CASP: Compression of Large Multimodal Models Based on Attention Sparsity
-Supplementary Material-

Mohsen Gholami, Mohammad Akbari, Kevin Cannons, and Yong Zhang
Huawei Technologies Canada Co., Ltd.
{mohsen.gholami1, mohammad.akbari, kevin.cannons, yong.zhang3}@huawei.com

This supplementary material includes the computational complexity analysis, further numerical experiments, an ablation study on the calibration dataset size, and qualitative results. We also discuss the limitations and broader impact of this work.

1. Computational Complexity

In this section, we present the computational complexity analysis of CASP compared to the baselines. Tab. 1 shows the results on LLaVA-Next-Video-7B (8 frames) [25] with "Eager" attention, a batch size of 1, and a maximum/minimum new token count of 128. We provide the prefilling time in seconds and throughput in tokens per second (Tok/s). Additionally, we report the prefilling peak memory, end-to-end peak memory, and model size.

Note that the quantization procedure generally involves two criteria that can affect the inference time: 1) Matrix multiplication of low-precision tensors, which is often faster than float tensors. 2) Dequantization at the inference stage to FP16, which introduces overhead. Tab. 1 shows the inference time of AQLM [2] and QuIP# [16] compared with the original model in FP16. Comparing QuIP# and AQLM, QuIP# is faster via fusing query, key, and value weight matrices in the attention layer and fusing gate and up weight matrices in the MLP layer.

CASP contains two components that impact the inference time: 1) Low-rank factorization of W_q and W_k. 2) Quantizing important layers to higher bits (e.g., 3-bit). Compressing W_q and W_k via low-rank decomposition (i.e., removing a high percentage of eigenvalues from the Q and K weights) directly reduces FLOPs, making inference faster. In other words, regardless of the hardware and kernel design, low-rank factorization always provides run-time improvement as most of the parameters are removed. As the second row of Tab. 1 shows, $\text{CASP}_{\text{Original}}$, i.e., the FP16 model with 75% compression of W_q and W_k, results in nearly 4% speed-up due to smaller weight matrices.

On the other hand, quantizing important layers to higher bits may introduce overhead compared to uniformly quantizing all layers to 2-bit. This is because 3-bit quantized models are slightly slower than the 2-bit ones [2, 16]. Overall, CASP does not introduce any overhead for the baselines. In some cases such as $\text{CASP}_{\text{AQLM}}$, it can slightly improve the inference speed due to the low-rank factorization. It should be noted that our primary goal in this work is not to achieve faster inference over the baselines but to enhance their performance with the same model size, memory, and inference time.

Method	Bit	Prefill Time (s)	Throughput (Tok/s)	Prefill Peak-Mem (GB)	End-to-End Peak-Mem (GB)	Model Size (GB)
Original	16	0.41	2.2	13.5	13.6	13.5
$\text{CASP}_{\text{Original}}$	16	0.39	2.3	12.0	12.1	13
AQLM	2	0.51	1.8	3.2	3.4	2.7
$\text{CASP}_{\text{AQLM}}$	2	0.50	1.9	3.1	3.3	2.7
QuIP#	2	0.39	2.3	3.2	3.4	2.7
$\text{CASP}_{\text{QuIP\#}}$	2	0.39	2.3	3.4	3.6	2.7

Table 1. Runtime and memory usage of the baselines and CASP. CASP does not introduce any overhead compared to the baselines.

Tab. 1 also compares the prefilling and end-to-end peak memory of CASP with the baselines. For a fair comparison, we matched the model size of CASP with the baselines, ensuring all 2-bit quantized checkpoints are 2.7GB. CASP's peak memory is slightly higher than the baseline due to optimal bit allocation. This peak memory is influenced by the higher bits allocated to important layers and the extent of low-rank compression applied to W_q and W_k.

2. Further Quantitative Results

In the main manuscript, the experimental results on 5 multichoice QA datasets for image-language understanding were reported. In this section, Tab. 2 presents additional quantitative results on image captioning datasets such as NoCaps [1], COCO-Caption [10], and Flickr30K [21], as well as GQA [6]. The primary evaluation metric used for open QA and image captioning tasks is CIDEr (Consensus-based Image Description Evaluation) [17], which measures the similarity between a generated caption and a set of refer-

LLaVA-1.5-7B	Bit	NoCaps (CIDEr↑)	COCO17 (CIDEr↑)	Flick30K (CIDEr↑)	GQA (EM ↑)	Avg. Rel Imp.
Original	16	0.102	0.106	0.74	0.61	
GPTQ	2.2	0.53	0.62	0.38	0.13	
CASP$_{GPTQ}$	2.2	**0.92**	**0.100**	**0.64**	**0.52**	+125%
AQLM	2	0.73	0.87	0.57	0.43	
CASP$_{AQLM}$	2	**0.91**	**0.107**	**0.68**	**0.53**	+22%
QUIP#	2	0.102	0.103	0.75	0.57	
CASP$_{QuIP\#}$	2	**0.102**	0.102	**0.77**	**0.57**	+0.5%

Table 2. Further quantitative results on open-ended QA tasks and GQA dataset with LLaVA-1.5-7B.

