Time Matters: A Comprehensive and Versatile Benchmark for Time-Aware Visual Question Answering

Anonymous ACL submission

Abstract

Visual Question Answering is a challenging task that has received increasing attention from both the computer vision and the natural language processing communities. Given an image and a question in natural language, reasoning over the visual elements of the image and general knowledge is required to infer the correct answer. However, most existing VQA benchmarks assume that answers are static once a question is posed, whereas in real-world scenarios, answers may change over time. For example, consider the question: "***In 2021/2024**, how long has it been since the Argentina national team last won the World Cup?*" The corresponding answers would be **35 years** and **1 year**, respectively, reflecting the Argentina national team's victory in the 2023 Qatar World Cup. To address this gap, we introduce a comprehensive and versatile benchmark for **T**ime-**A**ware **V**isual **Q**uestion **A**nswering (TAVQA), a crowdsourced collection comprising 11,088 unique questions. These questions demand a broad base of common sense, world knowledge, and deep visual understanding to answer effectively. Unlike existing VQA datasets, each question in TAVQA is time-aware, meaning that the answers are dynamic and models cannot rely on outdated information. Our analysis demonstrates that the TAVQA dataset is diverse and challenging for current VQA models. We hope this dataset will enable researchers to explore new avenues of investigation in this domain.

1 Introduction

Visual Question Answering (VQA) is a multidisciplinary AI research problem [Antol et al., 2015] that engages both the computer vision (CV) and natural language processing (NLP) communities [Wu et al., 2017], which involves answering the given natural language questions based on the images. To solve VQA questions, the model must be capable of visual recognition, language understanding, basic reasoning abilities, and fundamental knowledge of the world [Schwenk et al., 2022]. Therefore, the VQA task is an ideal way to evaluate the performance of Artificial Intelligence (AI) models in real-world scene understanding [Marino et al., 2019]. In the most common form of visual question answering, the model is presented with an image and a textual question about the image [Wu et al., 2017]. It must then determine the correct answer, typically a few words or a short phrase. Traditional VQA datasets, such as VQA [Antol et al., 2015] and VQA-2 [Goyal et al., 2017] primarily assess models' capabilities in visual perception and language comprehension, whereas, in the past few years, several benchmarks like KB-VQA [Wang et al., 2015], FVQA [Wang et al., 2017], OK-VQA [Marino et al., 2019], VCR [Zellers et al., 2019], and A-OKVQA [Schwenk et al., 2022] have incorporated questions that require more commonsense reasoning and external factual knowledge to provide accurate answers. These benchmarks aim to evaluate the models' ability to understand visual and natural language content and draw upon broader knowledge sources to enhance question-answering (QA) performance.

However, as illustrated in Figure 1, due to the dynamic and ever-changing nature of real-world facts, the answer can be completely different when the time constraint in the question changes [Zhu et al., 2023]. Recent studies [Kasai et al., 2024; Zhu et al., 2023; Chen et al., 2021; Zhang and Choi, 2021; Liska et al., 2022] have challenged the static nature of traditional question-answering systems by introducing datasets that explicitly incorporate temporal context. For instance, Kasai et al. [2024] extended the scope of time-aware QA to real-time QA by developing a dynamic platform that regularly updates questions and evaluates systems; Zhu et al. [2023] demonstrates that large language models (LLMs) struggle to answer factual

questions under time constraints accurately. They propose QAaP, a novel reasoning framework that treats question-answering as a programming task by representing the question and context as code. QAaP has significantly improved performance on various time-aware question-answering datasets. However, these efforts have predominantly focused on natural language question answering, where the model's input is limited to text. In contrast, many real-world queries are inherently multimodal, often involving textual and visual components. Thus, there is a pressing need to develop a time-aware VQA dataset that requires both textual and visual processing.

To address this gap, we propose a comprehensive **T**ime-**A**ware **V**isual **Q**uestion **A**nswering dataset, which contains 11,088 samples. Each sample comprises a textual question with a time specifier and a corresponding image. Several examples are shown in Figure 5. To improve the dataset's representativeness of real-world scenarios and to enhance its diversity, we collected four distinct types of data using four different methods: Wiki, Chart, News, and LLM-Self. Each type of data requires different capabilities from the model. For example, Chart-type data demands that the model have fine-grained image processing capabilities and strong reasoning skills. In contrast, Wiki-type data requires the model to possess sufficient commonsense knowledge. To the best of our knowledge, TAVQA is the first manually curated, time-aware dataset explicitly designed to challenge VQA models across diverse, real-world scenarios. By incorporating temporal elements and various data sources, the TAVQA dataset addresses gaps in VQA and pushes the boundaries of model reasoning, commonsense knowledge, and fine-grained image analysis.

Our contributions are summarised as follows:

- **Novel Benchmark.** We introduce a novel VQA benchmark with 11,088 unique time-aware questions, making it more challenging and reflective of real-world scenarios. Additionally, we provide manually annotated answers and images for each question, enabling researchers to fine-tune their models on our dataset. We conducted an in-depth analysis of our dataset across multiple dimensions to provide a comprehensive overview. Our findings demonstrate that the TAVQA dataset exhibits substantial diversity and poses significant challenges for various VQA model capabilities.
- **Comprehensive evaluation.** We selected a diverse set of recent advanced VQA models and systematically evaluated their performance on the challenging, time-aware questions within the TAVQA dataset. We thoroughly analyzed the experimental results and identified several noteworthy findings. Furthermore, we investigated a simple yet effective method to improve the models' performance on the TAVQA dataset.

2 Related Works

Visual Question Answering. Knowledge-based datasets are pivotal in VQA research, pushing systems to employ higher-level reasoning and external knowledge for problem-solving. They serve as vital resources and benchmarks for advancing knowledge-based VQA. KB-VQA [Wang et al., 2015] evaluates VQA models' performance on higher-level knowledge questions and explicit reasoning tasks using MS COCO images [Lin et al., 2014]. FVQA [Wang et al., 2017] provides image-question pairs and includes supporting facts from external knowledge bases, which are crucial for answering visual questions. OK-VQA [Marino et al., 2019] challenges models with categories like transportation, geography, and history, demanding an understanding of complex scenarios. KVQA [Shah et al., 2019] scales up, requiring multi-entity, multi-relation, and multi-hop reasoning, prompting visual entity linking to Wikidata. R-VQA [Lu et al., 2018], built upon Visual Genome, explores complex vision-language relationships through aligned supporting relation facts.

Recent VQA research progress, fueled by large-scale pre-training, explores various subfields, emphasizing the importance of common sense and reasoning. Mathew et al. [2021] introduces a tailored dataset for VQA on document images, emphasizing improved understanding of document structure. Ding et al. [2023] expands VQA to include document element recognition, layout understanding, contextual comprehension, and critical information extraction, advancing document understanding to multi-page questioning. Li et al. [2021] enhances model robustness by iteratively collecting adversarial data, thereby exposing vulnerabilities in VQA models and improving their resilience. Schwenk et al. [2022] broadens VQA datasets, incorporating questions on commonsense and knowledge about the world, while Gao et al.

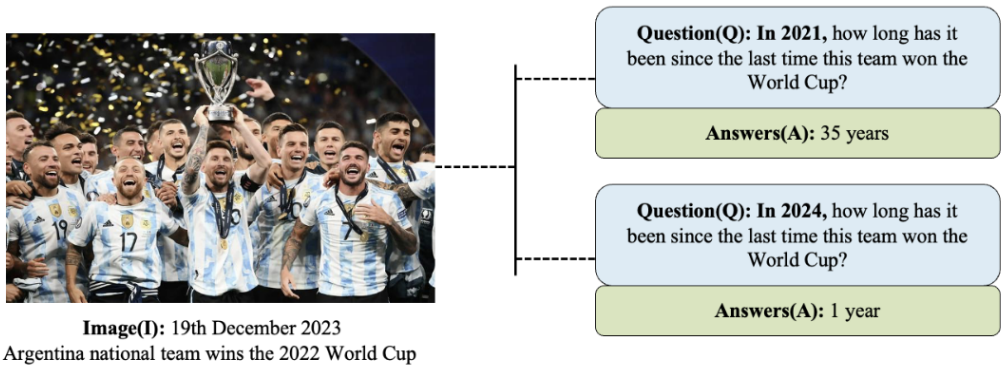


Figure 1: The answers to time-aware questions change over time.

