Analyzing the Sensitivity of Vision Language Models in Visual Question Answering

Monika Shah, Sudarshan Balaji, Somdeb Sarkhel, Sanorita Dey, Deepak Venugopal
University of Memphis, TN, USA
Adobe Research, San Jose, CA, USA
University of Maryland Baltimore County, USA
{mshah2, sbalaji, dvngopal}@memphis.edu, sarkhel@adobe.com, sanorita@umbc.edu

Abstract

We can think of Visual Question Answering as a (multimodal) conversation between a human and an AI system. Here, we explore the sensitivity of Vision Language Models (VLMs) through the lens of cooperative principles of conversation proposed by Grice. Specifically, even when Grice's maxims of conversation are flouted, humans typically do not have much difficulty in understanding the conversation even though it requires more cognitive effort. Here, we study if VLMs are capable of handling violations to Grice's maxims in a manner that is similar to humans. Specifically, we add modifiers to human-crafted questions and analyze the response of VLMs to these modifiers. We use three state-of-the-art VLMs in our study, namely, GPT-4o, Claude-3.5-Sonnet and Gemini-1.5-Flash on questions from the VQA v2.0 dataset. Our initial results seem to indicate that the performance of VLMs consistently diminish with the addition of modifiers which indicates our approach as a promising direction to understand the limitations of VLMs.

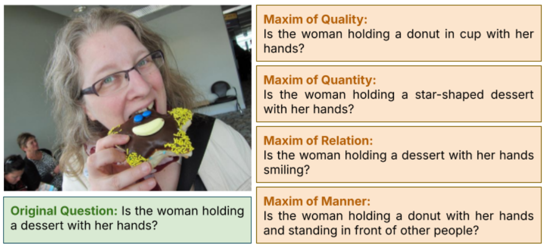
1 Introduction

Vision Language Models (VLMs) (Team et al., 2024; Liu et al., 2023; Hurst et al., 2024) that unify Large Language Models with computer vision have made significant advances in multimodal tasks such as image captioning (Yang et al., 2019; Cornia et al., 2020; Wang et al., 2022) and visual question answering (VQA) (Antol et al., 2015). However, we are just beginning to understand the reasoning capabilities and more importantly, the limitations of these models (Campbell et al., 2024). In this work, inspired by theories from cognitive science, we understand the behavior of VLMs in VQA when we *increase the cognitive load in comprehending questions*. Specifically, in Grice's classical theory of *cooperative principles* (Grice, 1975), it is known that humans acting cooperatively in a conversation typically need to follow a set of rules commonly known as *Grice's maxims*. These maxims make conversation more effective and ensure efficient communication. However, it is known from previous studies that even when these maxims are violated, humans can comprehend conversation easily (Davies, 2000). However, violations to Grice's maxims places greater cognitive burden on the listener (Jacquet et al., 2018).

In this work, we study how VLMs react when Grice's maxims are violated. Specifically, we treat VQA as a single utterance conversation where a human is asking the AI model a question to which the AI model responds with an answer. We introduce modifiers into human-crafted questions that adds greater detail to a question. At the same time, these details typically tend to violate Grice's maxims since they were not deemed to be essential when a human crafted the original question. While an AI model could benefit from the added information, processing modifiers will increase the reasoning required to answer the question. We add two types of modifiers, namely, *visual* and *relational* modifiers. The visual modifiers add more detail related to visual properties such as color, shape, etc., while relational modifiers add details related to spatial relationships.

We use VLMs to generate a modified question with either visual or relational modifiers. Next, we verify if the modified question changes human perception. That is, if humans can answer the modified question with an answer that is equivalent to the answer to an unmodified question, this implies that the modifier does not alter the question. Therefore, we would expect a VLM to be able to do a similar type of reasoning. We evaluate this on three state-of-the-art VLMs, *GPT-4o* (OpenAI, 2024), *Gemini-1.5-Flash* (Team et al., 2024) and *Claude-3.5-Sonnet* (Anthropic, 2024) on the VQA v2.0 dataset. That is, we generate modi-

Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM² 2025), pages 431–438
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics



Maxim of Quality: Is the woman holding a donut in cup with her hands?

Maxim of Quantity: Is the woman holding a star-shaped dessert with her hands?

Maxim of Relation: Is the woman holding a dessert with her hands smiling?

Maxim of Manner: Is the woman holding a donut with her hands and standing in front of other people?

Original Question: Is the woman holding a dessert with her hands?

