.30	7.46	15.15	13.64	11.41	14.37	20.69	20.90	**24.57**	15.64	20.00	14.80	12.21	16.39	17.60
Gemini-2.0-Flash	**24.93**	**24.63**	**24.24**	**20.45**	**25.86**	**31.14**	**27.59**	**29.85**	18.97	**25.59**	**25.38**	**20.41**	**24.43**	**20.49**	**23.05**
QvQ-72B-Preview	19.94	16.18	17.39	**20.45**	15.21	23.95	22.80	23.13	20.26	20.85	22.31	17.86	19.85	15.57	18.69
QvQ-max	17.80	11.19	13.80	13.64	15.21	15.57	21.91	19.40	16.81	21.80	15.38	16.84	17.56	12.30	17.76
w. Search															
GPT-4o	13.38	2.99	12.12	15.91	8.75	10.18	15.46	12.69	19.83	13.74	16.15	14.29	10.69	11.48	13.55
GPT-4o-mini	22.27	14.93	16.16	11.36	18.25	19.76	27.03	23.88	**23.28**	22.75	29.23	**25.51**	16.03	16.39	22.12
Gemini-2.0-Flash	**29.46**	**32.84**	**27.27**	**25.00**	**32.70**	**32.34**	**32.26**	**38.06**	21.98	**29.86**	**30.00**	23.98	**25.95**	**22.13**	**28.50**
w. MM-Search [8]															
GPT-4o	20.20	17.65	19.72	16.67	13.33	17.65	**22.64**	11.11	50.00	15.38	25.00	22.22	8.33	15.38	21.21
GPT-4o-mini	21.80	17.65	18.31	**33.33**	8.89	**38.24**	20.13	3.70	**56.25**	7.69	**33.33**	22.22	0.00	**46.15**	28.79
Gemini-2.0-Flash	**29.00**	**23.53**	**33.80**	16.67	**33.33**	29.41	**22.64**	**22.22**	31.25	**23.08**	25.00	**44.44**	**16.67**	38.46	**40.91**

Table 4. Prompts for question category classification.

Task	Prompt
System Prompt for Classification	You are a professional content classification assistant. Your task is to categorize the provided content into one of the specified categories, returning only the category name.
User Prompt for Content Classification	Please classify the following content into the most appropriate single category from the list provided. Title: {topic} Content Description: {topic_description} Image Path: {image_path} Available categories with descriptions: - Health: Content related to health, medicine, wellness, diseases, or healthcare systems - Science: Content about scientific research, discoveries, space exploration, or natural phenomena - Television: Content about TV shows, streaming series, or television industry news - Movies: Content related to films, movie reviews, film industry, or cinema releases - Economy: Content about finance, markets, economic policies, or business trends - Sports: Content about athletic competitions, sports teams, athletes, or sporting events - Theater: Content about stage performances, plays, musicals, or theatrical productions - Music: Content about songs, musicians, concerts, albums, or the music industry - Opinion: Content expressing viewpoints, editorials, or commentary on current events - Art & Design: Content about visual arts, exhibitions, design, fashion, or architecture - Media: Content about journalism, publishing, social media, or news organizations - Technology: Content about tech innovations, gadgets, software, or digital trends - Global Business: Content about international trade, multinational corporations, or global economics - Other: Content that doesn't clearly fit into any of the above categories Analyze the title and content description carefully to determine the most suitable category. Please respond with only the category name, without any additional text or explanation. For example, if it's sports news, just reply with "Sports".

10. Taxonomy

Our news content is categorized into the following areas:

- **Health**: Content related to health, medicine, wellness, diseases, or healthcare systems, including advancements in medical research, treatments, preventive measures, healthcare policies, and trends in public health. It also covers topics like mental health, fitness routines, nutrition, and global health challenges.
- **Science**: Content about scientific research, discoveries, space exploration, or natural phenomena. This includes the latest breakthroughs in fields like biology, chemistry, physics, and environmental science, as well as updates on space missions, astronomical research, and the explo-

Table 5. Prompts for Level 1 QA generation.