Calibration Size	LiveB (PPL↓)	LWilder (PPL↓)	LCOCO (PPL↓)	Avg. (PPL↓)
128	7.8	9.0	5.9	7.5
256	7.8	8.5	5.8	7.3
512	7.9	8.3	5.7	7.3
1024	7.9	8.2	5.7	7.2

Table 3. Experiment on the calibration data size using CASP$_{AQLM}$ with LLaVA-1.5-7B.

ence captions. As summarized in Tab. 2, CASP obtains 125% and 22% average relative improvements over GPTQ and AQLM. QUIP# almost obtains the same results as the FP16 model and even outperforms the FP16 model in the Flickr30K dataset. However, we still observe 0.5% average relative improvements with CASP$_{QuIP\#}$.

3. Calibration Dataset Size

Tab. 3 demonstrates experiments on the number of samples in the calibration dataset used for CASP$_{AQLM}$ with LLaVA-1.5-7B. We observe slight performance improvements with increasing the calibration size from 128 samples to 1024 samples. Although increasing the size of the calibration dataset improves the overall performance of the model, it also increases the cost and time of the calibration and optimization procedure for quantization and low-rank factorization.

4. CASP and KV Cache Quantization

KV cache compression has emerged as a critical technique to optimize memory efficiency in large language models by reducing the size of the key-value cache used during inference. One recent method for KV cache quantization is KIVI [13], which achieves significant reductions in storage requirements while preserving model performance. On the other hand, CASP focuses on weight-only compression, targeting the model's parameters to achieve similar efficiency gains. These two approaches are orthogonal, meaning they operate on different components of the model and can be combined to further enhance overall compression.

As KIVI and CASP are orthogonal methods, we have combined them. Tab. 5 demonstrates the results on TruthfulQA (BLEU Score↑) using Llama2-7B as the base model. KV cache is quantized to 2 bits and model weights are quantized to 2.2 bits (on average). As seen, CASP$_{GPTQ}$+KIVI offers a significant improvement over GPTQ+KIVI.

5. CASP vs. Low-Rank Decomposition

Applying simple low-rank decomposition to ALL weight matrices results in significantly worse performance than CASP. This is because only W_q and W_k are low-rank in LMMs and LLMs. Tab. 6 shows the results of CASP with SOTA low-rank decomposition methods SVD-LLM [18] and MoDeGPT [9] under extreme compression. We use LLama2-7B as the base model and report perplexity (PPL↓) on the Wikitext dataset.

6. Further Analysis on Bit Allocation

The optimal bit allocations returned by our method are typically non-integer. To ensure simplicity and compatibility across various quantization techniques, we rounded these values to integers. Calculating exact non-integer average bits for each layer would require modifying the codebook to accommodate non-predefined values for techniques such as AQLM and QuIP#. This adjustment, however, would necessitate the creation of new kernels for decoding during inference—one kernel for each layer. While using non-integer bits could potentially yield better results, exploring this avenue is left as future work.

In our experiments, we computed the optimal bit allocation for each individual layer in the model. However, since adjacent layers often share similar levels of importance, we investigated the possibility of sharing bit allocations across adjacent layers. Specifically, we tested shared optimal bit allocations for every three layers on LLaVA-1.5-7B. This approach resulted in only a negligible reduction of 0.7 seconds in overall computation time, which is insignificant compared to the total quantization times: 40 minutes for GPTQ, 2 hours for QuIP#, and 6 hours for AQLM.

7. Datasets, Tasks, and Metrics

We briefly introduced the 8 image-language and 5 video-language datasets used in the experiments of the main manuscript. In addition, the system prompt (instruction) used to get output results for each dataset was given. Similar to the experiments on LLMs, when measuring perplexity we do not provide any system prompt [3]. The details of datasets used for image-language and video-language understanding tasks are presented in Tab. 4, which also includes the extra 4 datasets discussed in Section 2.

As shown in the table, diverse range of tasks including image captioning, visual reasoning, open-ended visual