[2022] focuses on compositional reasoning. Gao et al. [2023] offers a dataset for perceptual video enhancement, enriching video quality assessment methods. These datasets drive VQA research, tackling diverse challenges and broadening the scope of inquiry.

Time-aware Question Answering in NLP. Time-aware QA prioritizes temporal factors in reasoning and answering, meeting real-time information needs. Compared to knowledge-based visual QA, it requires models to perceive temporal changes, posing greater challenges. Approaches include CoAG [Li et al., 2024b], leveraging adaptive graph networks for long-distance time-aware questions, TCQA [Son and Oh, 2023], training models with time-context data, and QAaP [Zhu et al., 2023], using language models to convert text into structured code. Studies like TSQA [Shang et al., 2022] and TEQUILA [Jia et al., 2018b] enhance models with time estimation modules and multi-stage processing. EXAQT [Jia et al., 2021] combines techniques like cluster Steiner trees for complex temporal questions. However, time-aware QA datasets remain scarce, including UrgentQA [Liu et al., 2009], TempQuestions [Jia et al., 2018a], and WikiData-based datasets [Chen et al., 2021]. RealTime QA [Kasai et al., 2024] presents a dynamic question-answering platform that regularly releases new questions and evaluates system performance. The platform focuses on real-world, up-to-date information, requiring QA systems to respond to queries about recent events or newly emerging information.

3 TAVQA Collection

In this section, we begin by defining time-aware vision questions (Section 3.1), followed by an explanation of the methods used to collect both the question and image datasets (Sections 3.2 and 3.3). Finally, we detail the annotation process applied to the collected data (Section 3.4).

3.1 Definition of Time Awareness

A typical VQA instance can be represented as a triple (Q, I, A), where: Q is the question, I is the image related to the question, A is the answer to the question, $A(C)$ as the answer based on the conditions in C.

A VQA instance is considered time-aware if it satisfies the following conditions:

- includes a time specifier T, represented as $T = (P, t)$, where P denotes a temporal preposition (e.g., "in," "between," "before") and t corresponds to a specific temporal reference (e.g., "2019" and "2000-2010").

- Two time specifiers $T_1 = (P_1, t_1)$ and $T_2 = (P_2, t_2)$ are considered distinct if at least one of P or t is different.

- There exist at least two distinct time specifiers, T_1 and T_2, such that $A(Q, I, T_1) \neq A(Q, I, T_2)$, where A denotes the answer to the question based on the given image and time specifier.

Time-aware VQA extends the traditional VQA framework by integrating temporal context into the answer-generation process. This additional dimension requires models to comprehend the content of images and questions and recognize how the answers may evolve over time.

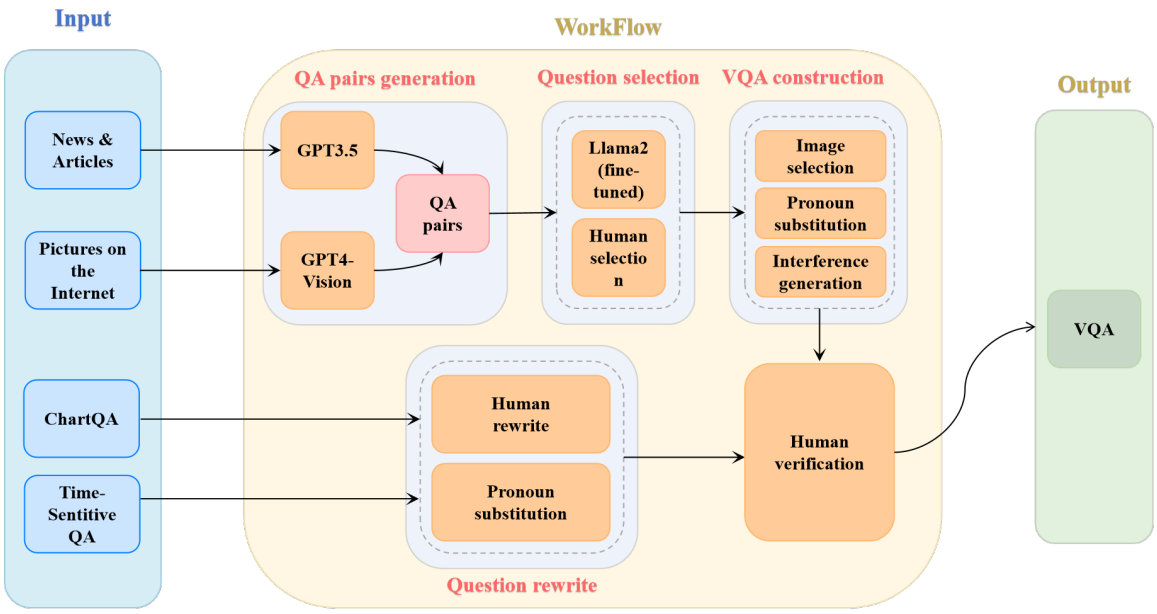


Figure 2: The processing framework flowchart for the TAVQA dataset.

3.2 Image Source

The images used in this study were primarily obtained through the Google API[1]. The process involved several steps: first, identifying replaceable nouns or concepts in the questions; second, using the Google API to search for related images; and finally, manually verifying the results.

3.3 Question Collection

To broaden the scope of knowledge in our dataset, we employ four methods of knowledge acquisition: Wikidata [Chen et al., 2021], News [Liska et al., 2022], Charts [Masry et al., 2022], and the knowledge stored within the large language model itself. Figure 2 illustrates the processing framework for our question collection.

3.3.1 News-based Question Generation

We obtained the $\{Q_i, T_i, A_i\}$ corresponding to the current news material by inputting the news material $News_i$ into ChatGPT-3.5 along with a carefully crafted prompt. In this process, we first collected a large-scale dataset of news articles[2] from 2018 to 2022 and applied a simple string-matching method to filter time-aware news items. However, more than 80% of the generated questions did not meet our criteria. To address this issue, we fine-tuned a Llama2 model [Touvron et al., 2023] using the LoRA technique [Hu et al., 2021], which was then utilized for automated question filtering. Finally, the remaining questions were manually reviewed and filtered for quality assurance.

3.3.2 LLM-Self-based Question Generation

To harness the knowledge embedded in large language models, we instructed the model to autonomously generate the required triplets $\{Q_i, T_i, A_i\}$ by inputting an image I_i directly into the model. For this, we employed the ChatGPT-4-Vision model, an advanced multimodal model demonstrating exceptional performance across various tasks. The images, categorized into Art, Literature, Technology, and Science, were sourced using the Baidu Image API[3].

3.3.3 Chart-based Question Generation

We recognize the substantial presence of time-aware information in graphical data. In parallel, Masry et al. [2022] introduced a chart-based QA dataset derived from real-world data. We manually adjusted the original QA questions to align with the time-aware vision questions definition (see Section 3.1) by filtering the chart data to isolate time-aware information. These modified questions were then incorporated into our dataset, contributing to our comprehensive data collection process.

3.3.4 Wiki-based Question Generation

This method is primarily inspired by the RealTime QA dataset [Kasai et al., 2024] and other Wikidata-

[1] https://images.google.com/
[2] http://data.statmt.org/news-crawl/README
[3] https://images.baidu.com/

based datasets to create time-aware VQA samples. Chen et al. [2021] programmatically extracted time-related entries from Wikidata and generated numerous questions with answers that change over time using pre-constructed question templates. We developed the required TAVQA dataset by manually refining the questions and incorporating relevant images into our dataset.

