On the Human-level Performance of Visual Question Answering

Chenlian Zhou[♡], **Guanyi Chen**[♠*], **Xin Bai**[♡], **and Ming Dong**[♠]
[♠]School of Computer Science,
[♡]Faculty of Artificial Intelligence in Education,
Central China Normal University
{g.chen, dongming}@ccnu.edu.cn, {myphyllis, xin_b}@mails.ccnu.edu.cn

Abstract

Visual7W has been widely used in assessing multiple-choice visual question-answering (VQA) systems. This paper reports on a replicated human experiment on Visual7W with the aim of understanding the human-level performance of VQA. The replication was not entirely successful because human participants performed significantly worse when answering "where", "when", and "how" questions in Visual7W compared to other question types. An error analysis discovered that the failure was a consequence of the non-deterministic distractors in Visual7W. GPT-4V was then evaluated using Visual7W and was compared to the human-level performance. The results embody that, when evaluating models' capacity on Visual7W, the performance is not necessarily the higher, the better.

1 Introduction

Visual Question Answering (VQA) often serves as a proxy for assessing the "Level of AGI" (Morris et al., 2023) of AI systems as it requires an inclusive range of abilities, including fine-grained image recognition, spatial awareness, action recognition and knowledge-based reasoning (Antol et al., 2015).

In recent years, a bank of VQA datasets has been released (Malinowski and Fritz, 2014; Antol et al., 2015; Gao et al., 2015; Zhu et al., 2016). Among these datasets, Visual7W (Zhu et al., 2016) is one of the most cited ones and has been used as a standard benchmark dataset for testing the capacity of vision and language models. This is because, on the one hand, Visual7W standardises VQA as multiple-choice questions (MCQ) that ground on images so that the assessment results are easy to quantify using, e.g., accuracy. On the other hand, Zhu et al. (2016) carried out a human experiment on Visual7W, and obtained remarkably high human performance (96.6%). This reveals that the better a model performs on Visual7W, the more closely it aligns with human-level intelligence. Nonetheless, since the outcomes of the human experiment are not publicly available, it is hard to conduct in-depth comparative studies between AI systems and humans on Visual7W.

This paper reports on an attempt to replicate the human experiment on Visual7W. Our study was unable to achieve the remarkably high level of human performance reported in Zhu et al. (2016), which has an accuracy of 96.6%. Instead, we found that humans indeed excel in certain question types but not in others, making the overall human performance way lower than the reported percentage.

Then, focusing on the "mistakes" made by humans, we carried out an error analysis and found that nearly all "mistakes" are results of non-deterministic or incorrect distractors in MCQs. At length, based on the outcomes of our human experiment, we examined one of the most advanced vision and language models, namely, GPT-4V (Achiam et al., 2023). Our analysis suggested that, in VQA evaluation, the performance is not necessarily the higher, the better on all question types because the human-level performance could be unexpectedly imperfect.

2 The Visual7W Dataset

Visual7W is one of the most widely used and cited VQA datasets due to its extensive coverage, easy-to-use and high quality. On the basis of 47,300 images from the COCO dataset (Lin et al., 2014), Visual7W collected 327,939 MCQs crowdsourcingly. [1] Concretely, each question comes with 4 candidate answers and falls into one of 7 question types, including "what", "where", "when", "who", "why", "how", and "which". Visual7W has two dif-

*Corresponding Author

[1]Visual7W is available at: https://ai.stanford.edu/~yukez/visual7w/.

Proceedings of the 31st International Conference on Computational Linguistics, pages 4109–4113
January 19–24, 2025.

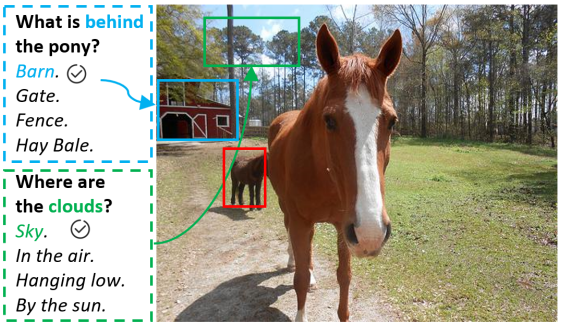

Figure 1: Example multiple choice VQA items in Visual7W.

Table 1: Human Performance on the Visual7W, in which QT is question type, Acc. is accuracy, PA is percentage agreement, κ is the kappa value. The "REF" column charts the results reported in Zhu et al. (2016). The human performance that is very different from REF is underlined.