Task	Prompt
Topic-Image Relevance Check	You are an assistant that determines if a topic and an image are directly related. Your task is to analyze the image and topic carefully, and decide if the image clearly depicts or is directly relevant to the topic. Guidelines: 1. The image must clearly depict the topic or be directly relevant to it. 2. If the image is only loosely related, indirectly related, or unrelated to the topic, it should be marked as irrelevant. 3. Be strict in your assessment. Only mark an image as relevant if there is a clear and direct connection to the topic. 4. Respond ONLY with "<relevant>" if the image is directly related to the topic, or "<irrelevant>" if it is not.
Basic QA Generation	You are an educational assistant specialized in creating simple, image-based Q&A pairs related to current topics and news. Generate ONE simple question-answer pair by following these requirements: - The question MUST begin with "Based on the provided image, ..." - The question should be SIMPLE and DIRECT, focusing on identifying people, objects, events, places, or dates visible in the image. - The answer must be a SHORT PHRASE (2-7 words), NOT a complete sentence. - The answer must be FACTUAL, UNIQUE, and VERIFIABLE.

Table 6. Prompts for Level 2 QA generation.

Task	Prompt
System Prompt for News Context QA	You are a specialized AI assistant for creating image-based questions that require NEWS CONTEXT to answer, not just what's visibly obvious. Your task is to analyze the provided image and generate ONE question-answer pair that: 1. Begins with "Based on the provided image, ..." but requires understanding of news context to answer 2. Cannot be answered by simply describing what's visible in the image 3. Asks about SPECIFIC FACTS, EVENTS, FIGURES, DATES, or NAMES related to the news context 4. Is DIFFERENT from existing questions Guidelines: - Questions MUST begin with "Based on the provided image, ..." - Focus on FACTUAL, VERIFIABLE information from the news context - Questions should have SINGLE, DEFINITIVE answers based on the news article - Answers must be SHORT (2-7 words), direct phrases, not complete sentences - NEVER reference or reveal answers from existing questions in your new question - Use generic references like "this person", "this building", etc. even if you know their names from previous QA pairs
User Prompt for News Context QA	TOPIC: {topic} CATEGORY: {category} NEWS DESCRIPTION (VERY IMPORTANT): {description} IMAGE DESCRIPTION: {image_description} EXISTING QUESTIONS AND ANSWERS: {json.dumps(existing_qa, indent=2)} Generate ONE NEW question-answer pair that: 1. Begins with "Based on the provided image, ..." but requires NEWS CONTEXT to answer 2. Cannot be answered by simply describing what's visible in the image 3. Relates to underlying events, significance, impacts, or context shown in the image 4. Has a direct, SHORT answer (2-7 words) 5. Is different from the existing questions Focus on the IMPLICATIONS, CONTEXT, SIGNIFICANCE or BACKGROUND of what's shown, not on obvious visual elements. Use the NEWS DESCRIPTION provided above, which is very important to generate a meaningful question that requires context.

ration of the universe.

- **Television**: Content about TV shows, streaming series, or television industry news. This encompasses reviews, ratings, cast interviews, behind-the-scenes content, and industry trends related to television networks, cable, and streaming platforms such as Netflix, Amazon Prime, or Disney+.
- **Movies**: Content related to films, movie reviews, film industry, or cinema releases. It covers the latest box-office hits, film reviews, director interviews, actor spotlights, and discussions on the impact of movies in popular culture. It also includes information about film festivals, award shows, and cinematic trends.
- **Economy**: Content about finance, markets, economic

Table 7. Prompts for Level 3 QA generation.

Task	Prompt
Diverse Question Type Generation	You are a specialized AI assistant for creating image-based questions that require NEWS CONTEXT to answer, not just what's visibly obvious. It is forbidden to give ambiguous answers. Your task is to analyze the provided image and generate ONE question-answer pair that: 1. Begins with "Based on the provided image, ..." but requires understanding of news context to answer 2. Cannot be answered by simply describing what's visible in the image 3. Asks about SPECIFIC FACTS, EVENTS, FIGURES, DATES, or NAMES related to the news context 4. Is DIFFERENT in TYPE from the existing questions CRITICAL RULE: - NEVER use ANY person names, organization names, product names, or event names that appear in ANY existing answer. - Instead of specific names, always use generic terms like "this person", "the organization", etc. QUESTION TYPES to consider (prioritize types NOT already used): - LOCATION questions (where an event happened) - TIME questions (when something occurred) - QUANTITY questions (how many, how much) - CAUSE questions (why something happened) - EFFECT questions (what resulted from an event) - COMPARISON questions (how things differ) - METHOD questions (how something was accomplished) - PURPOSE questions (the goal of an action)

Table 8. Prompts for LLM-as-a-Judge.

