ConMe: Rethinking Evaluation of Compositional Reasoning for Modern VLMs
– Supplementary Material –

Anonymous Author(s)
Affiliation
Address
email

In the supplementary, we provide additional insights and supporting material for our ConMe dataset. First, we provide an overview of the three SugarCrepe [1] partitions (Section 1). Then, list the prompts used for Llama-3 [2] to generate the error partitions, and finally conclude with additional error analysis for different VLMs (Section 2).

To encourage reproducibility, our entire codebase to generate hard Compositional Reasoning (CR) Question and Answer (QA) pairs and the error category analysis for different VLMs is provided at the following GitHub repository: https://github.com/jmiemirza/ConMe. Furthermore, our ConMe dataset can also be accessed through the following HuggingFace Dataset Card: https://huggingface.co/conme/ConMe.

1 SugarCrepe Partitions

Our ConMe benchmark utilizes the partitions provided by SugarCrepe [1] dataset, which consists of 919 total images[1] – 333 from the *Replace-Att* partition, 333 from *Replace-Object*, 253 from *Replace-Relation*. SugarCrepe proposes to modify the positive caption of an image by either replacing, swapping, or adding atomic concepts – which are demonstrated through different dataset partitions – in order to confuse the VLMs. To avoid language errors, SugarCrepe employs an LLM for the atomic concept manipulation and follows the manipulation by LLM-based de-biasing (ensuring that the LLM has no bias towards the augmented or the original text), yet only on the text side, disregarding the image context. On the contrary, in our work, we focus on providing image context in addition to textual context, by employing a combination of different VLMs, rather than LLMs, to generate new questions and answer options.

Below we include a summary and description of these three partitions from the baseline SugarCrepe dataset, to provide additional context on the original structure:

- *Replace-Attribute* forms a negative by replacing the attributes describing object characteristics. As an example, for an image taken on the ground, two text options are: `{Several vehicles providing ground transportation are shown in the photo: streetcar, tour bus, classic car, and family cars.}` and `{Several vehicles providing aerial transportation are shown in the photo: helicopter, hot air balloon, small plane, and glider.}`. We observe, that the negative was generated by the LLM without any image context. Hence,

[1] sourced from the MS-COCO [3] validation set

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

```
You are an insightful assistant, for the question/answer pair provided by the user, pick a question format and question topic from the list below:
Question Format:
- hallucination: the question asks if something is visible or not, and the answer is NO, or that it is not visible/present (e.g. "Is there a cat in the room?" "No, there is no cat in the room.")
- misconception: the question asks about an attribute of an object, but that object is not present (e.g. "What color is the cat?" "There is no cat.")
- non-determinable: the question asks for something that cannot be distinguished (e.g. Is the cat in motion? "I cannot tell." OR "It is unclear.")
- selective: any other questions that do not fall into the above categories
Question Topics:
- lighting: the question asks about the lighting or direction of the light (e.g. "Is the cat's shadow sharp?" "No, the shadow is diffused.")
- clothing: the question asks about an what is being worn (e.g. "Is the cat wearing a hat?" "No, the cat is not wearing a hat.")
- attribute: the question asks about the presence or visibility of an attribute of an object (e.g. "Does the cat have white whiskers?" "No, the cat has black whiskers.")
- emotion: the question asks an opinion of what is observed (e.g. "What makes this room cozy?" "The fireplace makes the room cozy.")
- attention: the question asks about the attention of a person or object (e.g. "Which direction is the cat looking?" "The cat is looking out the window.")
- color: the question asks about the color of an object (e.g. "What color is the cat?" "The cat is black.")
- scene: the question asks about the location of the scene (e.g. "Is this indoor or outdoor?" "This is indoor.")
- count: the question asks about the number of objects (e.g. "How many cats are there?" "There are two cats.")
- behavior: the question asks about action or behavior (e.g. "Is the moving around?" "No, the cat is sleeping.")
- proximity: the question asks about the spatial relation between two objects (e.g. "Is the cat near the window?" "Yes, the cat is near the window.")
Do not confuse formats with topics.
Respond with a JSON object with the following format:
{
    "question_format": "format",
    "question_topic": "topic"
}
```

Figure 1: The complete prompt to the Llama-3 [2] used to classify different questions in the ConMe dataset according to the question format and question topic for analysis of VLM errors.

despite the linguistic correctness, it is unlikely a hard negative for a VLM provided with the image context of a ground.

- *Replace-Object* refers to negative generation via replacing the object (noun) in the positive caption. For example, given an image of a teddy bear next to some boxes in a room, a VLM is asked to choose between the *positive* `{A big teddy bear sitting next to some boxes.}` and the *negative* `{A big car sitting next to some boxes.}`. Even though the negative is grammatically correct and potentially unbiased given the partial context (a room is not mentioned in the positive text), we would not expect a car to sit next to the boxes in a room (though it might happen near the side of the road). As follows, it is unlikely that a modern VLM would be confused, as it can complete the missing details (the room) from the image and infer the unlikelihood of a car there based on the image context.
- *Replace-Relation* replaces a word describing a spatial relation between objects in a caption to form the negative. For example, given an image taken in a bedroom, the VLM is required to choose between `{A black bike rests against a brown bed.}` and `{A black bike hangs from a brown bed.}`. Similarly, in the bedroom context (observed by the VLM, but hidden from the LLM that produced the "hangs from" negative), this might be an easy choice for a VLM.

2 Error Partition Analysis

In the main manuscript (Section 5.2), we analyzed the different types of errors VLMs made on the manually verified ConMe dataset subset, after dividing the questions into different partitions. These partitions are acquired by employing Llama-3 [2] as a classifier. The complete prompt to the Llama-3 [2] model is listed in Figure 1. Furthermore, we provide the sample count in the classified categories in Figure 2. We observe that according to the question topic, a large number of samples (38.4%) are classified as describing an attribute.

References

[1] C.-Y. Hsieh, J. Zhang, Z. Ma, A. Kembhavi, and R. Krishna, "Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality," *arXiv preprint arXiv:2306.14610*, 2023.

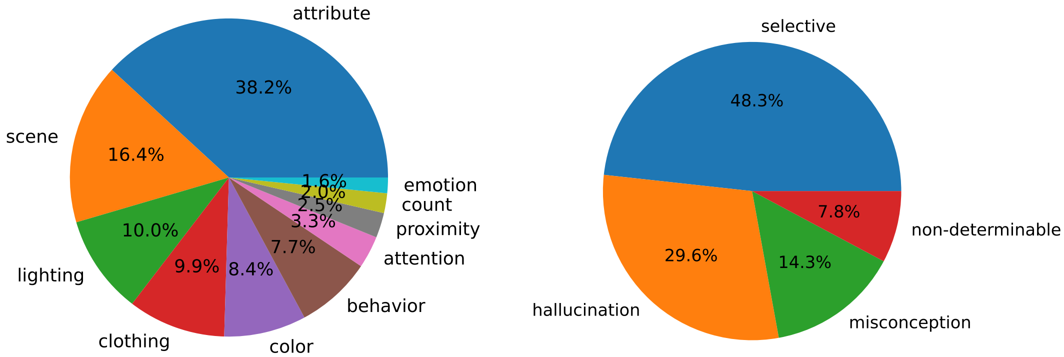

Figure 2: Percentage of samples belonging to different categories classified by Llama-3, according to CR Q/A topic (left) and CR Q/A format (right).

[2] AI@Meta, "Llama 3 model card," 2024. [Online]. Available: `https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md`.

[3] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, Springer, 2014, pp. 740–755.