	REF	Human		
QT	*Acc.*	*Acc.*	*PA*	κ
what	96.5	96.0	92	0.89
where	96.7	84.5	86	0.81
when	96.7	80.0	79	0.75
who	96.5	93.5	90	0.87
why	92.7	91.0	88	0.84
how	94.2	86.5	89	0.85

ferent VQA tasks: *pointing*, which involves "which" questions and where candidate answers are bounding boxes in the given images, and *telling*, which includes questions that are not of "which" type and where candidate answers are textual. Example samples are shown in Figure 1.

MCQs in Visual7W were collected by first asking crowdsourcing workers to produce QA pairs given images and certain question types, which were then filtered manually to ensure the quality and ensure that every pair can ground on the given images. Subsequently, another group of crowdsourcing workers were hired to produce the distractors for each QA pair, transforming it into an MCQ. However, this time, Visual7W deployed no further double-check on the quality of these distractors.

Finally, Zhu et al. (2016) conducted a human experiment asking participants to answer MCQs in Visual7W. The experiment obtained a remarkably high human-level performance on Visual7W. The first column of Table 1 depicts the human performance of each question type in the *telling* portion.

3 Human-level Performance on Visual7W

In this section, we introduce the setup for replicating the human experiment in Zhu et al. (2016). We then report the outcomes and compare them to what was reported in Zhu et al. (2016).

3.1 Material

Since the *telling* and the *pointing* sections in Visual7W are two very different tasks and consider that the *pointing* section is much smaller, we focused only on the *telling* section in this replication study. For each question type in the *telling* section, we randomly sampled 100 items, resulting in 600 test items in total.

3.2 Procedure and Participants

We shuffled all the items, and the candidate answers in each item. Then, these items were randomly divided into 6 groups and each group was completed by 2 participants. As a result, we recruited 12 participants, who all have backgrounds in science and can speak fluent English. Of these, 2 were identified as female and 10 were identified as male.

3.3 Results

We computed the average accuracy of humans' answers for each question type and the *Cohen's kappa* coefficient as well as the percentage agreements to measure the inter-annotator agreement (IAA). Table 1 charts all results. Our participants achieved "perfect" agreement (i.e., percentage agreement $> 90\%$ and $\kappa > 0.8$) on almost all question types. The only exception was the "when" type, which still showed substantial agreement, demonstrating the high quality of our human experiment. The outcomes of our human experiment are available at: https://github.com/a-quei/visual7w_human.

Compared to those reported in Zhu et al. (2016), our human participants performed equally well on "what", "who", and "why" questions but showed significantly lower performance when answering "where", "when", and "how" questions. For instance, for "when" questions, Zhu et al. (2016) reported an accuracy of 94.4%, while our participants were only correct on 80.0% of the time.

3.4 Error Analysis

To ascertain the quality of our human experiment and explore the root of these gaps, we conducted an error analysis, in which we manually checked every