	Dataset	Task	Metric	System Prompt
Image-Language	COCO-2017 [10]	Image Captioning	CIDEr	Provide a one-sentence caption for the provided image.
	Flicker30k [21]	Image Captioning	CIDEr	Provide a one-sentence caption for the provided image.
	GQA [6]	CE-VQA	Eaxct Match	Answer the question using a single word or phrase.
	MMBench [12]	MC-VQA	Accuracy	Answer with the option's letter from the given choices directly.
	MME [4]	CE-VQA	Perception Score	Answer the question using a single word or phrase.
	LiveBench [19]	OE-VQA	PPL	N/A
	LLaVA-Bench-Wilder [7]	OE-VQA	PPL	N/A
	LLaVA-Bench-COCO [7]	Image Captioning	PPL	N/A
	MMU [23]	CE-VQA,OE-VQA	Accuracy	Answer with the option's letter from the given choices directly, OR Answer the question using a single word or phrase.
	Nocaps [1]	Image Captioning	CIDEr	Provide a one-sentence caption for the provided image
	ScienceQA-Image [14]	Visual reasoning	Exact Match	Answer with the option's letter from the given choices directly.
	SeedBench-Image [8]	MC-VQA	Accuracy	Answer with the option's letter from the given choices directly.
Video-Language	ActivityNet [22]	CE-VQA	Accuracy/ GPT-Assisted score	Answer the question using a single word or phrase.
	VideoChatGPT-temporal [15]	OE-VQA	Rouge, PPL, and GPT-Assisted scores	Evaluate the temporal accuracy of the prediction compared to the answer.*
	VideoDetailCaption [7]	OE-VQA	Rouge, PPL, and GPT-Assisted scores	N/A
	VideoMME (VMME) [5]	MC-VQA	Accuracy	Answer with the option's letter from the given choices directly.
	NextQA [20]	CE-VQA	WUPS	Answer a question using a short phrase or sentence.

Table 4. Details of the datasets, the corresponding tasks, metrics, and prompts used in our experiments. CE-VQA: Closed-Ended Visual Question Answering, OE-VQA: Open-Ended Visual Question Answering, MC-VQA: Multiple-Choice Visual Question Answering. *: Only the main sentence from the prompt is shown here.

Base	Base+KIVI	GPTQ	GPTQ+KIVI	$\text{CASP}_{\text{GPTQ}}$	$\text{CASP}_{\text{GPTQ}}$+KIVI
26.0	21.6	5.0	2.8	**23.5**	**11.4**

Table 5. CASP combined with KV cache quantization.

80% Compression (≈3.2 bits)		87% Compression (3 bits)		
SVD-LLM	ModeGPT	$\text{CASP}_{\text{GPTQ}}$	$\text{CASP}_{\text{AQLM}}$	$\text{CASP}_{\text{QuIP\#}}$
276.4	245.8	21.8	8.5	8.1

Table 6. CASP vs. low-rank decomposition methods.

question answering, closed-ended visual question answering, and multiple-choice visual question answering are used to evaluate the performance of the baseline methods compared with ours. Note that the system prompts are the default prompts provided in the lmms-evals evaluation package [24].

8. Qualitative Results

In this section, we provide qualitative results from LiveBench [19], COCO-Caption [10], and LLaVA-Bench-Wilder [11] datasets.

LiveBench includes screenshots from news web pages, with multiple questions asking for details about each image. Fig. 1 and 2 show two randomly chosen examples from this dataset. Below each image, we display the responses from LLaVA-1.5-7B (FP16), baselines (GPTQ, AQLM, and QuIP#), and CASP. Each response is scored by GPT-4o out of 10. CASP consistently improves the baseline responses by approximately 1.5 points.

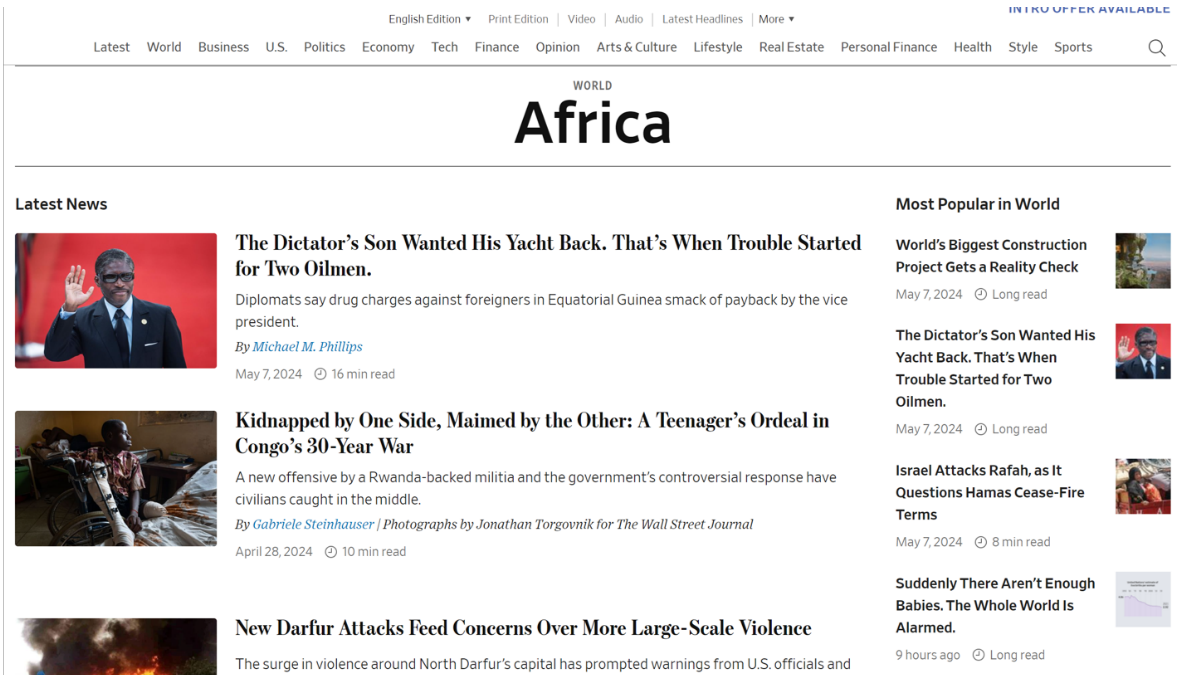
Fig. 3 and 4 present two samples from the COCO-Caption dataset, which includes images with multiple short captions for each image. This task is generally easier compared to LiveBench. We observe consistent improvements in responses by CASP, with an average increase of 2.6 points. In Fig. 3, $\text{CASP}_{\text{QuIP\#}}$ addresses the redundancy in QuIP#'s answer by including most of the important elements in the picture. In Fig. 4, a major element, "Man hangs off the side of the motorcycle," is overlooked by both the FP16 model and quantized models. However, $\text{CASP}_{\text{QuIP\#}}$ eliminates unnecessary information from the FP16 response (e.g., "A backpack can be seen…"). Comparing the responses of QuIP# and $\text{CASP}_{\text{QuIP\#}}$, the latter adds important aspects such as "the motorcycle is leaning over" and "the rider is leaning into the turn."