3.3.5 Distractors Generation

For the multiple-choice questions, we used ChatGPT-3.5 to generate the corresponding distractors. We iteratively refined the prompt to ensure the quality of the generated options. Finally, we manually sampled 300 items and confirmed that 95% met our criteria.

3.4 Question Annotation

We enlisted 40 annotators, all master's students in computer science, with a solid understanding of English and general knowledge. Their language skills and expertise ensured they were well-equipped to perform data labelling tasks accurately.

3.4.1 Annotation Criteria

We have developed a comprehensive set of annotation procedures and standards to ensure accurate data labelling by annotators. The primary tasks for data annotators are outlined as follows:

Noun Replacement in Questions Annotators are tasked with identifying and replacing specific nouns within the questions, including names of people, objects, countries, and other entities.

Image Selection Annotators are responsible for selecting suitable images from Google Image search results to provide visual context for the questions. If no appropriate image is available, the question is to be discarded.

Question Verification Annotators must ensure the accuracy of time-aware vision-question-answer pairs by verifying whether the visual information and question together enable correct inference using external knowledge. If not, they may adjust the question or answer. If the issue persists, the question is discarded.

4 Dataset Statistics

After data collection through the four methods, the TAVQA dataset is split into 8K/1K/2K for train, validation, and testing, as shown in Table 2. Figure 5 shows examples of different types in the TAVQA dataset. Each data instance contains a well-defined time specifier, a question in natural language, a related image, four-option candidates, and the ground truth answer.

Question and Answer Statistics. We first analyze the diversity of questions and answers in our dataset by calculating the unique words. As shown in Table 1, the TAVQA dataset contains 11,088 unique questions, 5,847 unique question words, and 7,532 unique answer words, indicating a variety of questions and answers.

Candidate Options Distribution. We further investigate the variety of candidates in our dataset. Figure 3 illustrates a long-tail distribution of candidate options, standard in open-labelled image tasks [Schwenk et al., 2022]. Only a limited set of options appear repeatedly, predominantly numeric values (e.g., 2014 and 2015). Most of the candidates in the TAVQA dataset belong to the long tail of the distribution, which indicates the variety of our dataset.

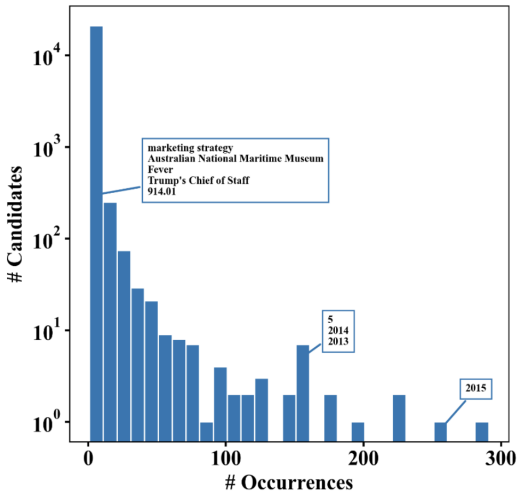


Figure 3: Candidate option distribution. Long-tail distribution observed.

Comparison with Other Datasets. In Table 1, we present the properties and statistics of the TAVQA dataset compared to other related datasets. While some datasets contain more questions than ours, this is expected, as our goal is to ensure that every question is time-aware (Section 3.1). Consequently, we filtered out a substantial number of entries during the data collection process that did not meet this criterion. Despite prioritizing this essential time-aware characteristic, we made efforts to keep the size of our dataset comparable to other datasets.

Table 1: Comparison of various knowledge-based VQA datasets. Data marked * are from Schwenk et al. [2022], while those marked with + are from our own statistics. Q: question, I: image, A: answer, DA: Direct Answer, MC: Multiple Choice.

	Q	I	Knowledge type	Time Awareness	Answer type	Avg. length (Q/A)	Unique words (Q/A)
KB-VQA*	2,402	700	Fixed KB	✗	DA	6.8/2.0	530/1,296
FVQA*	5,826	2,190	Fixed KB	✗	DA	9.5/1.2	3,010/1,287
OK-VQA*	14,055	14,031	Factoid	✗	DA	8.1/1.3	5,703/11,125
KVQA+	183,007	24,602	Fixed KB	✗	DA	7.4/1.6	8,497/20,897
RVQA+	198,889	60,473	common/world	✗	DA	4.4/1.4	5,909/6,671
A-OKVQA*	24,903	23,692	common/world	✗	DA/MC	8.8/1.3	7,248/17,683
TAVQA	11,088	7,096	common/world	✓	DA/MC	7.6/2.9	5,847/7,532

Table 2: TAVQA dataset partition

	News	Wiki	Chart	LLM-Self	Total
Train	3456	3368	1029	235	8088
Valid	438	428	103	31	1000
Test	831	852	262	55	2000

As illustrated in Figure 5, the model must demonstrate different capabilities when addressing various questions in TAVQA. The model requires a detailed understanding of images and strong reasoning abilities for Chart-type questions. In contrast, Wiki-type questions are primarily factoid (e.g., querying event dates from specific years). Therefore, the knowledge type in TAVQA emphasizes general reasoning skills and focuses on the model's ability to effectively leverage commonsense knowledge of the world.

The average question length in the TAVQA dataset is shorter than that of FVQA, OK-VQA, and A-OKVQA but longer than that of KB-VQA, KVQA, and RVQA. This is expected, as most time-aware questions require detailed descriptions, resulting in longer questions. Furthermore, our dataset's average length of correct answers is significantly higher than that of other datasets, as time-aware answers often need to convey more detailed information (e.g., "China plans to become carbon-neutral in 2050.").

To examine the prevalence of time-aware information in existing VQA datasets, we randomly sampled 1,000 questions from widely used datasets. Leveraging both large language model annotations and human annotations, we systematically analyzed the distribution of such information (see Section 3.1). As illustrated in Table 3, the average proportion of time-aware questions across these datasets is less than 2%, underscoring a notable deficiency in the representation of time-aware information in current VQA datasets. Please refer to Appendix C for more details.

Table 3: Time-aware question count in widely used VQA datasets. We sample 1K entries from each dataset and employ ChatGPT-3.5 for the automatic annotation.

Dataset	# GPT Positive	# Annotator Positive
FVQA	2	0
KB-VQA	10	0
OK-VQA	36	11
A-OKVQA	23	8

5 Benchmarking

In this section, we evaluate advanced Multi-modal Large-scale Language Models (MLLMs) in a zero-shot setting and report their performance on our TAVQA dataset.

5.1 Evaluation

TAVQA primarily frames the VQA task as a multiple-choice problem, requiring the model to select the correct option from a set of provided candidates. Accordingly, accuracy is reported as the primary evaluation metric. However, since MLLMs generate free-form answers, we also report their BLEU scores using the correct answer as a reference to provide a more comprehensive evaluation of the models' performance.

5.2 Setup

Selected Models. We select recently released advanced models, including BLIP2 series models [Li et al., 2023], InstructBLIP series models [Dai et al., 2023], LLaVA-Onevision [Li et al., 2024a], Phi-3.5-vision [Abdin et al., 2024], Llama-3.2-vision [Meta, 2024], Qwen-VL [Bai et al., 2023] and GPT-4o-mini [OpenAI, 2024].

Hyperparameters. The checkpoints for the open-sourced MLLMs are all obtained from HuggingFace[4] and executed using the Transformers[5] Python library. Please refer to Appendix D for further details regarding model configurations.

5.3 Main Results

Table 4 presents the test performance of models, with detailed statistics on the test performance within each knowledge category. We observed:

TAVQA is difficult for most of the VQA models. The overall average accuracy of the models is **33.68%**, with the best-performing model achieving an accuracy of **50.05%**. This performance is significantly lower compared to their results on other benchmarks. The BLIP2-opt-6.7b cannot answer more than 10 Chart-type questions correctly out of 262. This indicates that solving TAVQA requires more than developing a sophisticated model; it necessitates methods that effectively integrate temporal information.

