A. Fine-tuning Datasets

During our fine-tuning stage, VLMs are optimized using a variety of datasets with different tasks. These datasets, introduced in Table 2 , include InfographicVQA [25], Kleister Charity [31], WikiTableQuestions [28], VizWiz-VQA [10], and ST-VQA [4], and are briefly described as follows:

InfographicVQA (InfoVQA) [25]: This dataset is a collection of over five thousand infographic images, along with a large number of question-answer pairs. These infographics are sourced from various web domains and feature diverse layouts and designs. The InfographicVQA challenges vision language models to interpret and reason over complex visual documents, often necessitating understanding of graphical elements, data visualization, reasoning, and arithmetic skills.

Kleister Charity (KLC) [31]: This dataset consists of annual financial reports from UK charity organizations. The task involves key information extraction (KIE) such as charity names, addresses, charity numbers, and reporting dates. Primarily comprising scanned documents, this dataset poses challenges due to its length, diverse layout, and the necessity to interpret both text and structural features.

WikiTableQuestions (WTQ) [28]: This dataset includes question and answer pairs collected from thousands of HTML tables extracted from Wikipedia. The questions are designed to be complex, requiring multi-step reasoning and various data operations such as comparison, aggregation, and arithmetic computation.

VizWiz-VQA (VizWiz) [10]: Comprising a large number of question-answer pairs, this dataset features images captured by blind individuals using mobile phones and spoken questions about those images. Unique in terms of its image quality, which is often blurred, and the nature of its questions, a significant portion of the images are unanswerable due to poor image quality.

ST-VQA [4]: Designed specifically for understanding textual information within natural images, this dataset requires models to read and interpret scene text to accurately answer questions. It includes a large collection of images sourced from various public datasets such as COCO-text, Visual Genome, and ImageNet, challenging models to comprehend images across a wide range of scenarios and textual appearances within images.

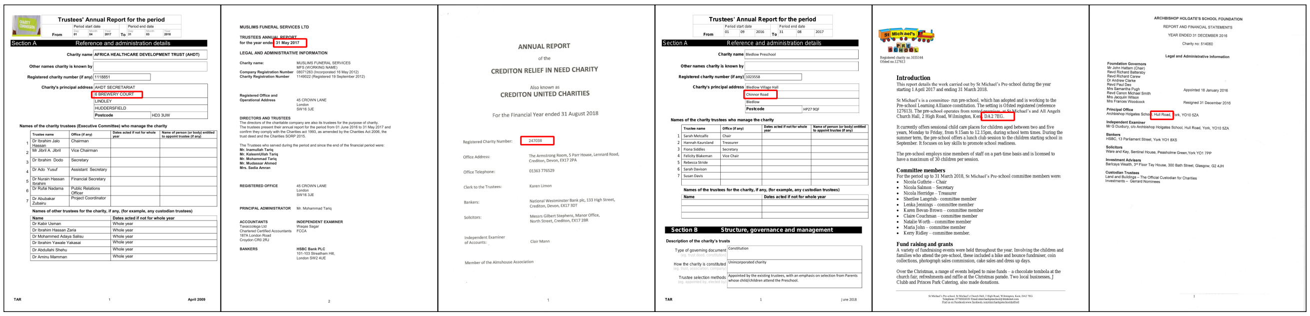
B. More Qualitative Results

Figure 5 presents additional quantitative results derived from various evaluation datasets comparing QA-ViT method [9] and our QID method, implemented with Qwen-VL-Chat [3]. These results highlight the effectiveness of our method in enhancing the vision model's ability to identify relevant visual cues and improve comprehension in both text-rich and natural scene environments.

Furthermore, we outline the limitations of our approach in Figure 6. Although our method aids in enhancing understanding of text-rich images, it does not significantly improve the model's reasoning and arithmetic capabilities. Consequently, our future research will focus on refining the model's ability to perform complex reasoning tasks more effectively in dense-text settings.

