

Prompting Large Vision-Language Models for Compositional Reasoning

Timothy Ossowski¹, Ming Jiang³, Junjie Hu^{1,2}

¹Department of Computer Science, ²Department of Biostatistics and Medical Informatics
University of Wisconsin, Madison, WI, USA

³ Department of Human-centered Computing, Indiana University, Indianapolis, IN, USA
ossowski@wisc.edu, mj200@iu.edu, junjie.hu@wisc.edu

Abstract

Vision-language models such as CLIP have shown impressive capabilities in encoding texts and images into aligned embeddings, enabling the retrieval of multimodal data in a shared embedding space. However, these embedding-based models still face challenges in effectively matching images and texts with similar visuo-linguistic compositionality, as evidenced by their performance on the recent Winoground dataset. In this paper, we argue that this limitation stems from two factors: the use of single vector representations for complex multimodal data, and the absence of step-by-step reasoning in these embedding-based methods. To address this issue, we make an exploratory step using a novel *generative* method that prompts large vision-language models (e.g., GPT-4) to depict images and perform compositional reasoning. Our method outperforms other embedding-based methods on the Winoground dataset, and obtains further improvement of up to 10% accuracy when enhanced with the optimal description.¹ More importantly, we provide a fine-grained error analysis of our method’s outputs, highlighting the key bottleneck in understanding image contents by existing VLMs.

1 Introduction

Recent advancements in vision-language models (VLMs) have rapidly accelerated progress in multimodal tasks such as visual question answering (Antol et al., 2015) and image captioning (Lin et al., 2014). Large vision-language encoders such as CLIP (Radford et al., 2021) and UNITER (Chen et al., 2020) have been trained to learn a joint embedding space for combining visual and textual information. These aligned multimodal embeddings have been widely used for zero-shot image-text retrieval (Plummer et al., 2015) and other challenging multimodal matching tasks (Thrush et al., 2022).

Notably among these approaches, CACR (Pandey et al., 2022) and IAIS (Ren et al., 2021) further improve the multimodal representations by incentivizing relation alignment during pretraining.

Despite remarkable advances, the embedding-based methods still encounter difficulties in various compositional reasoning tasks, particularly in the recent Winoground task (Thrush et al., 2022). This task evaluates the capability of VLMs to understand the compositional relation among objects in the images and their associated captions (see details in §3). One primary limitation of embedding-based methods is their reliance on compressing intricate compositional information from an image or a text into a *single vector representation*, which typically emphasizes object encoding but is limited in distinguishing the nuanced relationships between objects and their context in the image and caption. To address this limitation, we propose an *alternative generative* approach that utilizes the fine-grained reasoning capabilities of large generative models in comprehending image content and matching corresponding texts. In contrast to traditional generative methods that train specific modules for visual question answering (Wang et al., 2022b; Uehara et al., 2022), we use a *tuning-free prompt-based* method.

Specifically, our *keyword-guided compositional reasoning* method (KEYCOMP) prompts a VLM for depicting images based on the keywords detected from the query text, and then uses a stronger LLM to analyze the image description for matching corresponding texts. Our method design is mainly based on two considerations. First, our analysis (§4) shows that directly prompting generative VLMs like MiniGPT4 (Zhu et al., 2023) or BLIP-2 (Li et al., 2023) still poses a challenge for the model in identifying key image and text contents for further reasoning. Thus, proper guidance is necessary to instruct the VLM to focus on key image regions for image description. Second, we believe that existing LLMs (e.g., GPT-4) possess

¹Code is available at <https://github.com/tossowski/KeyComp>.

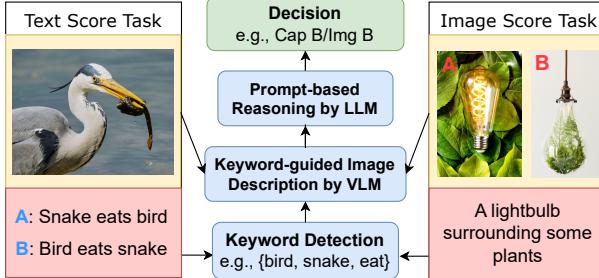


Figure 1: Illustration of our generative method for the Winoground task. Appendix C shows more detailed descriptions and model outputs. **Text Score Task:** Our method chooses the more appropriate caption given a single image. **Image Score Task:** Our method chooses the best image given a single caption.

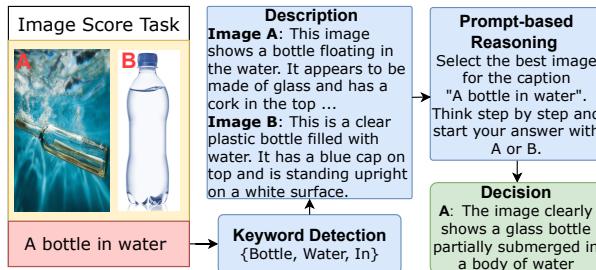


Figure 2: A detailed example for the image score task.

stronger language reasoning capabilities than the available VLMs, thus we use a LLM for multi-step reasoning instead of a weaker VLM used in concurrent prompting-based methods (You et al., 2023; Shen et al., 2023; Wu et al., 2023).

We conduct thorough quantitative and qualitative analyses of our method against existing embedding-based methods on Winoground. Overall, KEYCOMP achieves a state-of-the-art image score on Winoground, surpassing the best embedding-based method by a clear margin of 5.1% image score. Our method excels at multi-step reasoning for complex examples (Fig. 9) and unusual images (Fig. 17). More importantly, further error analysis of the failure cases (Appendix G) reveals a bottleneck in the image description quality of VLMs, shedding insights for future directions. Particularly, VLMs still struggle with describing spatial reasoning and LLMs may misinterpret VLM descriptions with complex syntax structures.

2 Method

This section first describes the Winoground tasks, and then introduces KEYCOMP in three steps (illustrated in Fig. 1 and Fig. 2).

Problem Definition Winoground (Thrush et al., 2022) has two fundamental matching tasks to as-

sess the visio-linguistic compositionality of VLMs. Given two aligned image-caption pairs (I_0, C_0) and (I_1, C_1) , a VLM is asked to select the correct caption from $\{C_0, C_1\}$ for a single image (either I_0 or I_1) in the text score task. In the image score task, a VLM is asked to select the correct image from $\{I_0, I_1\}$ for a single caption (either C_0 or C_1).

Step 1: Keyword Detection Our method first detects a keyword set $\mathcal{K} = \{w_1, \dots, w_k\}$ automatically from the caption text to help guide the VLM to describe the relevant contents of the images. The caption text is either a single caption in the image task or the concatenation of two captions in the text task. When detecting keywords, we use SpaCy² to identify nouns, verbs, prepositions, and adjectives. The intuition is that nouns in a sentence usually describe the key entities in the scene, and adjectives depict the attributes of the entities (e.g., color, size, shape), while verbs and prepositions usually reveal a relation between two entities in an image.