Figure 1: Original question of the green is the question that satisfies the Grice's maxim and the questions with modifiers that violates the Grice's maxim.

fied questions from each of these VLMs and evaluate the responses of each VLM to the modified questions. Our initial results seem to indicate that VLMs are sensitive to modifications to questions. In particular, we find that there is a consistent performance degradation in the presence of modifiers. In particular, when modifiers are added through Gemini-1.5-Flash, the performance degradation is more significant in all 3 VLMs.

2 Related Work

Following the original VQA task (Antol et al., 2015), several improved datasets for VQA have been developed (Goyal et al., 2017; Selvaraju et al., 2020; Tan and Bansal, 2019) to evaluate VQA systems. More recently, the trend has shifted towards incorporating LLMs within the evaluation process. For instance, (Zhou et al., 2023) uses ChatGPT to automatically evaluate model outputs on a Likert scale. The work in (Mañas et al., 2024) leverages LLMs to evaluate answers. Specifically, it formulates VQA as an answer-rating task where the LLM (Flan-T5 (Chung et al., 2024), Vicuna-v1.3 (Chiang et al., 2023) and GPT-3.5-Turbo) is instructed to score the correctness of a candidate answer given a set of reference answers. The work in (Britton et al., 2022) is related to our approach where it adds question modifiers to VQA and analyzes its effect on LXMERT (Tan and Bansal, 2019). However, there has not been a significant amount of work that relates the reasoning of VLMs in VQA grounded in principles of human cognition which is the direction we follow in this work.

3 Pragmatics in Visual Question Answering

Grice's classical theory of cooperative principles in pragmatics is widely used to characterize human conversation. Specifically, Grice developed principles that explain effective conversation between participants assuming that the participants have a common goal of understanding each other and therefore act cooperatively. These principles are summarized in four maxims, namely, the maxim of quality, quantity, relation and manner. The maximum of quality suggests that speakers should be as truthful as possible and only say what they believe to be true based on evidence. The maxim of quantity suggests that the right amount of information must be provided in a conversation, i.e., one should not add too much or too little information. The maxim of relation suggests that a speaker should stay relevant to the topic and the maxim of manner suggests the need to avoid ambiguity and focus on clarity.

While Grice's maxims characterize effective conversation, violation of Grice's maxims does not mean that the conversation is incomprehensible (Davies, 2000). Specifically, since participants are assumed to be acting cooperatively, if a speaker violates a maxim, then the burden of understanding falls on the listener. That is, the listener is expected to work harder (cognitively) to comprehend the intention behind utterances that violate the maxims. We use this principle as a way to understand the limitations of VLMs. Specifically, we think of the VQA task as a conversation that involves a single utterance between two participants, i.e., one participant is a human who asks a question to the AI model and the other participant is the AI model that needs to generate an answer. In cases where the human participant flouts Grice's maxims, there is an increased burden of understanding on the AI model to produce an answer that the human can agree on.

3.1 Adding Modifiers to VQA

Fundamentally, the purpose of modifiers in text is to add more detail. Modifiers can add more specifics to a description, clarify information to improve comprehensibility or can make text more engaging for a reader. At the same time, from a cognitive perspective, processing modifiers places greater demands on attention and reasoning capabilities. Specifically, it is known that to understand text with greater syntactic complexity (which can occur if modifiers are added to questions) the level of neural activity in the brain increases (Just et al., 1996). Further, if we consider our view of VQA as a conversation between a human and an AI model,

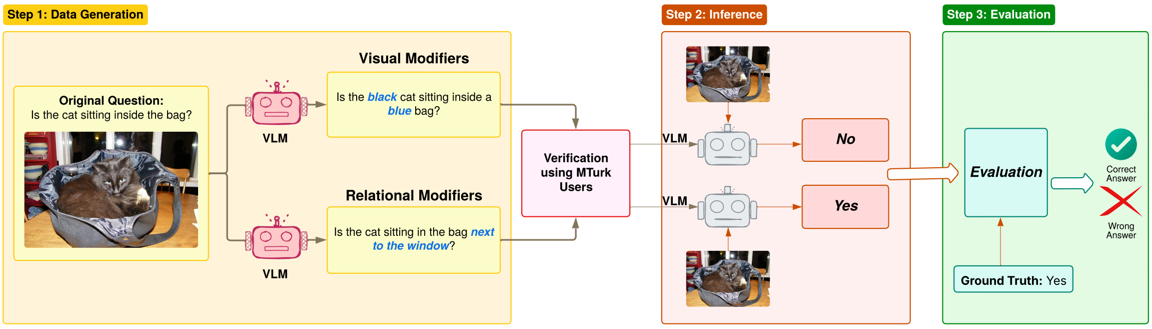


Figure 2: Illustrating our workflow. We generate modified questions from human-crafted questions using a VLM. Next, we verify if the modifier changes human perception of the question by comparing answers to the modified questions (collected through AMT) to the answers given to the original questions. For questions where the answers are alike, we evaluate if a VLM gives similar answers to the original and modified questions.