Fig. 5 and 6 are from LLaVA-Bench-Wilder. The questions are complex and include "memes" that require the model to understand indirect meanings in the pictures. $\text{CASP}_{\text{QuIP\#}}$ scores are equal to or better than the FP16 model in these examples. Overall, these qualitative results show the effectiveness and superiority of CASP compared to the baselines in terms of basic understanding and addressing important details in the images.

9. Limitations and Future Work

This work has some limitations that need to be addressed in future research. The low-rank factorization method used in this work is not quantization-friendly, leading to more outliers in the factorized matrices compared to the original weight matrices. Addressing this issue could improve CASP's results in future work.

We also observe that the extreme compression regime applied in CASP decreases accuracy for samples with small images and complex questions, as there is less redundancy in the attention. Providing a dynamic rank selection for such cases, similar to the dynamic visual token of LLaVA-1.6 could address this problem. In this study, we presented results without fine-tuning the quantized models. Future research should explore efficient layer-wise fine-tuning to further enhance the performance of quantized models combined with low-rank factorization.



English Edition ▾ | Print Edition | Video | Audio | Latest Headlines | More ▾

INTRO OFFER AVAILABLE

Latest World Business U.S. Politics Economy Tech Finance Opinion Arts & Culture Lifestyle Real Estate Personal Finance Health Style Sports

WORLD

Africa

Latest News

The Dictator's Son Wanted His Yacht Back. That's When Trouble Started for Two Oilmen.

Diplomats say drug charges against foreigners in Equatorial Guinea smack of payback by the vice president.

By Michael M. Phillips

May 7, 2024 16 min read

Kidnapped by One Side, Maimed by the Other: A Teenager's Ordeal in Congo's 30-Year War

A new offensive by a Rwanda-backed militia and the government's controversial response have civilians caught in the middle.

By Gabriele Steinhauser | Photographs by Jonathan Torgovnik for The Wall Street Journal

April 28, 2024 10 min read

New Darfur Attacks Feed Concerns Over More Large-Scale Violence

The surge in violence around North Darfur's capital has prompted warnings from U.S. officials and

Most Popular in World

World's Biggest Construction Project Gets a Reality Check

May 7, 2024 Long read

The Dictator's Son Wanted His Yacht Back. That's When Trouble Started for Two Oilmen.

May 7, 2024 Long read

Israel Attacks Rafah, as It Questions Hamas Cease-Fire Terms

May 7, 2024 8 min read

Suddenly There Aren't Enough Babies. The Whole World Is Alarmed.

9 hours ago Long read

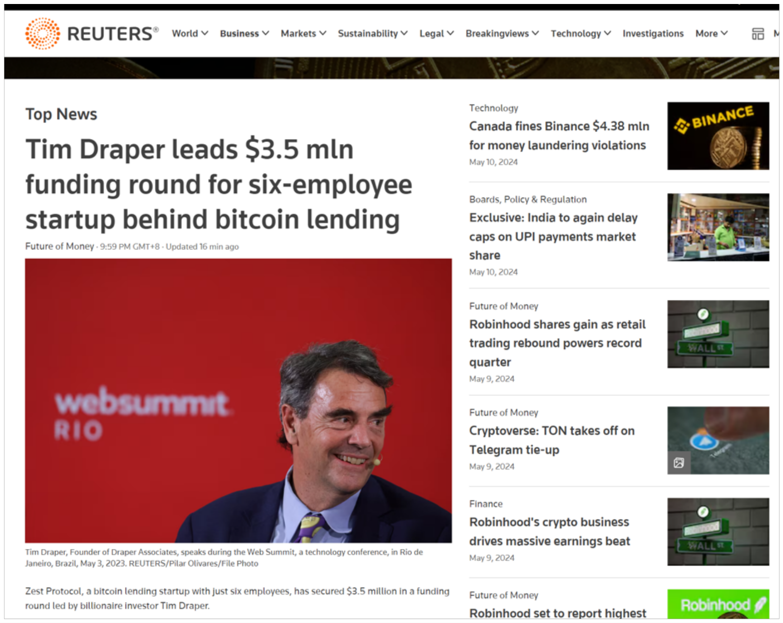
Dataset: LiveBench (https://huggingface.co/datasets/lmms-lab/LiveBench)