Step 2: Keyword-guided Description Next, we use a pretrained VLM f_{VLM} such as MiniGPT4 or BLIP-2 to generate text descriptions for images. To ensure the VLM focuses on relevant image details for image-text matching, we use the keywords \mathcal{K} to guide the generation of image descriptions. Specifically, we append all extracted keywords in \mathcal{K} to the end of a text “[Instruction]” to produce a prompt $P_{\mathcal{K}}$. The prompt is then used to generate a high-quality description of the entities and relations for an image I_a by Eq. (2). Example descriptions and prompts are presented in Appendix C and D.

$$P_{\mathcal{K}} = “[Instruction] w_1, \dots, w_k” \quad (1)$$

$$D_{\mathcal{K}, a} = f_{\text{VLM}}(I_a, P_{\mathcal{K}}), \quad a \in [0, 1] \quad (2)$$

Step 3: LLM Reasoning & Explanation As LLMs have demonstrated an impressive zero-shot language reasoning ability (Brown et al., 2020), we prompt a LLM f_{LLM} to perform reasoning on the generated image descriptions and the given captions, and select an answer for Winoground tasks. Additionally, inspired by recent chain-of-thought prompting (Wei et al., 2022), we also add another explanation instruction, such as “Think step by step”, to prompt the LLM to explain its answer selection. Specifically, in the image task for selecting the correct image from I_0, I_1 for a caption C_a , we construct a prompt $P_{\text{img}, a}$ by concatenating

²<https://spacy.io/>

C_a with a selection instruction, the two generated image descriptions and an explanation instruction. Similarly, we construct another prompt $P_{\text{txt},a}$ for an image I_a in the text task. Finally, we feed the prompt to the LLM to get a text output y containing the selection and the explanation.

$$P_{\text{img},a} = "[\text{T}] [C_a] [\text{T}] [D_{\mathcal{K},0}] [D_{\mathcal{K},1}] [\text{T}]" \quad (3)$$

$$P_{\text{txt},a} = "[\text{T}] [D_{\mathcal{K},a}] [\text{T}] [C_0] [C_1] [\text{T}]" \quad (4)$$

$$y = f_{\text{LLM}}(P_{t,a}), \quad t \in \{\text{img, txt}\} \quad (5)$$

where all “[T]” placeholders are the texts surrounding the key information in the template to construct the prompt. Appendix B shows all prompt variants.

3 Experimental Settings

Dataset & Evaluation The Winoground dataset consists of 400 items, each containing two image-caption pairs $(I_0, C_0), (I_1, C_1)$. While the images in each item may be completely different, the two captions $\{C_0, C_1\}$ have an identical set of objects, only in a different order. A model is evaluated by the following text, image, and group scores.

- **Text Score:** The model is asked to pick the corresponding caption from $\{C_0, C_1\}$ for a single image I_a . The model gets a score of 1 if and only if it picks the correct caption for both I_0 and I_1 .
- **Image Score:** The model is asked to pick the corresponding image from $\{I_0, I_1\}$ for a caption C_a . The model gets a score of 1 if and only if it picks the correct image for both C_0 and C_1 .
- **Group Score:** The model achieves a group score of 1 for the item if and only if it receives a text score of 1 and an image score of 1.

Methods in Comparison We compare our generative method with strong embedding-based methods, i.e., IAIS (Ren et al., 2021), CACR (Pandey et al., 2022) and CLIP (Radford et al., 2021). These methods choose the better caption or image by computing a similarity score between their multimodal embeddings and selecting the one with the highest score. In contrast, our generative method generates a text output for selection. We use string matching on the model output to extract the selection for evaluation. If the LLM generates an invalid output (e.g., “Neither”) indicating neither answer matches the query, we consider the prediction incorrect despite a slight underestimation of our method. Our experiments show that this invalid output occurs about 10% of the time in the test set.

Model Selection & Hyperparameters To produce higher-quality descriptions of images, we use an instruction-tuned VLM, i.e., MiniGPT4 (Zhu et al., 2023) with a backbone LLM of Vicuna-13b. All descriptions are generated using a temperature of 1 and a beam size of 10. With a lower beam size and temperature, we observe notably worse descriptions. For the reasoning component, we utilize OpenAI’s ChatGPT API with the GPT-3.5-turbo and GPT-4 models using a temperature of 1 and the default values of the other hyperparameters.

4 Results and Analysis

Overall Performance Table 1 reports the overall text, image, and group scores of our method compared with existing embedding-based approaches. Since LLMs are generative probabilistic models, they do not have deterministic outputs. Therefore, we report the average score across 3 different runs of the LLM reasoning (Step 3) along with the standard deviation of our method. We observe several findings: (1) Most notably, our method achieves a significantly better image score than prior works, reaching a new state-of-the-art image score. (2) Our text and group scores are also competitive with existing works, even when evaluating accuracy with strict string matching. (3) Despite the stochastic nature of our method, the standard deviation of LLM reasoning between runs rarely exceeds 1 – 2%. (4) However, we observe a significant difference in the image description quality from the VLM (Step 2). To estimate an upper bound, we instruct MiniGPT4 to sample 5 descriptions and manually select the best image description to feed into LLMs for reasoning. This leads to a large gain of 12.4% text score and 3.2% image score, indicating a potential direction of automatically selecting the best image descriptions for improvement.

Method	LLM	Text	Image	Group
KEYCOMP *	GPT-4	43.5 ± 0.7	28.7 ± 2.1	18.2 ± 0.9
KEYCOMP *	GPT-3.5	42.7 ± 0.8	27.8 ± 0.7	17.4 ± 0.3
KEYCOMP	GPT-3.5	30.3 ± 1.6	24.6 ± 1.2	12.4 ± 1.2
IAIS	-	42.5	19.5	16.0
CACR	-	39.3	17.8	14.3
CLIP	-	30.8	10.5	8.0

Table 1: Text, Image, and Group scores scaled by 100 in Winoground. We report the average scores and standard deviation in 3 different runs of LLM reasoning (Step 3). Asterisk indicates an upper bound, where we select the best caption among 5 generated descriptions.