adding modifiers to a human-crafted question is very likely to violate Grice's maxims which again results in the need for greater reasoning capability. For instance, consider the example shown in Fig. 1. The original question written by a human seems to follow Grice's maxims. However, by adding modifiers, we violate these maxims as illustrated in the example. Importantly though, each of the modified questions can be easily answered by humans even when they violate at least one of the maxims and have increased complexity of the question (for instance, in our AMT study, humans answered modified questions with answers similar to those in unmodified questions). Pragmatically, since the AI is interacting with humans (e.g.in standard VQA, we use human-generated questions (Antol et al., 2015)), such an interaction is likely to follow Grice's maxims assuming that humans are acting cooperatively and not maliciously. That is, if we consider the example shown in Fig. 1, it is unlikely that a human would ask the AI any of the questions where the modifiers violate the maxims. However, human reasoning is fairly robust to such modifications. Our goal is use these modifiers to help us explain if the reasoning mechanism of the model is equally robust. Specifically, the modifiers may i) describe new concepts such as the *star-shape* that describes the shape of the dessert, ii) add additional information such as where the woman is standing or iii) add ambiguity such as if the woman's facial expression describes a smile. The AI model could in theory use the additional context to improve the accuracy of its answers in the VQA task. In other cases, the accuracy may diminish either due to increased ambiguity or a lack of model capacity to process modifiers.

3.2 Evaluation Methodology

We add modifiers to human-written questions targeting specific properties in the image. Specifically, here, we consider two properties that are broad enough to describe the scene in an image in greater detail, i.e., *visual properties* and *relational properties*. Specifically, visual properties refer to attributes such as color, shape, texture, etc. for objects that are observed in the scene. Relational properties refer to spatial relationships in the scene such as *next to*, *on top of*, etc. We prompt a VLM (with the image and original question) to generate the modified question with a specific type of modification (i.e., visual/relational). We instruct it to add the modifier without changing the answer to the original question and also without altering the question type (e.g. a *what* question needs to remain a *what* question). Further, we also instruct it to not alter the context of the question significantly. Next, we use Amazon Mechanical Turk (AMT) to verify if violations to Grice's maxims alter human perception. Specifically, we ask a human to answer a question with a modifier and compare this answer to answers given to the original question. Note that for questions where the original answer has a *unique ground truth* (yes/no questions and numeric questions), it is easy to verify if the answer changes from the original answer. However, for a question that is open-ended, there could be multiple ground truth answers. For such cases, we use an LLM to compare answers to the modified and unmodified questions, and instruct it to quantify the similarity between them on a discrete 1-10 scale. An illustration of our evaluation workflow is shown in Fig. 2. More details about the prompts and the AMT study are presented in the appendix.

3.2.1 Results

We evaluate 3 well-known VLMs, *GPT-4o*, *Gemini-1.5-Flash* and *Claude-3.5-Sonnet* using the VQA v2.0 dataset (Goyal et al., 2017). We added visual and relational modifiers to 1000 questions from the test set of VQA v2.0. We selected these questions such that we had an equal number of instances corresponding to each question type (there are 55 question types, e.g. *what*, *why*, *is*, *how*, etc.). The sampled data we used consists of 500 yes/no and numeric questions (where the answer is a number) and 500 open-ended questions. We evaluated each VLM on modified questions generated from each of the 3 VLMs.