Larger models do not consistently demonstrate superior performance. Although BLIP2-opt-6.7b is significantly larger than BLIP2-flan-t5-xl (6.7B vs 3B parameters), its performance is markedly inferior, achieving only 26.49% of the latter's total accuracy. Notably, BLIP2-opt-6.7b struggles with processing chart images. This underscores the need for more sophisticated approaches to handling visual inputs beyond increasing model size. Most current VQA models exhibit poor performance in the Wiki and Chart types. The average accuracies of the VQA models tested on four types (LLM-Self, Wiki, Chart, and News) are **37.53%**, **29.17%**, **33.23%**, and **38.15%**, respectively. Notably, the models' performance on Wiki and Chart types approaches random selection (expected at 25%). We hypothesize that Wiki-type questions involving time specifiers, such as *"From 1803 to 1804, where was his workplace?"* (with a photo of Justinus Kerner) increases the challenge, as answering requires historical knowledge and the ability to reason through many candidates. This is more difficult than questions without time constraints, where multiple correct answers may exist. Chart-type questions often require identifying specific numbers at given times, but models may answer randomly due to pre-training on image captioning and fine-tuning for conversation, which limits their ability to interpret chart images. For instance, questions like *"Between 2010 and 2015, in which year does the blue bar show a value of 2.5?"* are particularly difficult.

5.4 Effectiveness of Time Specifier

We further investigate the extent to which time specifiers contribute to answering questions. This analysis excludes time specifiers from the questions, while all other parameters remain constant, consistent with the main experiment. As illustrated in Figure 4 (refer to Figure 14 for details), the performance of all models utilizing our manually labelled and well-defined time specifiers improved significantly, ranging from **6.16%** (InstructBLIP-flan-t5-xxl) to **10.79%** (GPT-4o-mini). This demonstrates that the time specifiers in our dataset play a crucial role in enhancing the accuracy of models when addressing time-aware questions.

5.5 In-Context Learning Experiment

As In-Context Learning (ICL) has proven effective in improving model performance across various downstream tasks, we are interested in whether providing the model with a few examples can improve its performance in the TAVQA benchmark. We conducted this experiment by showing the models different numbers of examples in the prompt template. Figure 4 reports the experimental results. It can be observed that:

A single demonstration significantly enhances the models' performance. On average, the performance of almost all models improved by **7.33%** and **7.31%** in the settings with and without time specifiers, respectively, after being provided with an example. The GPT-4o-mini model exhibited the most substantial improvement, with a gain of **26.53%** in the time specifier setting. Notably, it also improved by **20.42%** in the setting without time specifiers. These findings indicate that carefully chosen examples can improve model performance on challenging downstream tasks, with stronger models showing greater improvement. More demonstrations can harm models' performance. By providing an additional example (two in total) to the models, the average performance increased only by **0.28%** and decreased by **1.02%** in the two respective settings. Further increasing the number of examples to three resulted in even more remarkable performance degradation, with decreases of **3.69%** and **5.87%**. This decline may be attributed to the introduction of additional noise with more demonstrations, potentially making it

[4] https://huggingface.co/

[5] https://github.com/huggingface/transformers

Table 4: Zero-shot Multiple Choice experiment results. "#" represents "the number of". The Llama-3.2-vision model exhibits a low BLEU score but a high accuracy, as it generates additional sentences for answers (e.g., "The correct answer is..."). While the choices are correct, the extra text leads to a low BLEU score. The best performance is highlighted in **boldface**.

	# Correct / # Total					
Model	LLM-Self	Wiki	Chart	News	Total	BLEU
BLIP2-opt-2.7b	13/55	181/853	55/262	162/831	411/2000	0.1875
BLIP2-opt-6.7b	3/55	89/853	7/262	58/831	157/2000	0.1186
BLIP2-flan-t5-xl	16/55	212/853	67/262	324/831	619/2000	0.2603
BLIP2-flan-t5-xxl	21/55	242/853	72/262	335/831	670/2000	0.2778
InstructBLIP-vicuna-7b	16/55	208/853	80/262	248/831	552/2000	0.2247
InstructBLIP-vicuna-13b	20/55	236/853	72/262	312/831	640/2000	0.2554
InstructBLIP-flan-t5-xl	17/55	213/853	66/262	306/831	602/2000	0.2539
InstructBLIP-flan-t5-xxl	23/55	238/853	76/262	335/831	672/2000	0.2818
LLaVA-Onevision	23/55	309/853	74/262	376/831	782/2000	0.3026
Phi-3.5-vision	24/55	313/853	141/262	375/831	853/2000	0.3482
Llama-3.2-vision	19/55	**352/853**	93/262	430/831	894/2000	0.0958
GPT-4o-mini	29/55	306/853	122/262	395/831	852/2000	0.3631
Qwen-VL-2b	26/55	255/853	119/262	324/831	724/2000	0.3026
Qwen-VL-7b	**39/55**	329/853	**175/262**	**458/831**	**1001/2000**	**0.3960**

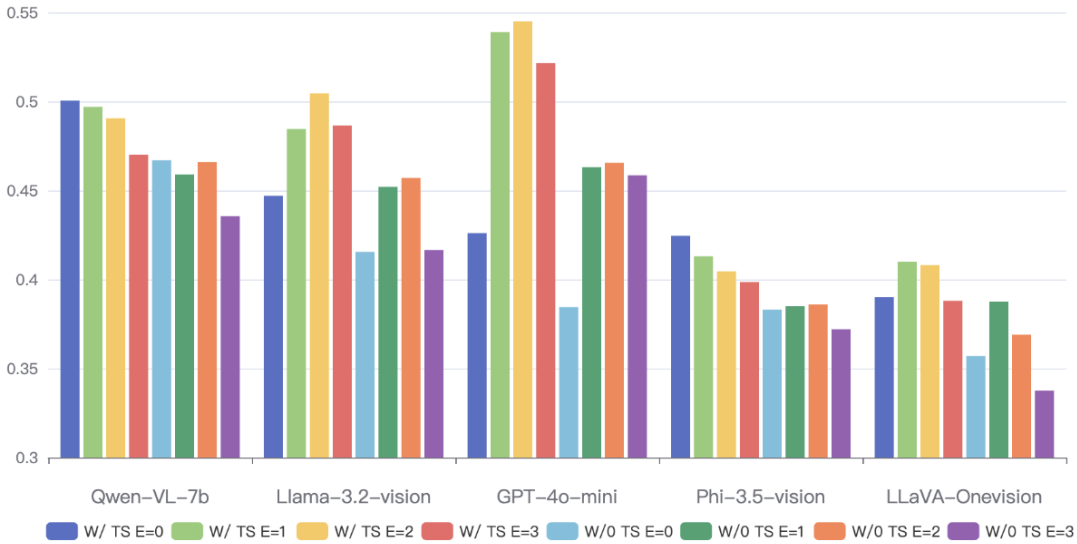


Figure 4: In-Context Learning Experiment results, with total accuracy as the metric. "W/" and "W/O" denote the presence and absence of time specifiers, respectively, while "E=" represents the number of provided examples.

more difficult for the models to identify essential information, thereby leading to performance degradation.

6 Conclusion

Although many benchmarks assess different aspects of VQA models, none specifically focus on time-aware visual questions commonly encountered in real-world situations. To fill this gap, we have created a new dataset, TAVQA, consisting mainly of questions with changing answers. A time specifier accompanies each question to narrow down the temporal context of the answer. We have evaluated a selection of recently released MLLMs and conducted a comprehensive analysis of the experimental results. While these models show promising performance on our dataset, they still have significant limitations in addressing time-aware questions, indicating substantial room for improvement.