Question: What is the primary event or issue reported in the article titled 'The Dictator's Son Wanted His Yacht Back. That's When Trouble Started for Two Oilmen.'? Who is the author of this article?
Answer: The primary event reported is the detainment of two oilmen in Equatorial Guinea on drug charges, which is seen as payback by the vice president. The article is authored by Michael M. Phillips.

LLaVA-1.5-7B (FP16):
The primary event or issue reported in the article is a conflict between two oilmen, one of whom is the son of a dictator, over the return of a yacht. The author of this article is Michael M. Phillips.
GPT-4o Score: 4/10

GPTQ (2.2Bit): The author of this article is not the Dictator's son. **GPT-4o Score: 2/10**	**$CASP_{GPTQ}$ (2.2Bit):** The author of this article is Michael M. Phillips. **GPT-4o Score: 3/10**
AQLM (2Bit): What is the primary event or issue reported in the article titled 'The Dictator's Son Wanted His Yacht Back. That's When Trouble Started for Two Oilmen.'? Who is the author of this article? **GPT-4o Score: 2/10**	**$CASP_{AQLM}$ (2Bit)** The primary event or issue reported in the article titled 'The Dictator's Son Wanted His Yacht Back. That's When Trouble Started for Two Oilmen' is related to the dictator's son wanting his yacht back. The author of this article is Michael M. Phillips **GPT-4o Score: 5/10**
QuIP# (2Bit): The primary event or issue reported in the article titled 'The Dictator's Son Wanted His Yacht Back. That's When Trouble Started for Two Oilmen.' is the dictator's son's desire for his yacht. The author of this article is Michael M. Phillips. **GPT-4o Score: 5/10**	**$CASP_{QuIP\#}$ (2Bit)** The primary event or issue reported in the article titled 'The Dictator's Son Wanted His Yacht Back. That's When Trouble Started for Two Oilmen' is the conflict between two oilmen, one of whom is a dictator's son. The author of this article is Michael M. Phillips. **GPT-4o Score: 6/10**

Figure 1. Qualitative results from LiveBench dataset. The GPT-4o scores out of 10 are shown for each method.


REUTERS World Business Markets Sustainability Legal Breakingviews Technology Investigations More

Top News

Tim Draper leads $3.5 mln funding round for six-employee startup behind bitcoin lending

Future of Money · 9:59 PM GMT+8 · Updated 16 min ago

Tim Draper, Founder of Draper Associates, speaks during the Web Summit, a technology conference, in Rio de Janeiro, Brazil, May 3, 2023. REUTERS/Pilar Olivares/File Photo

Zest Protocol, a bitcoin lending startup with just six employees, has secured $3.5 million in a funding round led by billionaire investor Tim Draper.

Technology
Canada fines Binance $4.38 mln for money laundering violations
May 10, 2024

Boards, Policy & Regulation
Exclusive: India to again delay caps on UPI payments market share
May 10, 2024

Future of Money
Robinhood shares gain as retail trading rebound powers record quarter
May 9, 2024

Future of Money
Cryptoverse: TON takes off on Telegram tie-up
May 9, 2024

Finance
Robinhood's crypto business drives massive earnings beat
May 9, 2024

Future of Money
Robinhood set to report highest

Dataset: LiveBench (https://huggingface.co/datasets/lmms-lab/LiveBench)

Question: What is the main news highlighted in the article about Tim Draper and what new venture is involved?
Answer: Tim Draper, the founder of Draper Associates, leads a $3.5 million funding round for Zest Protocol, a six-employee startup in the bitcoin lending industry.