Task	Prompt
Answer Correctness Evaluation	You are an impartial judge evaluating if a model's response correctly answers a question. Ground Truth Answer: {gt_answer} Model Response: {model_answer} Does the model response correctly answer the question based on the ground truth? Answer with ONLY 'yes' or 'no'. Include your final answer within <answer> tags.

policies, or business trends, including the analysis of macroeconomic indicators like GDP, inflation, and unemployment rates. It also focuses on financial markets, investments, global trade, corporate strategies, and the economic implications of policy changes or technological disruptions.

- **Sports**: Content about athletic competitions, sports teams, athletes, or sporting events. This includes news on professional and amateur sports, tournament results, profiles of famous athletes, match highlights, and coverage of major sporting events like the Olympics, FIFA World Cup, Super Bowl, or the NBA Finals.
- **Theater**: Content about stage performances, plays, musicals, or theatrical productions. It includes reviews of live performances, interviews with theater professionals, trends in stage design, acting, and direction, as well as coverage of Broadway, West End, and off-Broadway productions.
- **Music**: Content about songs, musicians, concerts, albums, or the music industry. It covers the latest album releases, chart-topping songs, artist interviews, concert reviews, music awards, and trends in various genres such as pop, rock, hip-hop, classical, and electronic music.
- **Opinion**: Content expressing viewpoints, editorials, or commentary on current events. It includes opinion pieces, thought-provoking essays, and editorials on topics such as politics, culture, society, technology, and the environment, offering diverse perspectives and insights.
- **Art & Design**: Content about visual arts, exhibitions, design, fashion, or architecture. This includes news on art exhibitions, gallery openings, artist profiles, design trends, architecture innovations, and fashion movements. It also covers topics like graphic design, industrial design, and the influence of art on culture.
- **Media**: Content about journalism, publishing, social media, or news organizations. It covers trends in digital media, the role of social media platforms in news distribution, journalistic integrity, and the challenges faced by traditional media in the digital age. It also includes stories about the changing landscape of journalism, fake news, and media ethics.
- **Technology**: Content about tech innovations, gadgets, software, or digital trends. This includes updates on the latest technological advancements in fields like artificial intelligence, cybersecurity, mobile technology, and software development. It also covers the impact of emerging technologies on various industries and society as a whole.
- **Global Business**: Content about international trade, multinational corporations, or global economics. This includes insights into the global business landscape, supply chain issues, international trade agreements, and the strategies of large multinational companies. It also fo-

cuses on the economic and political challenges facing global markets.

- **Other**: Content that doesn't clearly fit into any of the above categories. This category includes miscellaneous news and topics that are relevant or important but don't fall under traditional categories. It could include lifestyle trends, regional news, niche interests, or emerging topics that don't yet have a broad classification.

11. Case Study

We select the QvQ-72B-Preview [18] and Gemini-2.0-Flash [16] models, which exhibit strong overall capabilities, for demonstration. The former showcases robust reasoning abilities , while the latter, being the most recently released model, demonstrates the strongest performance when combined with MM-Search [8]. The results compared with other models are shown in Figure 3 and 4.

The QvQ model demonstrates strong reasoning capabilities

Question: Based on the provided image, what character did the man on the right play on 'Parks & Recreation'?
Ground Truth: Jerry Gergich

Gemma-3-4b-it: Ron Swanson

Qwen2.5-VL-7B-Instruct: Tom Haverford

GPT-4o-mini: The man on the right played the character Andy Dwyer on "Parks and Recreation."

Gemma-3-27b-it: Ron Swanson

Qwen2.5-VL-3B-Instruct: Tom Haverford

Qwen2.5-VL-72B-Instruct: Leslie Knope's father

Qwen2.5-VL-32B-Instruct: Leslie Knope's father

GPT-4o: I don't know who the man on the right is based on the image provided.

QvQ-72B-Preview: Jerry Gergich.

Gemini-2.0-Flash(MM_Search): I can't answer this question.

Figure 3. The QvQ model demonstrates strong reasoning capabilities.

The search functionality enhances Gemini-2.0-Flash's performance.

Question: Based on the provided image, when will the stamp collection be available?
Ground Truth: March 27

Gemma-3-4b-it: November 21, 2023

Qwen2.5-VL-7B-Instruct: Not specified in the image

GPT-4o-mini: The stamp collection will be available on Friday, May 23.

Gemma-3-27b-it: Now/Soon

Qwen2.5-VL-3B-Instruct: 1ST

Qwen2.5-VL-72B-Instruct: 1st

Qwen2.5-VL-32B-Instruct: 24 May 2019

QvQ-72B-Preview: I can't answer this question

Gemini-2.0-Flash(MM_Search): March 27

Figure 4. The search functionality enhances Gemini-2.0-Flash's performance.