7 Limitations

We have identified two areas of our dataset that warrant refinement. First, the temporal information in TAVQA is limited. For each question, only a single time specifier was collected; however, in real-world scenarios, time-aware questions may involve multiple valid time specifiers, each corresponding to different answers. Second, TAVQA lacks the rationale or reasoning steps that connect the question to the answer, posing potential challenges for model training.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. *arXiv preprint arXiv:2108.06314*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Yihao Ding, Siwen Luo, Hyunsuk Chung, and Soyeon Caren Han. 2023. Vqa: A new dataset for real-world vqa on pdf documents. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 585–601. Springer.

Difei Gao, Ruiping Wang, Shiguang Shan, and Xilin Chen. 2022. Cric: A vqa dataset for compositional reasoning on vision and commonsense. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5561–5578.

Yixuan Gao, Yuqin Cao, Tengchuan Kou, Wei Sun, Yunlong Dong, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. 2023. Vdpve: Vqa dataset for perceptual video enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1474–1483.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018a. Tempquestions: A benchmark for temporal question answering. In *Companion Proceedings of the The Web Conference 2018*, pages 1057–1062.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018b. Tequila: Temporal question answering over knowledge bases. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 1807–1810.

Zhen Jia, Soumajit Pramanik, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Complex temporal question answering on knowledge graphs. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 792–802.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, Kentaro Inui, and 1 others. 2024. Realtime qa: What's the answer right now? *Advances in Neural Information Processing Systems*, 36.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Jitong Li, Shaojuan Wu, Xiaowang Zhang, and Zhiyong Feng. 2024b. A context-enhanced adaptive graph network for time-sensitive question answering. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Linjie Li, Jie Lei, Zhe Gan, and Jingjing Liu. 2021. Adversarial vqa: A new benchmark for evaluating the robustness of vqa models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2042–2051.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, D'Autume Cyprien De Masson, Tim Scholtes, Manzil Zaheer, Susannah Young, and 1 others. 2022. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models. In *International Conference on Machine Learning*, pages 13604–13622. PMLR.

Yandong Liu, Nitya Narasimhan, Venu Vasudevan, and Eugene Agichtein. 2009. Is this urgent? exploring time-sensitive information needs in collaborative question answering. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 712–713.

Pan Lu, Lei Ji, Wei Zhang, Nan Duan, Ming Zhou, and Jianyong Wang. 2018. R-vqa: learning visual relation facts with semantic attention for visual question answering. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1880–1889.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. 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.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Meta. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/. Accessed: 2024-10-04.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence. Accessed: 20 September 2024.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8876–8884.

Chao Shang, Guangtao Wang, Peng Qi, and Jing Huang. 2022. Improving time sensitivity for question answering over temporal knowledge graphs. *arXiv preprint arXiv:2203.00255*.

Jungbin Son and Alice Oh. 2023. Time-aware representation learning for time-sensitive question answering. *arXiv preprint arXiv:2310.12585*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427.

Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. 2015. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570*.

Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163:21–40.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731.

Michael JQ Zhang and Eunsol Choi. 2021. Situatedqa: Incorporating extra-linguistic contexts into qa. *arXiv preprint arXiv:2109.06157*.

Xinyu Zhu, Cheng Yang, Bei Chen, Siheng Li, Jian-Guang Lou, and Yujiu Yang. 2023. Question answering as programming for solving time-sensitive questions. *arXiv preprint arXiv:2305.14221*.

A Examples of rejected questions

To ensure that all questions in the TAVQA dataset are genuinely time-aware, we eliminated approximately 80% of the instances during the processing phase outlined in Section 3.3. The majority of these instances were excluded for one of two reasons: (1) the answer to the question was static, and (2) the question allowed for an excessive number of correct answers. Examples of commonly rejected instances are presented in Figure 10.

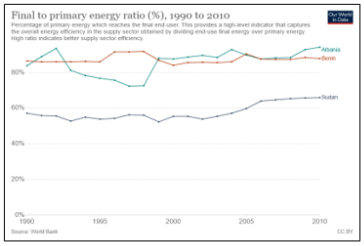
LLM Self

Question:
In which museum is this painting displayed ?
Time Specifier:
in 2023
Direct Answer:
Louvre Museum
Multiple Choice Answers:
(a) National Gallery (b) British Museum
(c) Louvre Museum (d) Metropolitan Museum of Art

News

Question:
How much has the number of customers purchasing a combination of new and old products on this platform increased ?
Time Specifier:
between 2016 and 2017
Direct Answer:
12%
Multiple Choice Answers:
(a) 15% **(b) 12%**
(c) 9% (d) 6%

Chart



Question:
In which year the final to primary energy ratio of Albania is the largest?
Time Specifier:
between 2001 and 2009
Direct Answer:
2004
Multiple Choice Answers:
(a) 2004 (b) 2003
(c) 2006 (d) 2008

Wiki

Question:
Which team did the player belong to?
Time Specifier:
from 1998 to 2001
Direct Answer:
VfB Lübeck
Multiple Choice Answers:
(a) FC St. Pauli (b) Karlsruher SC
(c) Hannover 96 **(d) VfB Lübeck**

Figure 5: Four different types of data in the TAVQA dataset from four sources (Section 3.3). LLM-Self: Generated using ChatGPT-4. News: Generated using the News. Chart: Generated based on the ChartQA dataset. Wiki: Generated based on the Time-aware QA dataset.

B Temporal sensitivity analysis in the existing dataset

In Table 3, we analyze the prevalence of time-aware questions in existing VQA datasets (FVQA, KB-VQA, OK-VQA, and A-OKVQA) based on evaluations by both large language models and human annotators. In this section, we provide examples from these datasets along with the rationale for the decisions, as illustrated in Figure 6, 7, 8, and 9.

C More statistics of TAVQA dataset

C.1 Term Frequency

We are particularly interested in identifying the question terms and answer terms that are most distinctive within each category to better understand the types of questions and responses present across these types. To accomplish this, we compute the relative frequency of terms by comparing their frequency within a specific category against their overall occurrence in the entire dataset. This method highlights the terms with the highest relative frequency in each type of question and answer. As shown in Figure 11, question and answer terms common in the "News" and "Wiki" categories, such as "help," "nationality," and "character," exhibit significant overlap. This suggests that models tasked with these categories may require enhanced commonsense reasoning and extensive world knowledge. In contrast, the "Chart" category presents distinctly different terms in questions and answers. Most questions in this type involve specific nouns, while the corresponding answers are primarily numerical. Thus, the model must comprehend the visual content—such as charts, as illustrated in Figure 5—and demonstrate advanced reasoning capabilities. These results underscore the diversity of our dataset, both in terms of questions and answers, which necessitates a wide range of model competencies to address the variety of tasks effectively.

C.2 Time range distribution

We are also interested in the time range of the time specifiers in our dataset, as it plays a crucial role in answering questions. Figure 13 presents the statistical results. The time specifiers span from 1900 to 2080, covering 180 years, further highlighting the TAVQA dataset's diversity. However, a substantial portion of the time points fall between 2000 and 2024, indicating that most questions and answers focus on the modern era.

C.3 Time specifiers distribution

Figure 12 shows the proportions of different time specifiers in our dataset. The vast majority of time specifiers are the first three ("in," "between A and B," and "from A to B."), which aligns with human common sense, as time-aware questions in daily life are typically expressed using these three forms. Other time specifiers primarily include "after," "before," "at," and "on."

D Additional Details for Multi-modal Large-scaled Language Models

We selected various recently released MLLMs and evaluated their performance in the TAVQA dataset. To ensure reproducibility, we manually set the random seed to 42 for all experiments, max new tokens of generation to 40, and the temperature to 0 for all models.

D.1 Description

- **BLIP2 Series Models.** BLIP2 utilizes off-the-shelf, pre-trained vision models with frozen parameters alongside large language models that maintain frozen parameters. It introduces a lightweight Querying Transformer to bridge the vision encoder and the text encoder. The model is trained on a dataset of 129 million images and achieves state-of-the-art (SOTA) performance across a range of multimodal tasks, including visual question answering (VQA) and image captioning.