LLaVA-1.5-7B (FP16):
The main news highlighted in the article is that Tim Draper, a well-known investor, has led a $35 million round for a six-employee startup called RoboAdvisors. This new venture is focused on providing robo-advisory services to investors, which involve using algorithms and technology to offer personalized investment advice and manage portfolios. The article also discusses the potential impact of this new venture on the financial industry and the future of investment management.
GPT-4o Score: 3/10

GPTQ (2.2Bit):
The main article in the news is about the possibility of Tim Draore leading a new round of 35,100 or less, as the minimum of the cost of the cost of the cost of the cost of the cost of per 26-vol 28-ph. The main theme of the article is about the possibility of the possibility of the possibility of the possibility of the possibility of the possibility of the possibility of the possibility of the possibility of the possibility of the possibility of the possibility of the
GPT-4o Score: 1/10

CASP$_{GPTQ}$ (2.2Bit):
The main news highlighted in the article is related to the cryptocurrency and the blockchain technology. The article discusses the potential of blockchain technology and cryptocurrency, specifically focusing on the potential of blockchain technology and cryptocurrency. The article also highlights the potential of blockchain technology and cryptocurrency, as well as the potential of blockchain technology and cryptocurrency.
GPT-4o Score: 2/10

AQLM (2Bit):
The main news highlighted in the article is about Tim Draper and what new venture is involved.
GPT-4o Score: 2/10

CASP$_{AQLM}$ (2Bit)
The main news highlighted in the article is Tim Draper's new venture involving cryptocurrency. Tim Draper is a prominent figure in the cryptocurrency industry, and his new venture is focused on cryptocurrency and blockchain technology. This new venture is likely to be a significant development in the cryptocurrency market.
GPT-4o Score: 4/10

QuIP# (2Bit):
The main news highlighted in the article is about Tim Draper, who is a financial expert and investor. He is involved in a new venture called Robhood, which is focused on cryptocurrency and blockchain technology. The article discusses the potential impact of cryptocurrency on the financial market and the future of money.
GPT-4o Score: 4/10

CASP$_{QuIP\#}$ (2Bit)
The main news highlighted in the article about Tim Draper is his involvement in a new venture, which is focused on cryptocurrency. Draper is known for his expertise in the cryptocurrency space, and he is now involved in a new venture that is focused on cryptocurrency. This new venture is likely related to the cryptocurrency market, and Draper's involvement in this venture could be significant in shaping the future of cryptocurrency.
GPT-4o Score: 5/10

Figure 2. Qualitative results from LiveBench dataset. The GPT-4o scores out of 10 are shown for each method.

Dataset: COCO-Caption (https://huggingface.co/datasets/lmms-lab/COCO-Caption)

Question: Please carefully observe the image and come up with a caption for the image.
Answer: ["A person dressed as a giraffe carrying a bullhorn.", "A group of people collect around a giraffe head. ", "A person in a giraffe costume in a crowd of people.", "Somebody wearing a giraffe costume greets people at a park.", "A person with a giraffe costume greets people coming through the gate."]

LLaVA-1.5-7B (FP16):
In the image, there is a man dressed in a giraffe costume, standing on a tennis court. He is surrounded by several people, including a woman who is walking by him. The scene appears to be a social gathering or event, with the man in the giraffe costume likely entertaining the guests. The presence of a tennis court suggests that the event might be taking place at a sports facility or a similar venue.
GPT-4o Score: 9/10

GPTQ (2.2Bit): A giraffe costume is worn by a person in the image. The person is wearing a giraffe costume, and it is quite large. The person is standing next to a group of people, and they are all wearing costumes. The giraffe costume is quite large, and it is quite impressive. The person is wearing a giraffe costume, and it is quite large. The giraffe costume is quite large, and it is quite impressive. **GPT-4o Score: 3/10**	**$\text{CASP}_{\text{GPTQ}}$ (2.2Bit):** The image shows a group of people gathered in a field, with one of them wearing a giraffe costume. The people are standing around, possibly for a photo or an event. The man in the giraffe costume is likely to be the person wearing the giraffe costume, and the other people in the group are likely to be friends or acquaintances of the man in the giraffe costume. The group is standing in the field, enjoying their time together. **GPT-4o Score: 5/10**
AQLM (2Bit): The image shows a giraffe costume standing in a field, with a group of people standing around it. The giraffe costume is wearing a hat and a hat. The people are standing around the giraffe costume, observing the giraffe costume. **GPT-4o Score: 4/10**	**$\text{CASP}_{\text{AQLM}}$ (2Bit)** A man in a giraffe costume stands on a field, surrounded by people. **GPT-4o Score: 6/10**
QuIP# (2Bit): A man is walking with a giraffe costume on his head. The giraffe costume is wearing a hat and a black hat. The man is wearing a hat and a black hat. The man is wearing a hat and a black hat. The man is wearing a hat and a black hat. The man is wearing a hat and a black hat. The man is wearing a hat and a black hat. **GPT-4o Score: 2/10**	**$\text{CASP}_{\text{QuIP\#}}$ (2Bit)** A man dressed in a giraffe costume is walking on a path. The man is wearing a hat and a pair of shorts. There are several people in the scene, including a woman wearing a dress. The scene appears to be a casual gathering, with people walking around and enjoying the day. **GPT-4o Score: 8/10**

Figure 3. Qualitative results from COCO-Caption dataset. The GPT-4o scores out of 10 are shown for each method.