Image Description Quality Matters. Fig. 3 displays the average text score of our method on var-

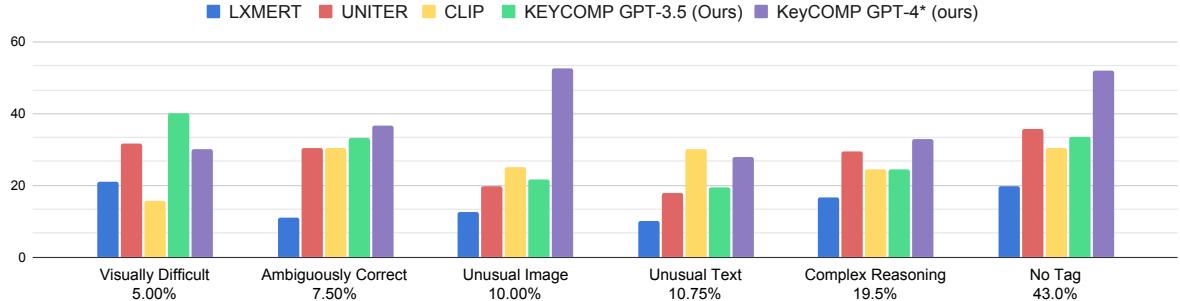


Figure 3: Fine-grained text score performance across different question categories. We give specific examples from each category in Appendix C. Percentages on the x-axis indicate each question type’s proportion of the dataset. To ensure representative results, question categories comprising less than 5% of the dataset are excluded.

ious question categories curated by Diwan et al. (2022). We provide correct examples from each category in Appendix C, D, and the image score results in Appendix E. Notably, keyword guidance improves image description quality for generative approaches. When providing high-quality image descriptions, our method obtains significant gains in the categories of unusual images (Fig. 17) and complex reasoning (Fig. 9). For instance in Fig. 9, when a VLM depicts a rabbit and a turtle correctly, the LLM has the commonsense knowledge to pick the faster one. When prompted with keywords (e.g., “people” and “windows”) in Fig. 17, the VLM is more likely to describe an unusual crayon drawing correctly. This suggests that KEYCOMP has the potential to surpass embedding-based methods when external knowledge is necessary for complex reasoning or when precise image information from keywords is required for unusual images.

Error Analysis and Findings We also manually perform fine-grained error analysis on our method’s outputs and illustrate three main categories of errors in Appendix G. We identify a bottleneck of image content understanding capability of VLMs and highlight three key findings: 1) the VLM often struggles to describe spatial relationships between two objects, especially for two objects with similar colors (e.g., bushes and hedge in Fig. 19); 2) the LLM occasionally infers the wrong answer due to the misinterpretation of a detailed VLM description with a complicated syntax structure (see Fig. 20); 3) the VLM produces inaccurate descriptions of scene elements that are out of focus or missing parts of objects (e.g., a man’s head is outside of Fig. 21).

Prompt and Model Ablations Table 2 presents the effect of different prompts on our method. We observe that guiding VLM output with keywords results in the most significant improvement (② vs ⑤). Tuning the prompts for ChatGPT (③-⑤) pro-

vides further gains, suggesting future improvement with more sophisticated prompting. Prompting the LLM with chain of thought instructions yields a small gain of 1.3% text score and 2.3% image score over a simpler selection prompt (⑤ vs ③). Using a VLM to answer questions directly (①) leads to inferior performance, with only 2.0 group score and 11.9% lower image score than KEYCOMP. We also explore different VLM/LLM sizes in Appendix F.

ID	Prompt Method	Text	Image	Group
①	MiniGPT4 end-to-end	20.8 ± 0.3	12.5 ± 1.7	2.0 ± 0.3
②	No Keyword Guidance	21.6 ± 1.0	21.8 ± 0.8	9.4 ± 1.0
③	Multiple Choice	29.0 ± 0.3	22.3 ± 0.4	11.9 ± 0.4
④	Explanation Prompting	26.3 ± 1.3	23.9 ± 1.5	11.5 ± 0.6
⑤	KEYCOMP	30.3 ± 1.6	24.6 ± 1.2	12.4 ± 1.2
⑥	Multiple Choice*	35.2 ± 1.2	27.5 ± 0.7	14.8 ± 0.7
⑦	Explanation Prompting*	36.1 ± 0.4	28.8 ± 0.6	15.9 ± 0.6
⑧	KEYCOMP *	42.7 ± 0.8	27.8 ± 0.7	17.4 ± 0.3

Table 2: Average performance and standard deviation of prompt variants. To isolate the effect of the prompt, the bottom three rows use the same image description. We show detailed prompt templates in Appendix B. Asterisk indicates an upper bound, where we select the best caption among 5 generated MiniGPT4 descriptions.

5 Discussion and Future Works

We propose to use VLMs to describe key entities and relations in images to perform compositional reasoning. We empirically demonstrate the effectiveness of our method against embedding-based approaches and end-to-end VLM methods on Winoground. More importantly, we showcase successful examples, categorize common errors made by generative approaches, and identify a key bottleneck of image content understanding of existing VLMs. We hope to shed insights into future works of image content understanding regarding (1) effective prompting strategies to guide VLMs to focus on key image regions; (2) spatial reasoning of objects by VLMs; and (3) accurate interpretation of out-of-focus or partial objects by VLMs.

Limitations

Our method shows promising results over embedding based methods and end-to-end VLMs. However, small errors in the early steps can accumulate and lead to erroneous reasoning. Using VLMs in a pipeline, as reported by (You et al., 2023), is often limited by their performance. This bottleneck explains why we noticed significant improvement when we manually selected the best descriptions generated by MiniGPT4. Additionally, designing a universal prompt for image reasoning can be challenging, and the optimal prompt may change with model updates. Appendix B illustrates the effect of changing prompts, and future work may consider automatic prompt learning techniques such as prefix tuning (Li and Liang, 2021). Lastly, our method uses a non-deterministic reasoner, resulting in slight output variations even with 0 temperature. Although the variance is low, this may pose issues for downstream tasks.

References

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.
- 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.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Jun Chen, Deyao Zhu, Kilichbek Haydarov, Xiang Li, and Mohamed Elhoseiny. 2023. Video chatcaptioner: Towards the enriched spatiotemporal descriptions. *arXiv preprint arXiv:2304.04227*.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.
- Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, pages 104–120. Springer.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.
- Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. 2022. Why is winoground hard? investigating failures in visuolinguistic compositionality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.
- Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962.
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*.
- Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975.
- Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 2021 Conference of the Association for Computational Linguistics*.
- Tsung-Yi Lin, Michael Maire, Serge 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 13*, pages 740–755. Springer.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.

- Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023. Chameleon: Plug-and-play compositional reasoning with large language models. *arXiv preprint arXiv:2304.09842*.
- Oscar Mañas, Pau Rodriguez, Saba Ahmadi, Aida Nematzadeh, Yash Goyal, and Aishwarya Agrawal. 2023. Mapl: Parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Rohan Pandey, Rulin Shao, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. Cross-modal attention congruence regularization for vision-language relation alignment. *arXiv preprint arXiv:2212.10549*.
- Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.
- Shuhuai Ren, Junyang Lin, Guangxiang Zhao, Rui Men, An Yang, Jingren Zhou, Xu Sun, and Hongxia Yang. 2021. Learning relation alignment for calibrated cross-modal retrieval. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics*.
- Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueteng Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*.
- Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650.
- Sanjay Subramanian, Medhini Narasimhan, Kushal Khangaonkar, Kevin Yang, Arsha Nagrani, Cordelia Schmid, Andy Zeng, Trevor Darrell, and Dan Klein. 2023. Modular visual question answering via code generation. In *Proceedings of the 2023 Conference of the Association for Computational Linguistics*.
- Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*.
- Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.
- Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 200–212.
- Kohei Uehara, Nan Duan, and Tatsuya Harada. 2022. Learning to ask informative sub-questions for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4681–4690.
- Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022a. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR.
- Ruonan Wang, Yuxi Qian, Fangxiang Feng, Xiaojie Wang, and Huixing Jiang. 2022b. Co-VQA : Answering by interactive sub question sequence. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2396–2408, Dublin, Ireland. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in neural information processing systems*.
- Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. 2023. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.
- Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

Haoxuan You, Rui Sun, Zhecan Wang, Long Chen, Gengyu Wang, Hammad A Ayyubi, Kai-Wei Chang, and Shih-Fu Chang. 2023. Idealgpt: Iteratively decomposing vision and language reasoning via large language models. *arXiv preprint arXiv:2305.14985*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Appendix

A Related Work

Vision Language Models Language Models and Vision Models by themselves have shown impressive performance in their respective tasks, but combining them to perform vision-language reasoning remains a challenging problem. Popular encoder-based VLMs such as CLIP (Radford et al., 2021) or ALIGN (Jia et al., 2021) perform contrastive learning on large datasets of image-text pairs, displaying remarkable zero-shot transfer to unseen tasks such as image classification (Deng et al., 2009) and image-text retrieval (Plummer et al., 2015). GLIP (Li et al., 2022) improves the pretraining stage of CLIP by introducing phrase level grounding, allowing for richer semantic representations for the object detection task (Ren et al., 2015). Recent sequence-to-sequence VLMs such as OFA (Wang et al., 2022a) and FLAVA (Singh et al., 2022) pretrain on a larger variety of vision-language tasks and objectives to allow for more sample-efficient pretraining and cross-task trasnfer.

Other recent works combine the information from different modalities by introducing lightweight tunable parameters to connect frozen image and text encoders. Flamingo (Alayrac et al., 2022) freezes a vision encoder and inserts trainable cross attention layers to a frozen LLM to integrate visual features, achieving state of the art few-shot performance. BLIP-2 (Li et al., 2023) also freezes a vision encoder and LLM, aligning the modalities with a lightweight transformer called the Q-former. Works such as *Frozen* (Tsimpoukelli et al., 2021) and MAPL (Mañas et al., 2023) propose methods which freeze a LLM, but train a visual encoder to represent images as continuous embeddings, resulting in a multimodal few-shot learner. Inspired by the success of instruction tuning LLMs with human feedback with ChatGPT (Ouyang et al., 2022), MiniGPT4 (Zhu et al., 2023) and LLaVA (Liu et al., 2023b,a) further improves the quality of BLIP-2 outputs and LLaMA-2 by instruction tuning with additional descriptions generated by ChatGPT.

Large Language Models as Reasoners Concurrent work has explored the idea of using large language models to connect visual foundation models. For example, HuggingGPT (Shen et al., 2023), CHAMELEON (Lu et al., 2023), MM-React (Yang et al., 2023), and Visual ChatGPT (Wu et al., 2023) utilize ChatGPT as a controller, allowing it to delegate visually demanding tasks to foundational visual models. IdealGPT (You et al., 2023) proposes an iterative approach to decompose complex visual tasks into a series of sub-questions and answers. ChatCaptioner (Chen et al., 2023) has BLIP-2 interact conversationally with ChatGPT to create more informative image descriptions. ViperGPT (Surís et al., 2023), VisProg, (Gupta and Kembhavi, 2023), and CodeVQA (Subramanian et al., 2023) use code LLMs such as CodeX (Chen et al., 2021) to write python programs which invoke VLMs to solve challenging visual tasks without any task-specific training. Different from these works, we replace the ChatGPT controller with keyword-guided detailed descriptions of the contents of an image, delegating reasoning to a more powerful LLM instead of a VLM. We also avoid iteratively prompting our models to avoid potential error accumulation.

B Prompt Variations

In Tables 3 and 4 we display all the different prompt variations used in our experiments.

Name	VLM Prompt ($P_{\mathcal{K}}$)	ChatGPT Prompt ($P_{\text{txt},a}$)	Performance
KEYCOMP	Describe the image in detail using these keywords: {keywords}	Select the best caption for this image: A: {caption_0} B: {caption_1} Think step-by-step and start your answer with A or B. Even if you are unsure make a guess.	30.3
Explanation Prompting	Describe the image in detail using these keywords: {keywords}	Select the best caption for this image: A: {caption_0} B: {caption_1} Start your answer with A or B. Even if you are unsure make a guess. Briefly explain your decision in 1-2 sentences.	26.3
Multiple Choice	Describe the image in detail using these keywords: {keywords}	Select the best caption for this image: A: {caption_0} B: {caption_1} Start your answer with A or B. Even if you are unsure make a guess.	29.0
No Keyword Guidance	Describe the image in detail	Select the best caption for this image: A: {caption_0} B: {caption_1} Start your answer with A or B. Even if you are unsure make a guess. Briefly explain your decision in 1-2 sentences.	21.6
MiniGPT4 End-to-End	Which caption is more appropriate: A: {caption_0} B: {caption_1} Answer in one sentence. Even if you are unsure make a guess. Briefly explain your decision in 1-2 sentences.	None	20.8

Table 3: Average performance of different prompt variants for text score across 3 different runs.

Name	VLM Prompt ($P_{\mathcal{K}}$)	ChatGPT Prompt ($P_{\text{img},a}$)	Performance
KEYCOMP	Describe the image in detail using these keywords: {keywords}	Statement: {caption} Image A: {description_0} Image B: {description_1} Think step by step and fill in the blank: Situation {{}} is most consistent with the statement {caption} because ...	24.6
Explanation Prompting	Describe the image in detail using these keywords: {keywords}	Statement: {caption} Image A: {description_0} Image B: {description_1} Fill in the blank: Situation {{}} is most consistent with the statement {caption} because ... Explain your decision in 1-2 sentences.	23.9
Multiple Choice	Describe the image in detail using these keywords: {keywords}	Image 1: {image_0_description} Image 2: {image_1_description} Consider the caption: {caption} Select the better image for this caption: A: {image_1} B: {image_2} Start your answer with A or B.	22.3
No Keyword Guidance	Describe the image in detail	Statement: {caption} Image A: {description_0} Image B: {description_1} Think step by step and fill in the blank: Situation {{}} is most consistent with the statement {caption} because ...	21.8
MiniGPT4 End-to-End	Given the following images: A: ImageContent B: ImageContent Which image is more appropriate for the caption {caption}? Answer with A or B.	None	12.5

Table 4: Average performance of different prompt variants for image score across 3 separate runs.