- **InstructBLIP Series Models.** InstructBLIP incorporates instruct-tuning into the field of multi-modal large-scaled language models. Building on the BLIP2 framework enhances the model's ability to follow user instructions more accurately and effectively.

- **LLaVA-Onevision.** LLaVA is trained on 158K unique language-image instruction-following instances and 595K image-text pairs, enabling the model to follow user instructions in a multi-turn interaction. Initially, the vision and text encoder parameters are frozen, and a lightweight projector is trained to bridge the two. Subsequently, the vision encoder remains frozen while the projector and pre-trained language model (LLM) are fine-tuned. LLaVA-OneVision introduces an improved vision representation approach that

Dataset: FVQA **Question: Tell me the name of the cosmetics shown in this image?** Answer: lipstick	Dataset: FVQA **Question: Which object in this image has a tail?** Answer: kite	Dataset: FVQA **Question: What object in this image is commonly eaten for lunch?** Answer: sandwich	Dataset: FVQA **Question: Which food in this image is designed in Italian?** Answer: pizza
Decision by human: **No.** The answer is static and do not change over time. Decision by ChatGPT: **No.** The type of cosmetic product in a specific image does not change over time.	Decision by human: **No.** The answer is consistent, based solely on the image's visual content. Decision by ChatGPT: **No.** The answer will remain consistent regardless of when the question is asked, as it relies solely on the visual content of the image.	Decision by human: **No.** The answer solely relies on the image's content. Decision by ChatGPT: **No.** The answer is based on the visual content of the image and does not change over time.	Decision by human: **No.** The answer depends on the image's content, not the question's timing. Decision by ChatGPT: **No.** The answer depends on the visual content of the image rather than on the timing of the question.

Figure 6: Examples of the time-aware question analysis in FVQA. (Section 4).

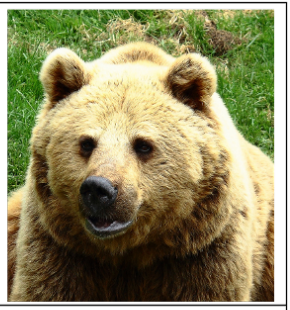
			
Dataset: KB-VQA **Question: what is the animal in the middle** Answer: giraffe	Dataset: KB-VQA **Question: how many peoples in this image** Answer: 2	Dataset: KB-VQA **Question: tell me the largest object** Answer: people in the middle	Dataset: KB-VQA **Question: what is the color of the biggest object in this image** Answer: brown
Decision by human: **No.** The animal's identification relies on its appearance, which is unchanging. Decision by ChatGPT: **No.** The identification of the animal in the image is based on its physical appearance, which does not change over time.	Decision by human: **No.** The answer relies on the image's content, so the number of people remains constant. Decision by ChatGPT: **No.** The answer is based on the visual content of the image, which does not change over time.	Decision by human: **No.** The answer depends on the image's content. Decision by ChatGPT: **No.** The answer is based solely on the visual content of the image, and the relative size of objects within the image does not change over time.	Decision by human: **No.** The answer depends on the image's content, with the largest object's color unchanged. Decision by ChatGPT: **No.** The answer relies on the visual content of the image, specifically the color of the largest object, which does not change over time.

Figure 7: Examples of the time-aware question analysis in the KB-VQA. (Section 4).

Dataset: OKVQA **Question: Is the boy swimming or doing another water activity?** Answer (First one): another activity	Dataset: OKVQA **Question: Which breed of dog it this?** Answer (First one): doberman	Dataset: OKVQA **Question: What objects is this person using to help them ski?** Answer (First one): poles	Dataset: OKVQA **Question: What would the red bottle next to the sink probably contain?** Answer (First one): dish soap
Decision by human: **No.** The answer relies solely on the image's content. Decision by ChatGPT: **No.** The answer depends solely on the visual content of the image and does not change over time.	Decision by human: **No.** The solely answer relies on the dog's appearance. Decision by ChatGPT: **No.** The answer is based on the dog's physical appearance in the image and does not change over time.	Decision by human: **No.** The answer depends solely on the image's visual content. Decision by ChatGPT: **No.** The answer is based entirely on the visual content of the image and does not change over time.	Decision by human: **No.** The answer relies on the image's context. Decision by ChatGPT: **No.** The answer is based on the visual context in the image and typical usage patterns, which do not change over time.

Figure 8: Examples of the time-aware question analysis in the OK-VQA. (Section 4).

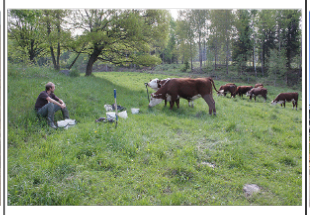
			
Dataset: A-OKVQA **Question: What is the occupation of the person driving?** Answer: farmer	Dataset: A-OKVQA **Question: What mood do the cows seem to be in?** Answer: happy	Dataset: A-OKVQA **Question: What is the man putting on the bus?** Answer: bow	Dataset: A-OKVQA **Question: When did the namesake of this theater die?** Answer: 2009
Decision by human: **No.** The interpretation remains constant, based on the image's content. Decision by ChatGPT: **No.** This interpretation does not change over time, as it relies on the visual content of the image.	Decision by human: **No.** The answer solely relies on observing the cows' demeanor Decision by ChatGPT: **No.** The answer might involve subjective observation of the cows' body language or demeanor, which doesn't change over time.	Decision by human: **No.** The answer relies on the image's content, which remains unchanged. Decision by ChatGPT: **No.** The answer depends on the visual content of the image, which remains the same regardless of when the question is asked.	Decision by human: **No.** The event is a fixed point in history, so the answer remains unchanged. Decision by ChatGPT: **No.** The event (the death of the theater's namesake) has already occurred and is a fixed point in history, the answer will not change over time.

Figure 9: Examples of the time-aware question analysis in dataset A-OKVQA. (Section 4).

Question: Who travelled to this place?	**Question: Which country is set to become the first in the world to print a warning on this?**	**Question: In an interview with Closer Weekly, he said he regrets not using his birth name professionally. What is it?**	**Question: How many suitcases are in the picture?**
Candidates: A. Elon Musk B. Oprah Winfrey C. Angela Merkel D. Mark Zuckerberg	Candidates: A. United Kingdom B. Australia C. New Zealand D. Canada	Candidates: A. Ramon Estévez B. Nicolas Coppola C. Michael Mann D. Michael Douglas	Candidates: A. 1 B. 2 C. 3 D. 4
Rejected reason: There are too many correct answers to this question.	Rejected reason: The answer to this question is fixed.	Rejected reason: A person's birth name does not change over time.	Rejected reason: The answer to this question is fixed.

Figure 10: Some examples of rejected instances during our question collection stage (Section 3.3).

Category	Highest Relative Frequency Question words	Highest Relative Frequency Answer words
LLM-Self	Coal, DHL, Kyoto, COP3, Skype	company, IPO, Machu, Picchu, telescope
News	character, tournament, raised, nearly, help	Snow, Cup, Liberty, seven, fold
Chart	line, bigger, Regular, Casual, Agriculture	44068, 1930, 1940, 1974, 54132
Wiki	occupant, nationality, employer, spouse, educated	Belgrade, Attack, Valparaiso, Female, governors

Figure 11: Highest relative frequency across different types of the TAVQA dataset.