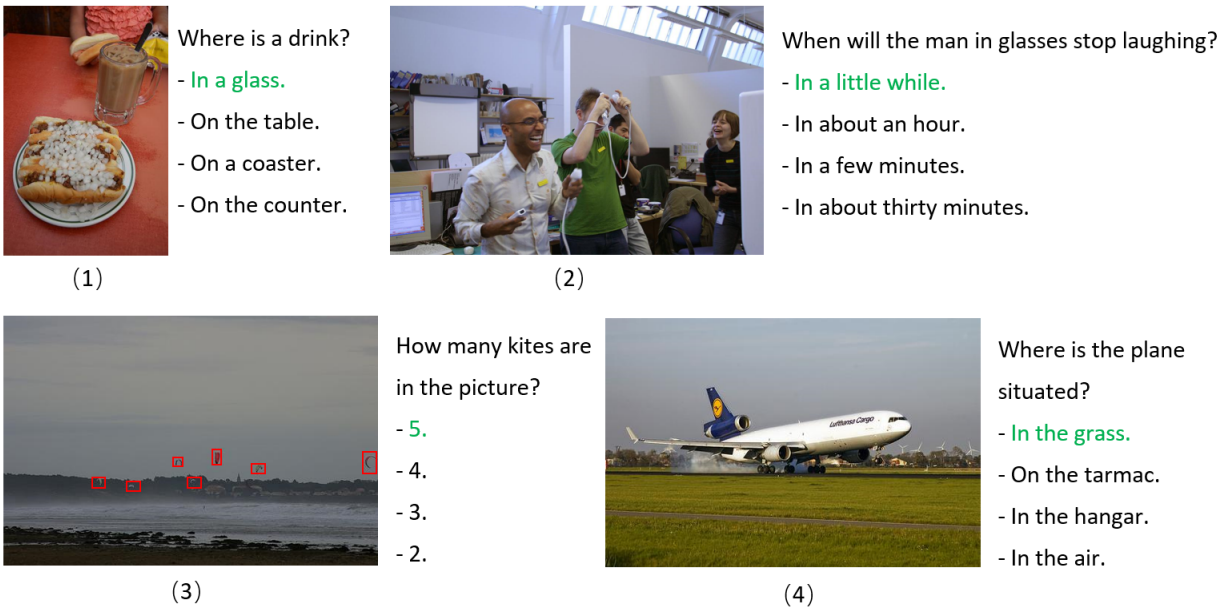

Figure 2: Example test items that our participants did not answer correctly. The ground truth answers are in green. Note that we hereby put all ground truth answers as the first options of all MCQs, but during the experiment, the options were randomly shuffled.

incorrect answer of our participants. Except for 3 cases caused by carelessness, we found all the rest of the "incorrect" answers to be very reasonable.

We have identified the major reason is that most MCQs that our participants answered incorrectly appear to have multiple reasonable answers in addition to the ground truth. We call MCQs as such as *non-deterministic* MCQs. The presence of these non-deterministic MCQs in Visual7W may be caused by the fact that, according to Zhu et al. (2016), the distractors in Visual7W were written by crowdsourcing workers based solely on the questions and the ground truth answers, without having the access to the paired images. However, intuitively, many "where", "when", and "how" questions need to be grounded on the images. As a result, even though the written candidate answers may seem very different from the ground truth, they are not necessarily incorrect answers, and they may turn out to be considered correct once the paired images are provided, making the MCQs non-deterministic. Additionally, as aforementioned, as no further filtering was applied during the construction of MCQs in Visual7W, these non-deterministic MCQs were ultimately included.

We then analysed these non-deterministic MCQs together with their paired images and found that the non-deterministic could be attributed to the ambiguity of the question. In the example of Figure 2(1), given the image, "a drink" in the question could refer to either "the liquid in the glass" or "the glass of liquid", making both the option "in a glass" and the option "on the table" valid answers. Furthermore, the non-deterministic could also be caused by the lack of necessary context. For example, the question "*when will the man stop laughing*" Figure 2(2) is almost unanswerable without giving further context, such as what he is laughing for or, usually, how long he laughs every time.

Moreover, we also observed several items whose ground truth are incorrect. The example in Figure 2(3) asks the number of kites. With a closer look at the given image, we identified at least 7 kites (which are marked in Figure 2(3)) while the ground truth says 5 and the correct answer is not even included in this MCQ. The example in Figure 2(4) shows an item whose ground truth states the plane is situated in the grass. Yet, upon closer inspection of the image, it is evident that the plane is actually situated "on the tarmac" but not "in the grass".

4 Comparing GPT-4V with Human

To check whether the findings from examining the models on Visual7W are consistent with those from comparing the models to humans, we used Visual7W to evaluate GPT-4V in a zero-shot setting and compared its outputs to that of humans.

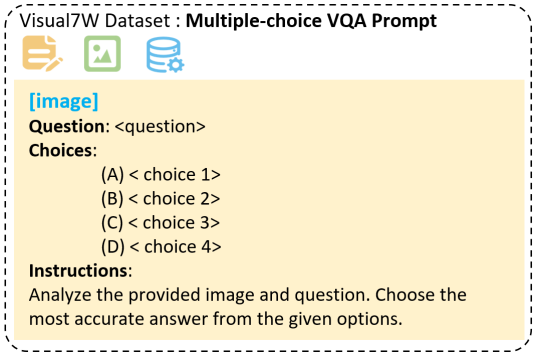

Figure 3: An example prompt we used for assessing GPT-4V on Visual7W.

Table 2: The performance of GPT-4V on the Visual7W, in which PAG is the percentage agreement between human and GPT-4V.

QT	*Acc.*	*PAG*	$P(\bar{H}\|\bar{G})$
what	86	85	28.58
where	82	80	55.56
when	79	69	42.86
who	86	88	35.71
why	84	82	43.75
how	65	60.5	22.86

4.1 Setup

We asked GPT-4V to accomplish the multiple-choice VQA task using exactly the same 600 items mentioned in § 3.1. After several pilot studies on a small set of data, we ended up with the prompt depicted in Fig.3.