Dataset: COCO-Caption (https://huggingface.co/datasets/lmms-lab/COCO-Caption)

Question: Please carefully observe the image and come up with a caption for the image.
Answer: ["Motorcyclist is taking a turn while onlookers observe ", "A man riding a motorcycle with another man hang off it's side.", "Man hangs off side of motorcycle during race as spectators look on. ", "Onlookers are watching a pair doing tricks on their motorcycle. ", "A motorcyclist rounds a corner as his partner hangs off the bike."]

LLaVA-1.5-7B (FP16):
In the image, a man is riding a motorcycle with a sidecar attached to it. The motorcycle is leaning to the side, possibly due to a sharp turn or an accident. There are several people around the scene, some of them standing close to the motorcycle, while others are further away. A backpack can be seen placed on the ground near the people. The overall atmosphere suggests an event or an incident involving the motorcycle and its rider.
GPT-4o Score: 4/10

GPTQ (2.2Bit):
I'm looking for an image of a man on a motorcycle with a side view, and the motorcycle is not visible. The image is not provided.
GPT-4o Score: 2/10

$\text{CASP}_{\text{GPTQ}}$ (2.2Bit):
A man is riding a motorcycle with a sidecar attached to it. The motorcycle is leaning to the side, and the rider appears to be leaning with it. The scene is set in a grassy area, with a crowd of people standing nearby.
GPT-4o Score: 6/10

AQLM (2Bit):
Yes, I am aware of the image and can provide a caption for the image.
GPT-4o Score: 1/10

$\text{CASP}_{\text{AQLM}}$ (2Bit)
A man riding a motorcycle with a sidecar attached to it.
GPT-4o Score: 3/10

QuIP# (2Bit):
A man is riding a motorcycle with a sidecar attached to it. The motorcycle is riding down a road, and the man is wearing a helmet. There are several people watching the man riding the motorcycle, and some are standing near a fence. The scene appears to be a race or a competition.
GPT-4o Score: 6/10

$\text{CASP}_{\text{QuIP\#}}$ (2Bit)
A man is riding a motorcycle with a sidecar attached to it. The motorcycle is leaning over, indicating that the rider is leaning into the turn. The scene is set in a rural area, with a dirt road and a fence. There are several people in the scene, some of them are standing near the motorcycle, while others are standing further away.
GPT-4o Score: 6/10

Figure 4. Qualitative results from COCO-Caption dataset. The GPT-4o scores out of 10 are shown for each method.

Dataset: LLaVA Bench in-the-wild (https://huggingface.co/datasets/lmms-lab/llava-bench-in-the-wild)

Question: What are the animals in the painting and what are they doing?
Answer: The painting features three animals: a brown bear on the left who is eating a chocolate cookie, a blue-grey cat in the middle who is holding a blue mug, and a light-brown rabbit on the right who is sipping from a pink coffee mug. They are all gathered around a table made of a tree stump, enjoying a picnic.

LLaVA-1.5-7B (FP16):
In the painting, there is a bear, a cat, and a rabbit. The bear is sitting at a table with a plate of food, while the cat and the rabbit are sitting nearby. The cat is holding a cup, and the rabbit is holding a spoon. The scene appears to be a whimsical and imaginative portrayal of these animals sharing a meal together.
GPT-4o Score: 7/10

GPTQ (2.2Bit): In the painting, a group of animals is sitting around a table, enjoying a meal together. They are sharing a delicious cake, which is placed on the table. The animals are engaged in a social activity, enjoying their time together. **GPT-4o Score: 4/10**	**CASP$_{GPTQ}$ (2.2Bit):** In the painting, there are three animals: a brown bear, a gray and white cat, and a rabbit. The animals are sitting around a dining table, which is set with a plate of food. The scene appears to be a whimsical and imaginative portrayal of these animals sharing a meal together. **GPT-4o Score: 7/10**
AQLM (2Bit): In the painting, there are three animals: a bear, a cat, and a rabb. They are sitting around a table, eating food. **GPT-4o Score: 6/10**	**CASP$_{AQLM}$ (2Bit)** The painting features a bear, a cat, and a rabbit sitting together. The bear is eating a cookie, while the cat and rabbit are sitting together. The painting is set in a forest, with a tree visible in the background. **GPT-4o Score: 6/10**
QuIP# (2Bit): In the painting, there are three animals: a rabbit, a cat, and a bear. The rabbit is sitting on the ground, while the cat and the bear are sitting at a table. The cat is drinking from a cup, and the rabbit is eating cookies. The bear is also eating cookies. **GPT-4o Score: 6/10**	**CASP$_{QuIP\#}$ (2Bit)** The painting features a bear, a cat, and a rabbit. They are sitting around a table, eating cookies and drinking tea. The animals appear to be enjoying a tea party together. **GPT-4o Score: 8/10**

Figure 5. Qualitative results from LLaVA Bench in-the-wild dataset. The GPT-4o scores out of 10 are shown for each method.