C Text Score Task for Question Categories

We compile several examples of our method, with one example from each question category defined by a recent study (Diwan et al., 2022).

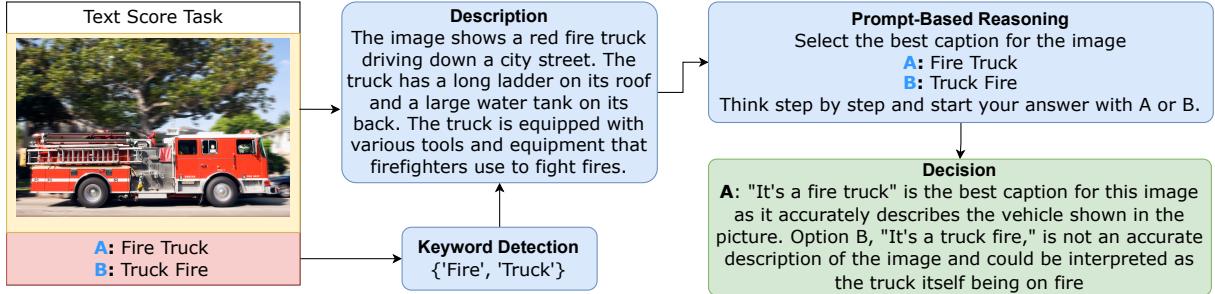


Figure 4: Non-Compositional Question. The swapped words ("Fire" and "Truck") do not necessarily contain the same semantic entities, so compositional reasoning may not be required to solve the question.

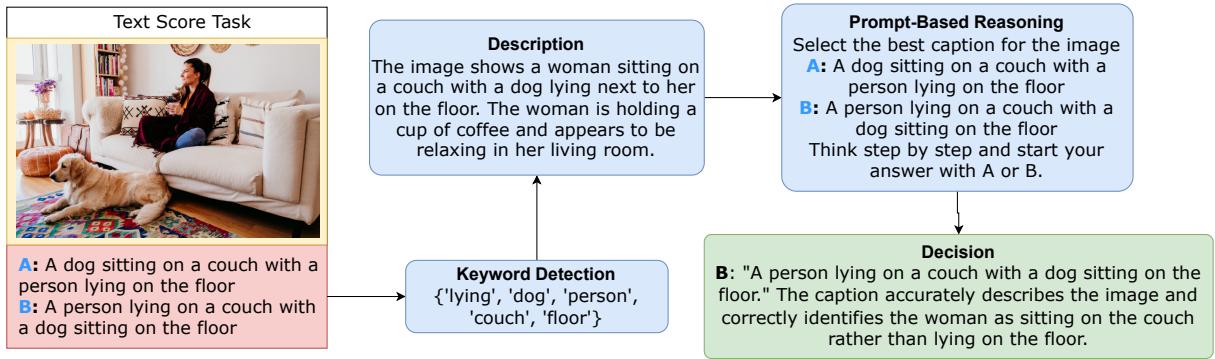


Figure 5: Ambiguously Correct Question. Note that the correct caption B describes the woman as lying on the couch when she is sitting, but the LLM is still able to pick the ambiguously correct caption.

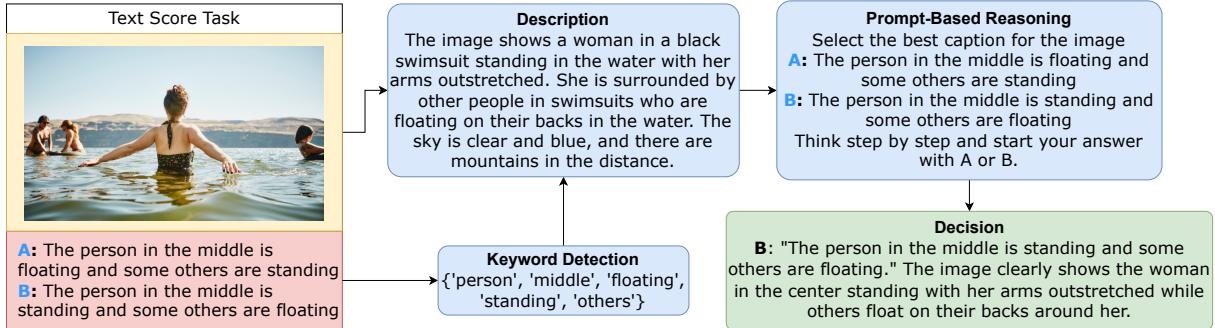


Figure 6: Visually Difficult Question. To correctly answer, the model needs to identify the person floating on the right.

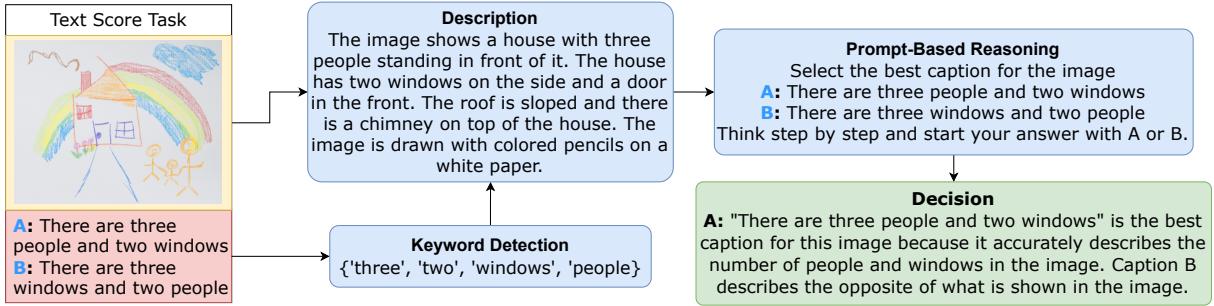


Figure 7: Unusual Image Question. The VLM is more likely to describe an unusual crayon drawing correctly, based on the detected keywords “windows” and “people” in the prompt, while embedding-based methods are not feasible to include any hints, and are sensitive to domain shifts in the image distribution.