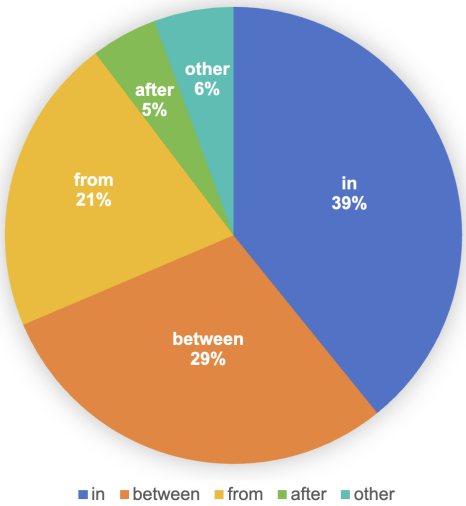


Figure 12: Time specifiers distribution. The top three time specifiers are "in," "between A and B, "and "from A to B." A and B denote a specific time, respectively.

enables the model to process multiple images and videos more effectively.

- **Phi-3.5-vision.** Phi-3.5-Vision is a 4.2-billion-parameter model derived from phi-3.5mini. It was pre-trained on a large-scale text-image pair dataset and post-trained on multi-modal instruction-tuning datasets. It demonstrates strong performance on reasoning tasks and is proficient at processing both single-image and multi-image, as well as text, prompts.

- **GPT-4o-mini.** GPT-4o mini supports text and vision processing and is equipped with knowledge up to October 2023. This is an API-only model.

- **Llama-3.2-vision.** The architecture of Llama-3.2-vision is based on the combination of Llama 3.1 LLMs, combined with a vision tower and an image adapter. The text models used are Llama 3.1 8B for the Llama 3.2 11B Vision model. It was trained on a massive dataset of 6 billion image-text pairs, with a diverse mix of data. This makes it an excellent candidate for fine-tuning downstream tasks.

- **Qwen-VL.** Qwen2-VL can handle arbitrary image resolutions, mapping them into a dynamic number of visual tokens, offering a more human-like visual processing experience. It decomposes positional embeddings into components to capture 1D textual, 2D visual, and 3D video positional information, thereby enhancing its multimodal processing capabilities. Qwen2-VL achieves state-of-the-art performance on visual understanding benchmarks.

D.2 Checkpoints URLs

We list the source of checkpoints of open-sourced models below:

- **BLIP2-opt-2.7b:** : https://huggingface.co/Salesforce/blip2-opt-2.7b
- **BLIP2-opt-6.7b:** : https://huggingface.co/Salesforce/blip2-opt-6.7b
- **BLIP2-flan-t5-xl:** : https://huggingface.co/Salesforce/blip2-flan-t5-xl
- **BLIP2-flan-t5-xxl:** : https://huggingface.co/Salesforce/blip2-flan-t5-xxl
- **InstructBLIP-vicuna-7b:** : https://huggingface.co/Salesforce/instructblip-vicuna-7b
- **InstructBLIP-vicuna-13b:** : https://huggingface.co/Salesforce/instructblip-vicuna-13b
- **InstructBLIP-flan-t5-xl:** : https://huggingface.co/Salesforce/instructblip-flan-t5-xl
- **InstructBLIP-flan-t5-xxl:** : https://huggingface.co/Salesforce/instructblip-flan-t5-xxl
- **LLaVA-Onevision:** : https://huggingface.co/lmms-lab/llava-onevision-qwen2-7b-ov-chat

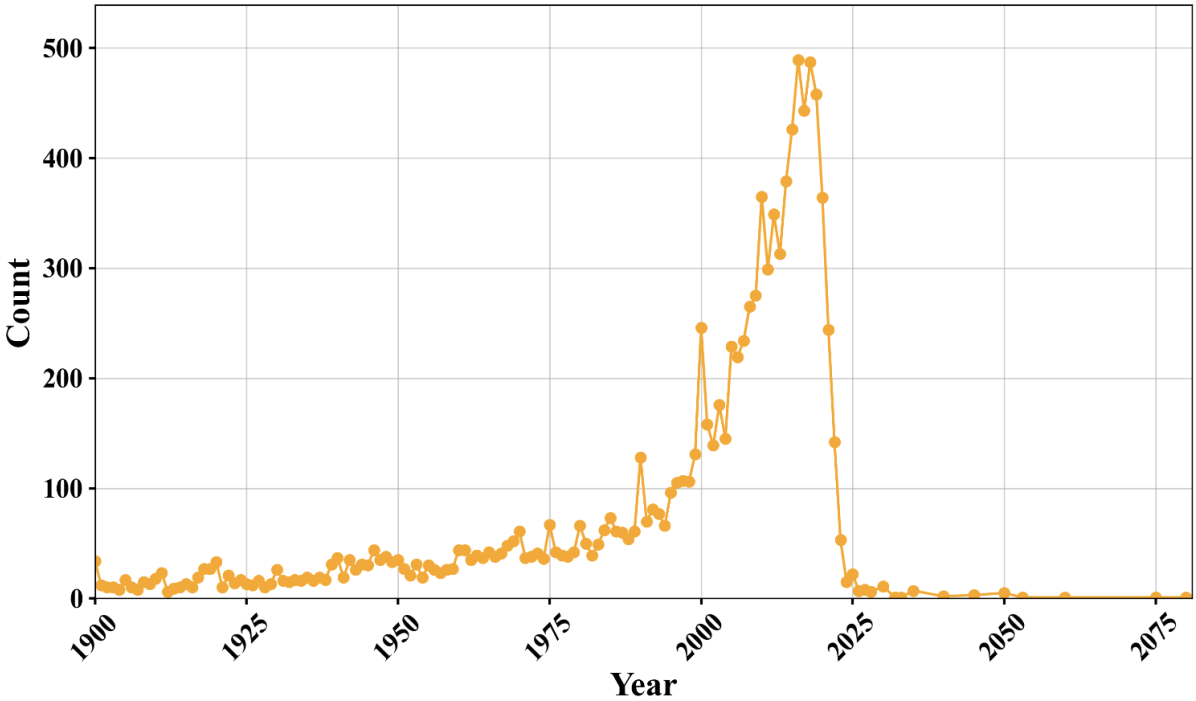


Figure 13: The time range of time specifiers in TAVQA. The time point information in time specifiers is mostly between 2000 and 2025.

- **Phi-3.5-vision:** : https://huggingface.co/microsoft/Phi-3.5-vision-instruct
- **Llama-3.2-vision:** : https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct
- **Qwen-VL-2b:** : https://huggingface.co/Qwen/Qwen2-VL-2B-Instruct
- **Qwen-VL-7b:** : https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct

E Additional Experiment

E.1 Effectiveness of Visual Modality

We further investigate the extent to which visual inputs contribute to answering time-aware visual questions. In this ablation study, we remove the image input from the original setting and keep only the textual questions, while all other experimental conditions (models, prompting strategy) remain unchanged, consistent with the main experiment. Specifically, for BLIP2-flan-t5-xxl and InstructBLIP-flan-t5-xxl, which require an image as an input, we feed a blank image in the w/o-image setting. As shown in Figure 15, the performance of all multimodal models drops noticeably when the image is removed, ranging from **0.4%**(GPT-4o-mini) to *12.0%*(Qwen-VL-7b). This trend indicates that visual information plays a critical role in the TAVQA dataset.

E.2 The Bias Analysis Of Answer For Temporal Ranges

We analyze the bias of answer for temporal ranges by evaluating model performance in different time ranges. The result is shown in Figure 16. Each range contains 10 time-sensitive questions, which are randomly sampled from the TAVQA dataset. These results reveal a clear recency bias: most models achieve higher accuracy on recent periods than on earlier decades. For example, performance in the 2021+ range is consistently higher than that in 1900–1949 for all models, with the performance gap ranging from 2 (Qwen-VL-7B) to 5 (Qwen-VL-7B, GPT-4o-mini) correctly answered questions. This trend indicates that current vision–language models are more strongly aligned with contemporary visual and factual knowledge, while their ability to reason about historical images is limited. Larger models such as Qwen-VL-7B show relatively more stable performance across time ranges, whereas smaller models exhibit stronger temporal skew. Overall, these results demonstrate that temporal bias remains a significant challenge for time-aware visual question answering.