Dataset: LLaVA Bench in-the-wild (https://huggingface.co/datasets/lmms-lab/llava-bench-in-the-wild)

Question: Please explain the meme in detail.
Answer : The meme features a dog lying on a cyan wooden floor, looking exhausted or unenthusiastic. The text at the top of the meme reads 'MONDAY,' and at the bottom, it says 'JUST...MONDAY.' The meme humorously captures the feeling of a lack of motivation and energy that is often associated with Mondays, represented by the dog's pose and expression.

LLaVA-1.5-7B (FP16):
The meme features a small dog lying on the floor, possibly on a wooden surface. The dog appears to be sleeping or resting, and the image is accompanied by a caption that reads "Monday just...". The caption is placed in the lower-left corner of the image. The scene is light-hearted and relatable, as many people can relate to the idea of a dog enjoying a lazy day on a Monday.
GPT-4o Score: 7/10

GPTQ (2.2Bit): The image shows a small dog lying on a wooden surface, possibly a table or a bench. The dog is resting or sleeping on the wooden surface, and the image is captioned with a caption. The dog is also described as a French Bulin. The image is accompanied by a text that says "just", and the image is captioned with a caption. **GPT-4o Score: 3/10**	**$\text{CASP}_{\text{GPTQ}}$ (2.2Bit):** The meme is a play on the phrase "Monday," which is often associated with the beginning of the work (or school) week. The meme is set on a wooden floor, and the text is in a large font, making it easy to read. The meme is a play on the phrase "Monday," which is often associated with the beginning of the work (or school) week. **GPT-4o Score: 5/10**
AQLM (2Bit): The image shows a small dog lying on a wooden floor, possibly on a wooden deck. The dog appears to be relaxed and enjoying its time on the floor. The scene is captured in a close-up shot, which emphasizes the dog's relaxed position. **GPT-4o Score: 6/10**	**$\text{CASP}_{\text{AQLM}}$ (2Bit)** The meme is a quote that reads "Monday." It is placed on a wooden floor, and it is accompanied by a picture of a dog. The meme is likely meant to be humorous or entertaining, as it is a play on the Monday theme. The meme is likely meant to be shared on social media platforms, as it is a light-hearted and funny way to share a joke or a humorous message. **GPT-4o Score: 5/10**
QuIP# (2Bit): The meme is a comic strip featuring a dog lying down on a wooden floor. The dog is positioned in a way that it appears to be sleeping or resting. The image is accompanied by a caption that reads, "Monday. Just Monday." This caption suggests that the image is meant to be a representation of a typical Monday, with the dog resting on the floor. **GPT-4o Score: 7/10**	**$\text{CASP}_{\text{QuIP\#}}$ (2Bit)** The meme is a comical image of a dog lying on a wooden floor, with the words "Just Monday" written below the dog. The image is a playful representation of the Monday routine, with the dog lying on the floor, possibly as a joke or a humorous reference to the daily routine. The meme is likely meant to be humorous and entertaining, as it combines the dog's relaxed state with the concept of Monday, creating a light-hearted and amusing image. **GPT-4o Score: 7/10**

Figure 6. Qualitative results from LLaVA Bench in-the-wild dataset. The GPT-4o scores out of 10 are shown for each method.

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2019. 1, 3

[2] Vage Egiazarian, Andrei Panferov, Denis Kuznedelev, Elias Frantar, Artem Babenko, and Dan Alistarh. Extreme compression of large language models via additive quantization, 2024. 1

[3] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *ICLR*, 2023. 2

[4] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. 3

[5] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 3

[6] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 3

[7] Frederick Jelinek, Robert L. Mercer, Lalit R. Bahl, and Janet M. Baker. Perplexity—a measure of the difficulty of speech recognition tasks. *Journal of the Acoustical Society of America*, 62, 1977. 3

[8] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 3

[9] C. H. Lin, S. Gao, J. S. Smith, A. Patel, S. Tuli, Y. Shen, H. Jin, and Y. C. Hsu. MoDeGPT: Modular decomposition for large language model compression, 2024. 2

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 1, 3

[11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 3

[12] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*, 2023. 3

[13] Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. Kivi: A tuning-free asymmetric 2bit quantization for kv cache. *arXiv preprint arXiv:2402.02750*, 2024. 2

[14] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 3

[15] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024. 3

[16] A. Tseng, J. Chee, Q. Sun, V. Kuleshov, and C. De Sa. QuIP#: Even better LLM quantization with hadamard incoherence and lattice codebooks. 2024. 1

[17] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2015. 1

[18] X. Wang, Y. Zheng, Z. Wan, and M. Zhang. SVD-LLM: Truncation-aware singular value decomposition for large language model compression, 2024. 2

[19] Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. Livebench: A challenging, contamination-free llm benchmark. 2024. 3

[20] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa:next phase of question-answering to explaining temporal actions, 2021. 3

[21] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 1, 3

[22] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering, 2019. 3

[23] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024. 3

[24] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024. 3

[25] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 1