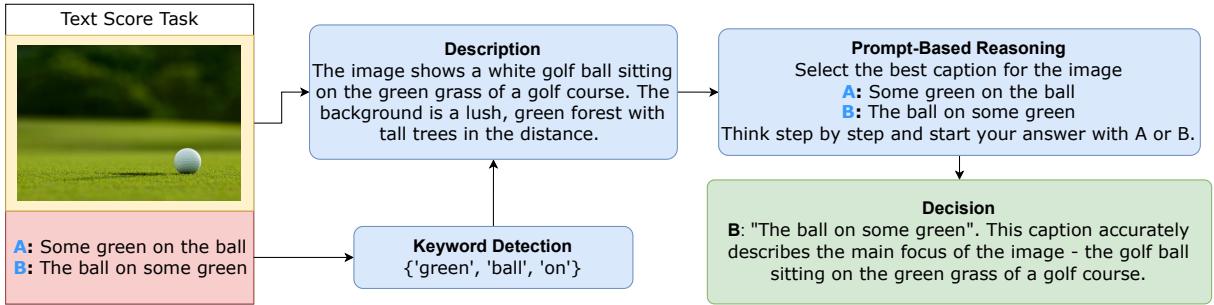


Figure 8: Unusual Text Question. The caption “The ball on some green” is considered less common compared to an alternative such as “the ball on some grass”.

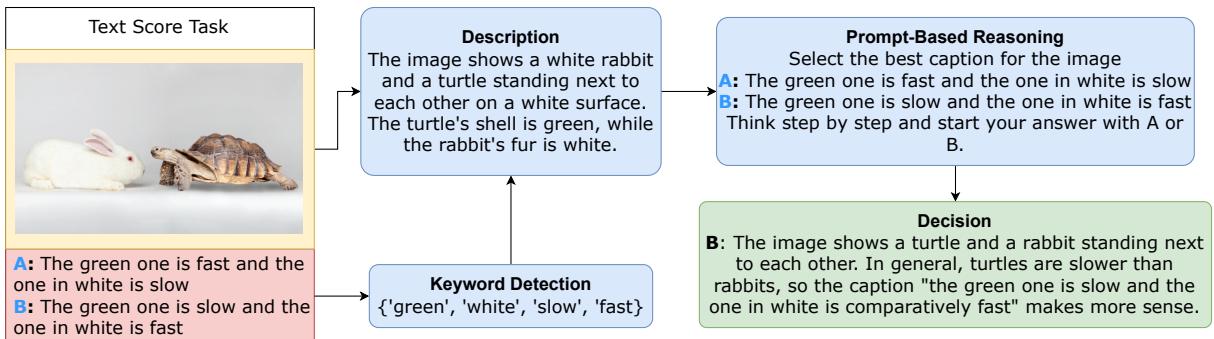


Figure 9: Complex Reasoning Question. This requires the knowledge that rabbits are typically faster than turtles.

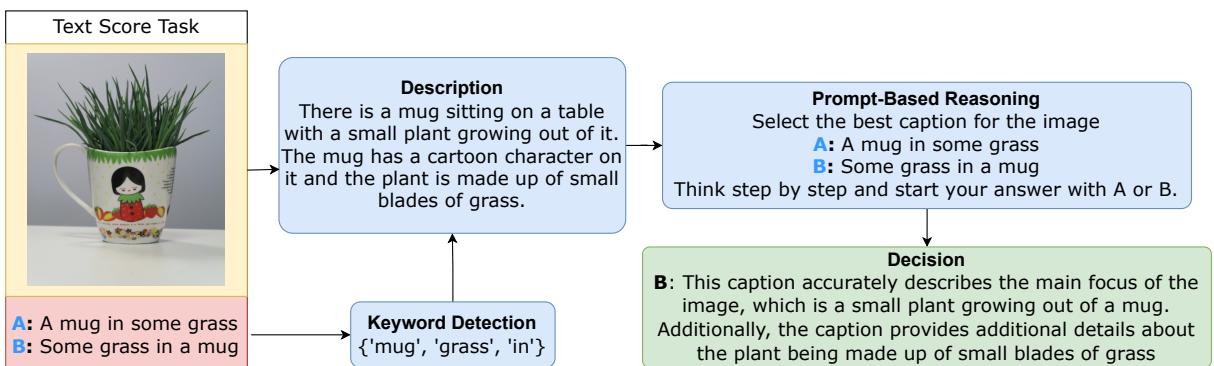


Figure 10: No Tag Question.

D Image Score Task for Question Categories

In this section we illustrate the details of our method when applied to the image score task. We include one example from each question category, using the same image caption pairs used in Appendix C.

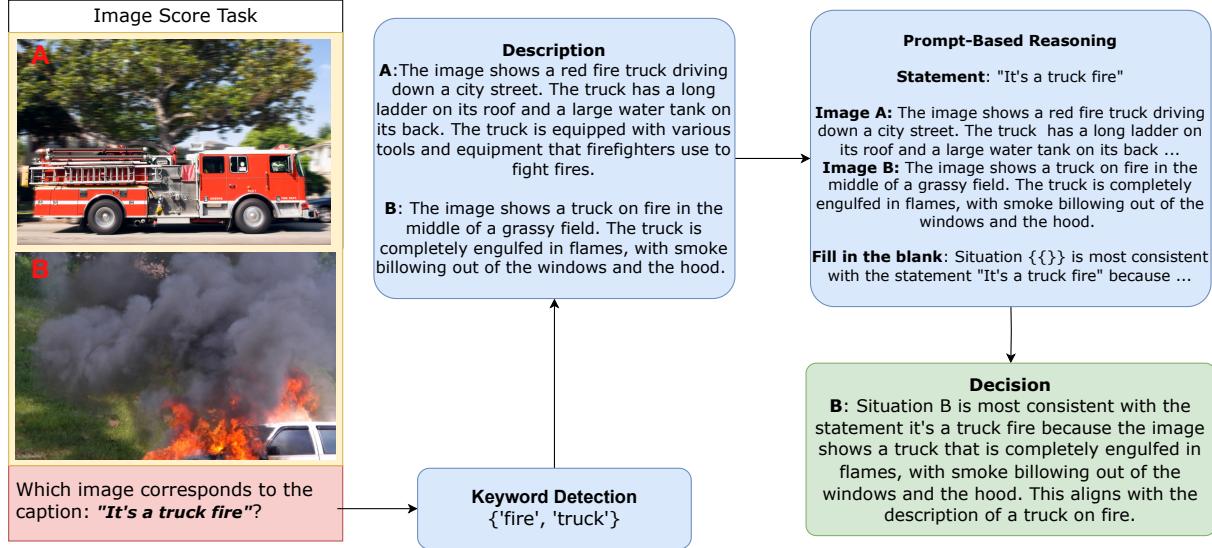


Figure 11: Non-Compositional Question for the image score task.