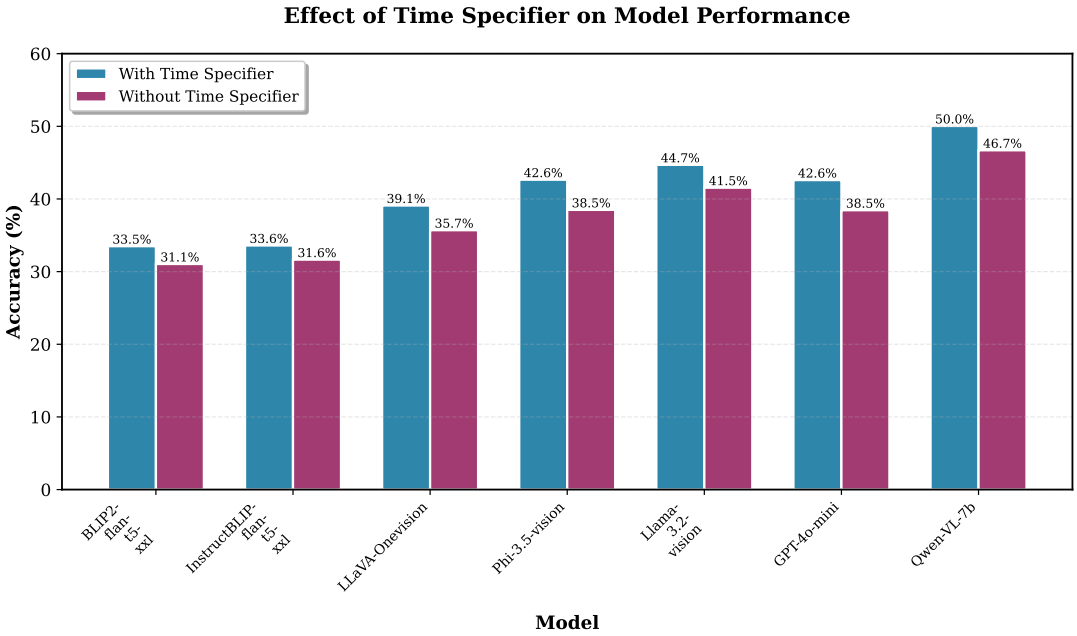


Figure 14: Models answering questions with and without Time Specifier (TS) as described in Section 5.4. w/ and w/o denote with and without, respectively. This experiment is conducted in a zero-shot manner.

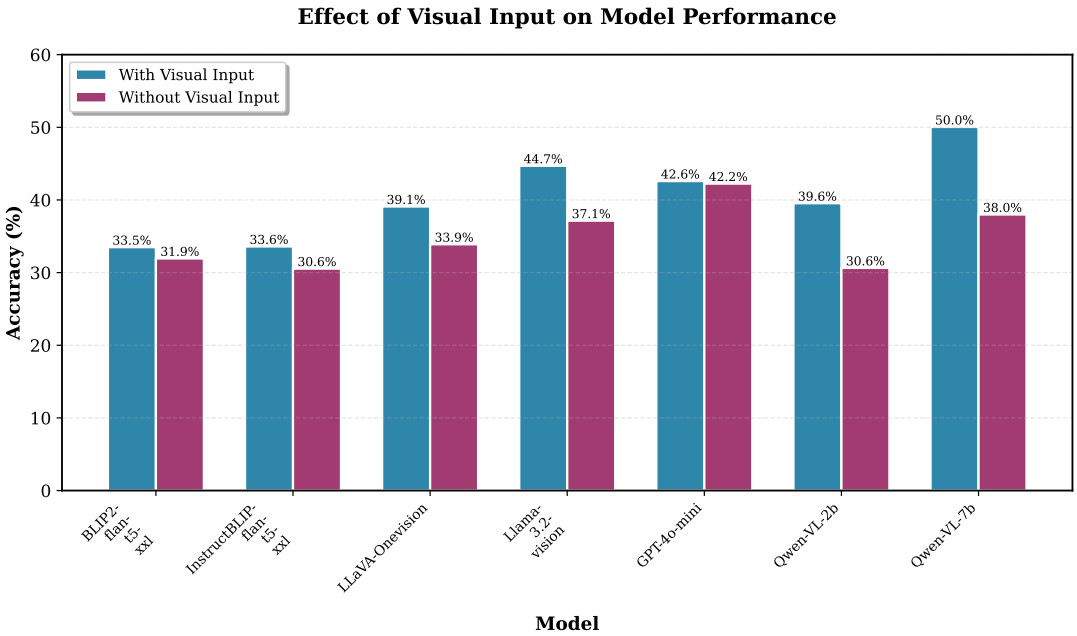


Figure 15: Models answering questions with and without Visual Inputs (VI). w/ and w/o denote with and without, respectively. This experiment is conducted in a zero-shot manner.

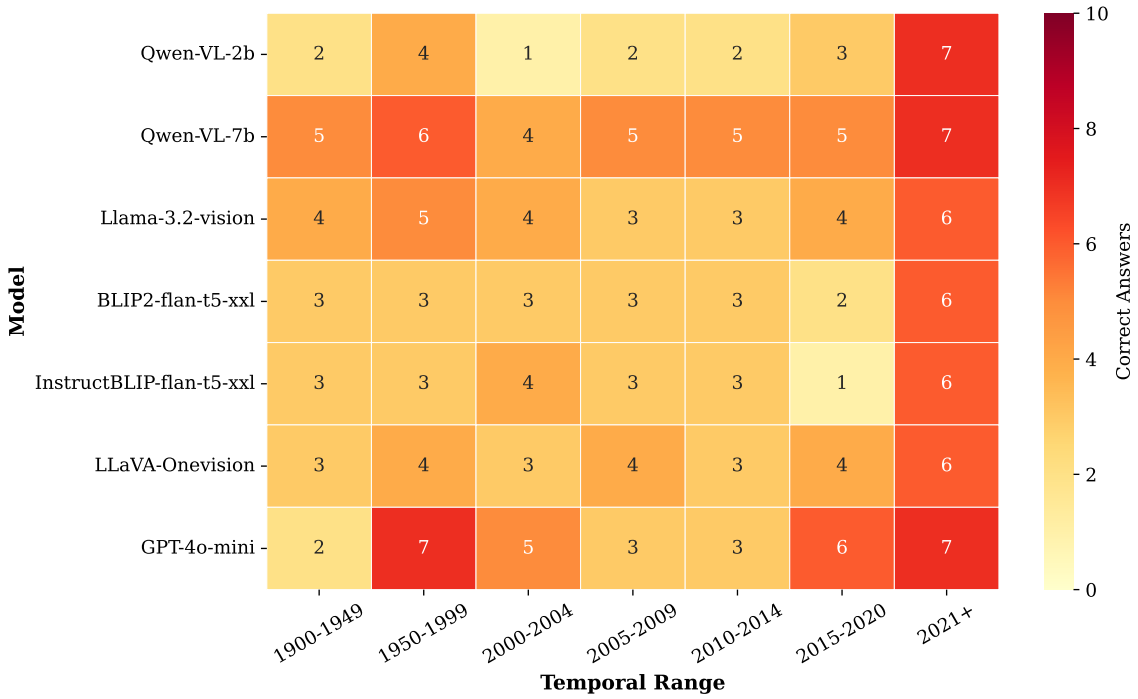


Figure 16: The performance of multiple vision-language models across various temporal ranges. Each range consists of ten questions, and the numerical entries represent the total count of correct answers.

Please answer this VQA multiple choice question. You must choose the answer you think is the most correct from the Candidates and give it to me, **you only have to give me one Candidate answer! **

Question:

in 2021, how long has it been since the last time this team won the World Cup?

Candidates:

["35 years", "1 year", "10 years", "5 years"]

Answer:

Figure 17: Construction of samples inputted into the model for testing.