C. Broader Impact

The enhanced capabilities of vision-language models (VLMs) offer substantial promise for improving document comprehension in environments with extensive textual content. However, the interaction between question embeddings and vision representations remains relatively unexplored. Our approach encourages this interaction with limited fine-tuning samples while preserving the structural integrity of pre-trained VLMs. It also minimizes the necessity for extensive retraining, thereby reducing the computational resources required for deploying sophisticated AI solutions. Additionally, our method's efficiency with limited data can decrease both the time and costs involved in annotating large datasets, enhancing the accessibility and affordability of advanced document understanding technologies. We encourage the research community to further explore and adopt our QID for text-intensive tasks, anticipating significant benefits in various applications.



Question: What is the value for the address?
Qwen-VL-Chat (QA-ViT): 10
Qwen-VL-Chat (QID): 6 Brewery Court

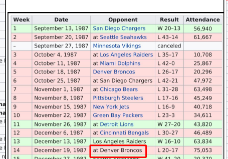
Question: What is the value of reported date?
Qwen-VL-Chat (QA-ViT): 2017-09-30
Qwen-VL-Chat (QID): 2017-05-31

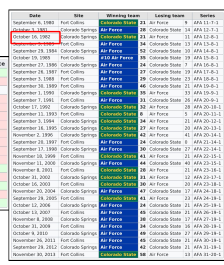
Question: What is the value of charity number?
Qwen-VL-Chat (QA-ViT): 1176045
Qwen-VL-Chat (QID): 247038

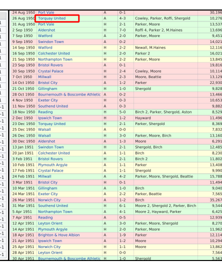
Question: What is the address indicated?
Qwen-VL-Chat (QA-ViT): Bledlow Village Hall
Qwen-VL-Chat (QID): Chinnor Road

Question: What is the postal code?
Qwen-VL-Chat (QA-ViT): WR13 7JG
Qwen-VL-Chat (QID): DA2 7EG

Question: What is the value of address?
Qwen-VL-Chat (QA-ViT): 10
Qwen-VL-Chat (QID): Hull Road







Question: What was the total number of points by Franco Pellizotti?
Qwen-VL-Chat (QA-ViT): 20
Qwen-VL-Chat (QID): 15

Question: What date was the only game played on soldier field?
Qwen-VL-Chat (QA-ViT): 10 October 27
Qwen-VL-Chat (QID): October 23

Question: Which manager was previous to George Case?
Qwen-VL-Chat (QA-ViT): Irvin Noren
Qwen-VL-Chat (QID): Bob Lemon

Question: Which opponent of the Kansas city chiefs in 1987 saw a total of more than 70,000 in attendance?
Qwen-VL-Chat (QA-ViT): San Diego Chaargers
Qwen-VL-Chat (QID): Denver Broncos

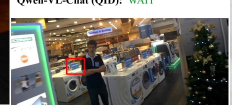
Question: What date comes just after October 3, 1981 on the list?
Qwen-VL-Chat (QA-ViT): October 10, 1982
Qwen-VL-Chat (QID): October 16, 1982

Question: Which team did they face next after their first game against port vale?
Qwen-VL-Chat (QA-ViT): Walsall
Qwen-VL-Chat (QID): Torquay United

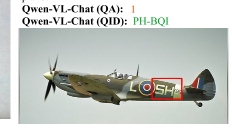
Question: What was the name of the last episode of the season?
Qwen-VL-Chat (QA-ViT): The first time
Qwen-VL-Chat (QID): Home Again!