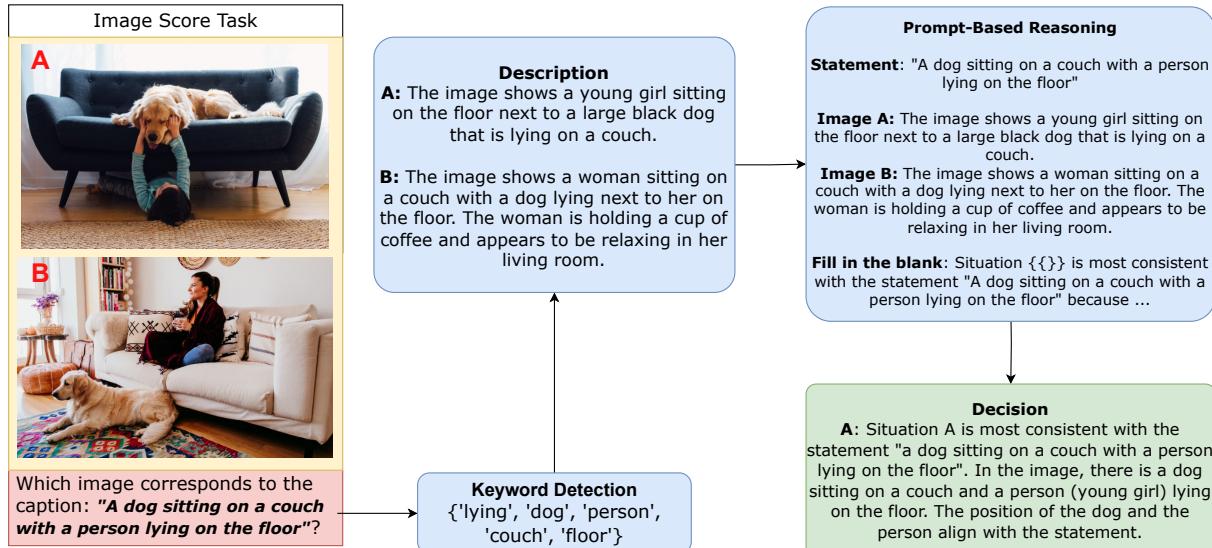


Figure 12: Ambiguously Correct Question. Refer to Appendix C for the ambiguity contained in image B.

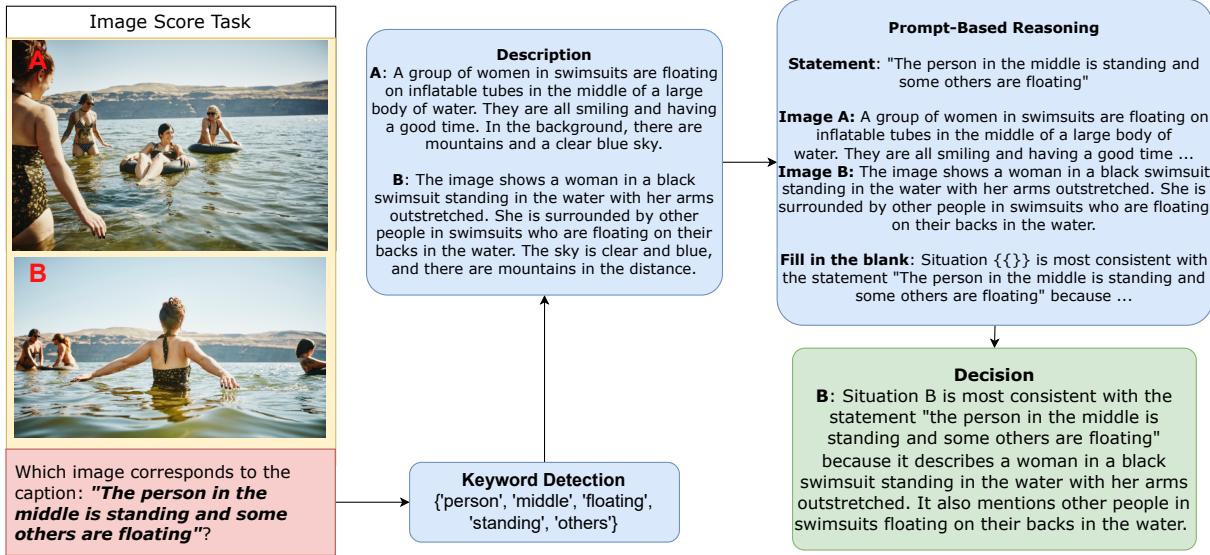


Figure 13: Visually Difficult Question for the image score task.

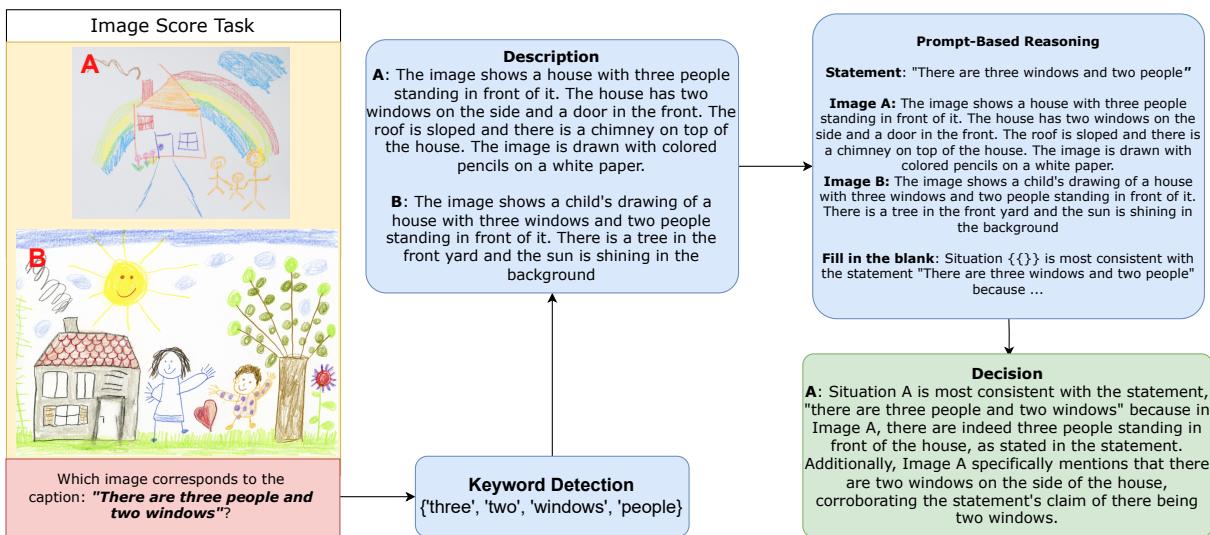


Figure 14: Unusual Image Question. Although both images are cartoons, both descriptions match the contents of the images accurately.