Tables 1, 2 show the % change in accuracies of answers given by the VLM when modifiers are added to the original questions. The results in Table 1 correspond to yes/no and numeric questions where we can evaluate the answers exactly since these questions have a unique ground truth answer. As seen from our results, the positive values of % change in almost all cases indicates that the models performed worse on modified questions regardless of which VLM performed the modification. Modifiers added by Gemini-1.5-Flash seemed to be the hardest ones to process for all 3 VLMs since the average % change over all the VLMs was the largest. The modifiers added by Claude-3.5-Sonnet seemed to be easier to process for all 3 VLMs since the average % change in accuracy was the smallest across all the models. This seems to indicate that Claude-3.5-Sonnet may not add substantially detailed modifiers compared to the other models. Interestingly, Gemini-1.5-Flash performed worse with self-modified questions compared to modifications by other VLMs both for visual and relational modifiers. In the case of GPT-4o, self-modified questions did not result in a significant change to the model's accuracy as compared to its change in accuracy on questions modified by other VLMs. This indicates that GPT-4o can handle specific forms of modifications which is built into its prior but struggles with other forms of modifications. While our results indicate that some VLMs perform better than others, the specific reasons for why this may be the case is still unclear. We plan to explore this in future.

For the open-ended question results shown in Table 2, we use GPT-4o to evaluate the similarity between human-generated answers to the original question (which we collected using AMT) and the answers given by the model to the original and modified questions. In this case, we only compare the texts using GPT-4o. For each question, we used answers from 3 AMT workers and considered all the 3 similarity scores provided by GPT-4o on a discrete scale between 1 and 10. Table 2 shows the % difference between these scores for answers given by the VLM to the original questions with those given by the VLM for modified questions. Overall, similar to our earlier result, in all cases, the models performed worse on modified questions (positive % change values) regardless of which VLM performed the modification. Further, consistent with our results on yes/no and numeric questions, modifications by Gemini-1.5-Flash were the hardest to process (largest average % change) for all 3 VLMs while Claude-3.5-Sonnet modifications were the easiest to process (smallest average % change). There was no consistent pattern to indicate whether the models performed better/worse on modifications of open-ended questions compared to the yes/no, numeric questions. However since the open-ended questions are scored approximately, the results in Tables 1 and 2 may not be directly comparable.

Significance tests. We use a paired test to evaluate if the response of a VLM changes significantly when a modifier is added. Specifically, for yes/no and numeric questions since the answer can be compared exactly with the ground truth to obtain a binary outcome, we use the *McNemar's test*. The McNemar's exact test is used to evaluate if there is a significant difference in a dichotomous dependent variable between two groups. It is used frequently to evaluate drug effects (Trajman and Luiz, 2008), and has been shown to have low type I error (Dietterich, 1998). To run this test, we pair binary outcomes obtained by comparing the VLM's answer prior to and after question modification with the ground truth.

Our results showed that in most cases there was significant change in the VLM response ($p < 0.05$). However, when the modifiers were added using Claude-3.5-Sonnet, the change in responses of Claude-3.5-Sonnet/GPT-4o was insignificant ($p \geq 0.05$) which again indicates that Claude-3.5-Sonnet may be limited in its ability to add detailed modifiers. The responses of GPT-4o did

Model / Modifier	GPT-4o		Gemini-1.5-Flash		Claude-3.5-Sonnet	
	Visual	*Relational*	*Visual*	*Relational*	*Visual*	*Relational*
GPT-4o	1.06%	2.91%	**8.22%**	**8.22%**	-0.26%	3.18%
Gemini-1.5-Flash	8.71%	7.08%	**11.44%**	**13.07%**	5.17%	6.26%
Claude-3.5-Sonnet	4.86%	3.78%	**8.91%**	**6.21%**	1.35%	3.51%

Table 1: % change in accuracy for questions with yes/no answers and numeric answers (larger values indicate the model performed worse on modified questions). The column headings indicate which VLM was used to generate modified questions and the row headings indicate the VLM we are evaluating. The values in red show the worst performing VLM model/modifier combination when adding visual modifiers and the values in blue show the worst performing model/modifier combination for relational modifiers.

Model / Modifier	GPT-4o		Gemini-1.5-Flash		Claude-3.5-Sonnet	
	Visual	*Relational*	*Visual*	*Relational*	*Visual*	*Relational*
GPT-4o	4.58%	**6.87%**	**8.10%**	6.08%	1.96%	4.54%
Gemini-1.5-Flash	5.82%	4.91%	**8.13%**	**6.59%**	3.43%	6.31%
Claude-3.5-Sonnet	6.12%	5.96%	**8.72%**	**8.32%**	3.89%	7.16%

Table 2: % change in accuracy for questions with open-ended answers (larger values indicate the model performed worse on modified questions). The column headings indicate which VLM was used to generate modified questions and the row headings indicate the VLM we are evaluating. The values in red show the worst performing VLM model/modifier combination when adding visual modifiers and the values in blue show the worst performing model/modifier combination for relational modifiers.

not significantly change on self-modified questions ($p \geq 0.05$) with yes/no or numeric answers which again may indicate that GTP-4o performs well only when it has a strong prior on the type of modification. One alternate possible explanation is that perhaps GPT-4o stores the context in our interaction with it when generating modified questions and this somehow could influence its response to the modified questions (though we used a separate session for generating modified questions).