In addition to accuracy, to compare GPT-4V to humans quantitatively, we computed (1) *PAG*: percentage agreement between humans and GPT-4V, and (2) the proportion of GPT-4V failed questions that at least one of our human participants also failed. Formally, this is the conditional probability $P(\bar{H}|\bar{G})$, where $\bar{H}$ means at least one of our human participants failed to answer the question correctly.

4.2 Results

Table 2 reports the performance of GPT-4V. Focusing on the accuracy of GPT-4V alone, we noticed that GPT-4V achieved its best accuracy at 86% on "what" and "who" questions, while it did not work well on "when" and "how" questions. Nonetheless, if we compared GPT-4V to humans using the accuracy numbers in Table 1 as well as PAG and $P(\bar{H}|\bar{G})$, we would have different observations.

First, this time, the GPT-4V's performance on "where" and "when" questions is closest to the human-level performance among all question types, even though its accuracy scores on these two question types are not the highest. More importantly, GPT-4V also has high $P(\bar{H}|\bar{G})$ scores on "where" and "when" questions, suggesting GPT-4V confused at similar non-deterministic MCQs as humans. [2]

Second, although, as discussed, GPT-4V have the highest accuracy score on "what" questions, one of the biggest performance gaps is also identified on "what" questions. More specifically, for the 100 "what" questions, humans can easily answer 96 out of 100 questions correctly, while GPT-4V can only handle 86 of them. A low $P(\bar{H}|\bar{G})$ score indicates that it makes errors that are rare for humans, suggesting that they behave differently.

Last, the comparison strengthens the conclusion above that GPT-4V is not good at answering "how" questions. In comparison to humans, it accurately handles 20 fewer "how" questions, even though humans' proficiency in answering "how" questions is already one of the lowest among all question types. Moreover, it also receives the lowest PAG and $P(\bar{H}|\bar{G})$ scores on "how" questions, which means most of its mistakes are not the ones that humans would also make, concluding GPT-4V's inability to handle "how" questions. We further found that most "how" questions ask "how many". Thus, GPT-4V's inability to handle "how" questions should be the result of its low ability to counting (Golovneva et al., 2024).

In short, due to more non-deterministic MCQs in some question types than the other, the performance on Visual7W is not the higher, the better. As shown in the examples of "where" and "when" questions, a model can have a lower accuracy score on a question type only if the question type has more non-deterministic MCQs.

5 Conclusion

In this paper, we attempted to reproduce a human experiment about multiple-choice VQA on Visual7W (Zhu et al., 2016). In our replication, Human participants performed significantly worse than what was reported in Zhu et al. (2016) on three question types: "where", "when", and "how". An error analysis revealed that this discrepancy

[2]Note that, after a double check, GPT-4V has a low PAG score on "when" questions because it confused similar MCQs with humans but selected different answers. This is rational since human participants also have a relatively low IAA on "when" questions.

may be due to the lack of quality control during the construction of distractors, which led to many non-deterministic MCQs in Visual7W. Finally, we tested GPT-4V using Visual7W and compared its outputs to humans', in which we showed that higher performance on Visual7W does not necessarily equate to a better model. We hope that our findings encourage others to focus more on comparing models to human-level performance instead of merely the ground truth in corpora during the VQA evaluation.

Limitations

The current work investigates the human-level performance of VQA. One limitation is that we only focused on multiple-choice VQA and a single dataset, namely, Visual7W. In future, we plan to extend the work to other main-stream VQA datasets and open-ended VQA. We also plan to extend the work to include a wider range of subjects as well as a closer look at the errors (van Miltenburg et al., 2020).

Moreover, it is worth noting that the experimental setup in this replication study is not identical to that of Zhu et al. (2016) due to the lack of necessary details.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question. *Advances in neural information processing systems*, 28.

Olga Golovneva, Tianlu Wang, Jason Weston, and Sainbayar Sukhbaatar. 2024. Contextual position encoding: Learning to count what's important. *arXiv preprint arXiv:2405.18719*.

Tsung-Yi Lin, Michael Maire, Serge J. 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*, pages 740–755.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in neural information processing systems*, 27.

Meredith Ringel Morris, Jascha Sohl-dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. 2023. Levels of agi: Operationalizing progress on the path to agi. *arXiv preprint arXiv:2311.02462*.

Emiel van Miltenburg, Wei-Ting Lu, Emiel Krahmer, Albert Gatt, Guanyi Chen, Lin Li, and Kees van Deemter. 2020. Gradations of error severity in automatic image descriptions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 398–411, Dublin, Ireland. Association for Computational Linguistics.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004.