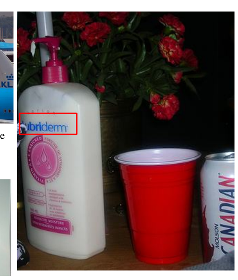
Question: What is the location of the next festival?
Qwen-VL-Chat (QA-ViT): TBA, Hawaii
Qwen-VL-Chat (QID): Tumon, Guam, USA

Question: What word is illuminated on the street sign?
Qwen-VL-Chat (QA-ViT): LOOK
Qwen-VL-Chat (QID): WAIT



Question: What number is written on the blue plane?
Qwen-VL-Chat (QA): 1
Qwen-VL-Chat (QID): PH-BQI





Question: What is written above the white square?
Qwen-VL-Chat (QA-ViT): AQUITAINE
Qwen-VL-Chat (QID): REGION

Question: What is the visible brand name of the product?
Qwen-VL-Chat (QA-ViT): VIRTU
Qwen-VL-Chat (QID): VITU

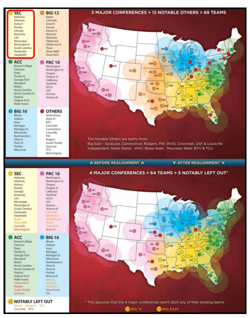
Question: What is the brand of washing machines behind the man?
Qwen-VL-Chat (QA-ViT): Indesit
Qwen-VL-Chat (QID): Samsung

Question: What is the word in red text on the front of the can?
Qwen-VL-Chat (QA-ViT): Campbell's
Qwen-VL-Chat (QID): Tomato

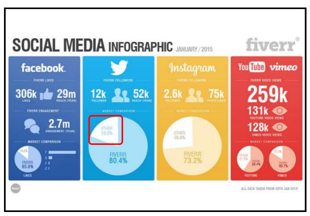
Question: What letters come after the red and blue target circle?
Qwen-VL-Chat (QA-ViT): k12
Qwen-VL-Chat (QID): SH

Question: What is the brand of the lotion?
Qwen-VL-Chat (QA-ViT): Eubriderm
Qwen-VL-Chat (QID): Lubriderm

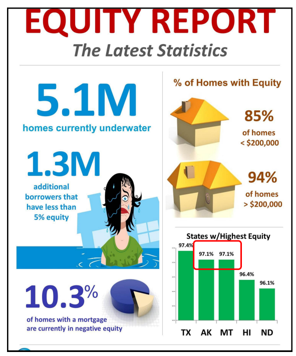
Figure 5. More qualitative results between between QA-ViT and our QID with Qwen-VL-Chat model. Image regions with answers are highlighted.



Question: How many teams did SEC conference have before re-alignment?
Qwen-VL-Chat (QID): 14
Ground Truth: 12



Question: What is the percentage market share others have in comparison to Fiverr in Twitter?
Qwen-VL-Chat (QID): 80.4%
Ground Truth: 19.6%



Question: How many states have 97.1% equity?
Qwen-VL-Chat (QID): 3
Ground Truth: 2

Iteration	Year	Dates	Location	Theme
1st	1972[10]	6 May-20 May	Suva, Fiji	"Preserving culture"
2nd	1976[11]	6 March-13 March	Rotorua, New Zealand	"Sharing culture"
3rd	1980[12]	30 June-12 July	Port Moresby, Papua New Guinea	"Pacific awareness"
4th	1985[13]	29 June-15 July	Tahiti, French Polynesia	"My Pacific"
5th	1988[14]	14 August-24 August	Townsville, Australia	"Cultural interchange"
6th	1992[15]	16 October-27 October	Rarotonga, Cook Islands	"Seafaring heritage"[16]
7th	1996[17]	8 September-23 September	Apia, Sāmoa	"Unveiling treasures"
8th	2000[18]	23 October-3 November	Nouméa, New Caledonia	"Words of past, present, future"[19]
9th	2004[20]	22 July-31 July	Koror, Palau	"Nurture, Regenerate, Celebrate"[21]
10th	2008[22]	20 July-2 August	Pago Pago, American Samoa	"Threading the Oceania *'Ula*"
11th	2012	1–14 July	Honiara, Solomon Islands	"Culture in Harmony with Nature"[23]
12th	2016	TBA	Tumon, Guam	"TBA"
13th	2020[24]	TBA	TBA, Hawaii	"TBA"

Question: What is the average length of the festival as of 2012?
Qwen-VL-Chat (QID): 10 years
Ground Truth: 14 days

Figure 6. Failure cases of QID on documents and questions require arithmetic and reasoning skills. Image regions with answers are highlighted.