For open-ended questions, since the comparison between the ground truth and a VLM's answer does not yield a dichotomous value, we use the *Wilcox signed-rank* test (since the data was not normally distributed) instead of the McNemar's test. The results were very similar our findings with the McNemar's test. Claude-3.5-Sonnet/GPT-4o showed no significant change in responses ($p \geq 0.05$) when the modifier was Claude-3.5-Sonnet, and GPT-4o had insignificant change when answering self-modified questions. We plan to further investigate if there are specific linguistic characteristics of the modifiers that makes a question either harder or easier to answer.

4 Conclusion

In this work, we studied if VLMs are sensitive to modifications to questions in VQA. Specifically, adding modifiers increases details in a question, but when viewed from the perspective of cooperative principles, they can violate Grice's maxims. Humans can accurately ignore irrelevant details to answer questions even with these violations. We studied if VLMs could do the same in VQA by generating modified questions from human-crafted questions that preserve the original answer. We used 3 state-of-the-art VLMs in our study and showed that in most cases, adding modifiers to questions degrades the performance of the VLM. Based on these initial results, we plan to develop more detailed experiments to understand the types of modifications that VLMs are better at processing. Further, while our current results reveal that the performance of VLMs drops in the presence of modifiers, it is not yet clear as to why such a drop occurs. In future work, we plan to analyze the reasons for why some VLMs tend to perform more poorly than others in modified questions.

5 Limitations

Following are the limitations associated with this work.

1. This work assumes that human-written questions follow Grice's maxims of conversation. However, it may be the case that since humans are asking an AI a question (as op-

posed to talking to fellow humans), some of these maxims are violated even in human-generated questions.

2. Since the internal details of how VLMs handle prompts are not clearly known, there could be some bias associated with self-modified questions. That is, if a VLM tries to answer its own modified question since it would have access to the previous prompts (instructing it to add modifiers), it may be able use it in the response to modified questions. Even though, we provided the modification as a separate prompt, there could be some bias in the results of self-modified questions if the prompts are not completely independent.

3. Since open-ended questions do not have a unique ground truth answer, the evaluation we used may have a bias compared to those which have a unique ground truth answer.

Acknowledgments

This research was supported by NSF award #2008812, and awards from the Gates Foundation and Adobe. The opinions, findings, and results are solely the authors' and do not reflect those of the funding agencies.

References

Anthropic. 2024. Claude 3.5 sonnet. https://claude.ai. Large language model.

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.

William Britton, Somdeb Sarkhel, and Deepak Venugopal. 2022. Question modifiers in visual question answering. In *Language Resources and Evaluation Conference*.

Declan Campbell, Sunayana Rane, Tyler Giallanza, Camillo Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven Frankland, Tom Griffiths, Jonathan D Cohen, and 1 others. 2024. Understanding the limits of vision language models through the lens of the binding problem. *Advances in Neural Information Processing Systems*, 37:113436–113460.

Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, and 1 others. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10578–10587.

Bethan Davies. 2000. Grice's cooperative principle: Getting the meaning across. *Leeds Working Papers in Linguistics and Phonetics*, 8(1):26.

Thomas G Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923.

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.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Baptiste Jacquet, Jean Baratgin, and Frank Jamet. 2018. The gricean maxims of quantity and of relation in the turing test. In *2018 11th international conference on human system interaction (hsi)*, pages 332–338. IEEE.

Marcel Adam Just, Patricia A Carpenter, Timothy A Keller, William F Eddy, and Keith R Thulborn. 1996. Brain activation modulated by sentence comprehension. *Science*, 274(5284):114–116.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. 2024. Improving automatic vqa evaluation using large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4171–4179.

OpenAI. 2024. Hello gpt-4o (may 13 version). https://openai.com/index/hello-gpt-4o/. Large language model.

Ramprasaath R Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Tulio Ribeiro, Besmira Nushi, and Ece Kamar. 2020. Squinting at vqa models: Introspecting vqa models with sub-questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10003–10011.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

A Trajman and RR Luiz. 2008. Mcnemar $\chi 2$ test revisited: comparing sensitivity and specificity of diagnostic examinations. *Scandinavian journal of clinical and laboratory investigation*, 68(1):77–80.

Yiyu Wang, Jungang Xu, and Yingfei Sun. 2022. End-to-end transformer based model for image captioning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2585–2594.

Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10685–10694.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.

Appendix A: VLM Prompts

Prompt to generate modified questions targeting visual properties:

Instruction: Your task is to generate 1 different modified version of the original question about an image, ensuring that each modification preserves the original answer from the provided question and provide the type of the visual attribute that was added to the original question.

Given an image and its original question, create 1 unique modification by adding different types of visual attributes to the objects in the original question. The visual attributes can be of the following types:

- Physical properties (size, color, shape etc.) of the object
- Appearance characteristics (texture, pattern etc.) of the object
- Visual state (new, old, clean, dirty etc.) of the object

NOTE: You are not limited to the categories mentioned above. You are free to categorize as you see fit.

IMPORTANT: When adding visual attributes to questions, ensure that your modifications don't inadvertently reveal or hint at the correct answer.

**For the visual attribute categories, please use clear, specific labels such as:

- Color (when referring to color attributes)
- Texture (when referring to surface qualities)
- Size (when referring to dimensions)
- Shape (when referring to form)
- Pattern (when referring to visual arrangements)
- Visual state (when referring to condition)
- Physical property (when referring to other physical characteristics)

This helps maintain consistency in your categorization.**

Rules: Each modification MUST:

- Preserve the core meaning of the original question
- Yield the same answer as the original question
- Be distinctly different from other modifications
- Use natural, grammatically correct language

Avoid:

- Repeating the same modifier type across the 3 versions
- Making assumptions about details not visible in the image
- Changing the fundamental subject or action in the question

Output: Modified Questions [LIST]: [Question1] **Visual attribute [LIST]: [category1]**

Example 1: Original Question: Is the dog skateboarding? Modified Question [LIST]: [Is the small dog skateboarding?] Visual attribute [LIST]: [size]

Example 2: Original Question: Is there graffiti shown on the concrete wall? Modified Question [LIST]: [Is there colorful graffiti shown on the concrete wall?] Visual attribute [LIST]: [color]

IMPORTANT: When adding visual attributes to questions, ensure that your modifications don't inadvertently reveal or hint at the correct answer. The visual attributes should add detail without changing the difficulty level of the question or providing clues that make the answer obvious.

Prompt to generate modified questions targeting relational properties:

Your Task: Generate 1 different modified version of the provided question, ensuring that each modification uses a different relational modifier (positional relationships, for e.g. in front of, on, next to, in, etc.) while preserving the original answer.

Instruction: Given an original question, create 1 unique modification by adding different relational modifiers to the objects in the original question. Each modification must preserve the core meaning and yield the same answer as the original question.

Rules:

Each modification MUST:

- Use a different relational modifier (e.g., on, under, in front of, next to, in, among, etc.)
- Preserve the core meaning of the original question
- Yield the same answer as the original question
- Be distinctly different from other modifications
- Use natural, grammatically correct language

Avoid:

- Changing the fundamental subject or action in the question
- Making assumptions about details not provided in the original question
- Using non-relational modifiers (like color, size, shape, etc.)

Output:

Modified Questions [LIST]: [Question1] Relational Modifier [LIST]: [Modifier1]

Example: Original Question: Is the dog skateboarding? Modified Question [LIST]: [Is the dog skateboarding on the sidewalk?] Relational Modifier [LIST]: [on the sidewalk]

NOTE: DO NOT CHANGE THE MAIN CONTENT IN THE QUESTION. When adding relational attributes to questions, ensure that your modifications don't inadvertently reveal or hint at the correct answer. The relational attributes should add detail without changing the difficulty level of the question or providing clues that make the answer obvious.

Appendix B: AMT Details for verification

We used three workers to answer each question. Following is the instruction provided to AMT users to verify the modified questions generated by LLMs;

Instruction: You will see an image and two questions; Q1 (Original Question) and Q2 (Modified Question). The answer for Q1 is shown. Is the same answer correct for Q2?

Q1: Is there a coffee cup?
Answer: Yes

Q2: Is there a white coffee cup?
Answer: Yes

Select one of these options:
○ Correct Answer
○ Incorrect Answer
○ Answer is incorrect in both Q1 and Q2

We consider the modified questions that has the same answer or *correct* response from AMT users as the verified questions.