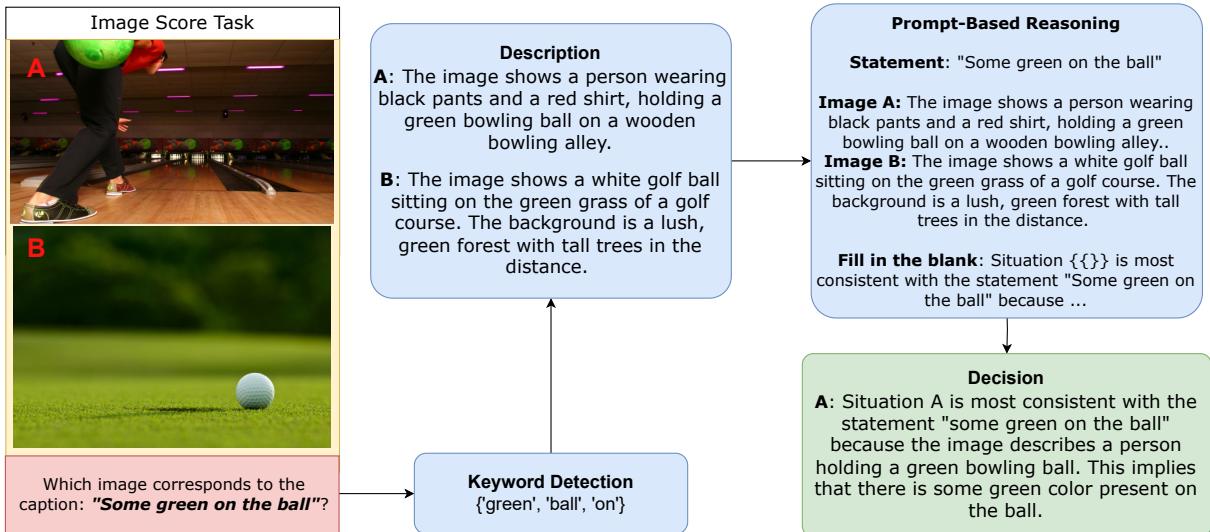


Figure 15: Unusual Text question for the image score task.

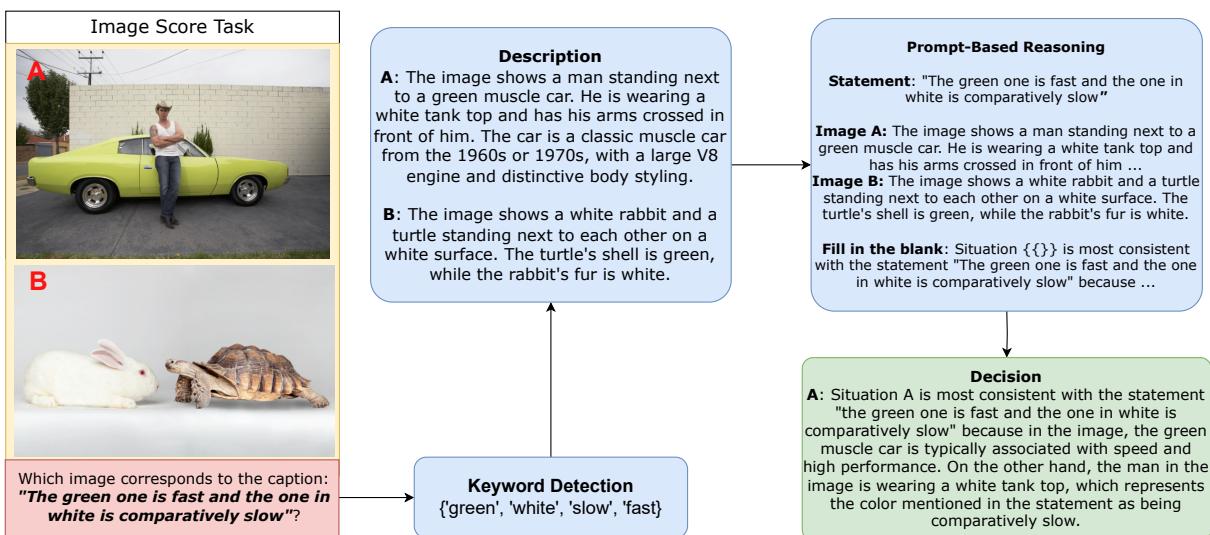


Figure 16: Complex Reasoning question for the image score task.

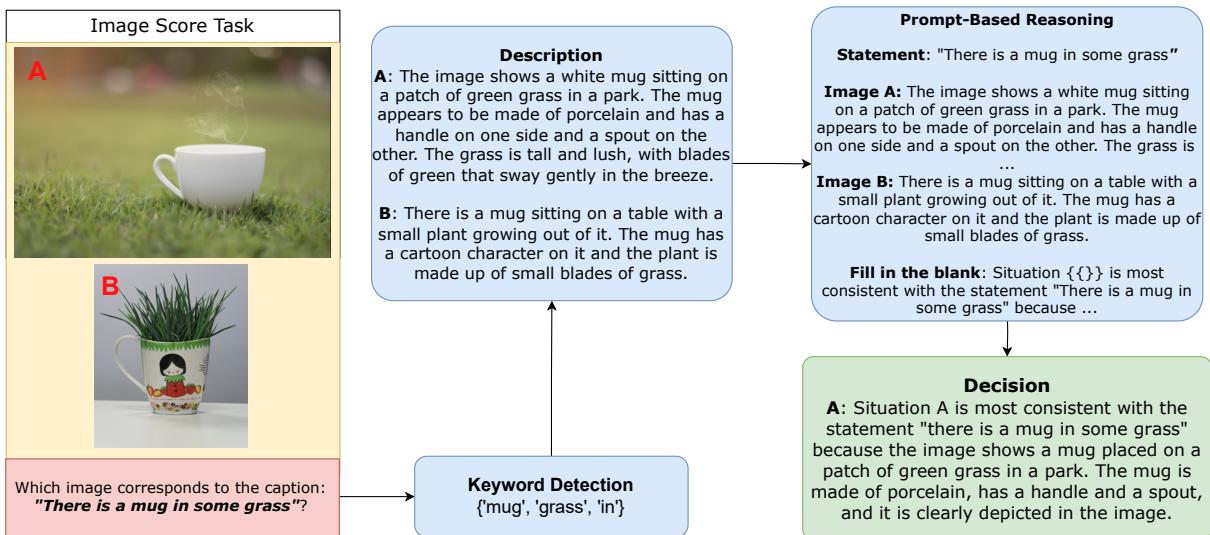


Figure 17: NoTag question for the image score task.

E Fine Grained Image Scores

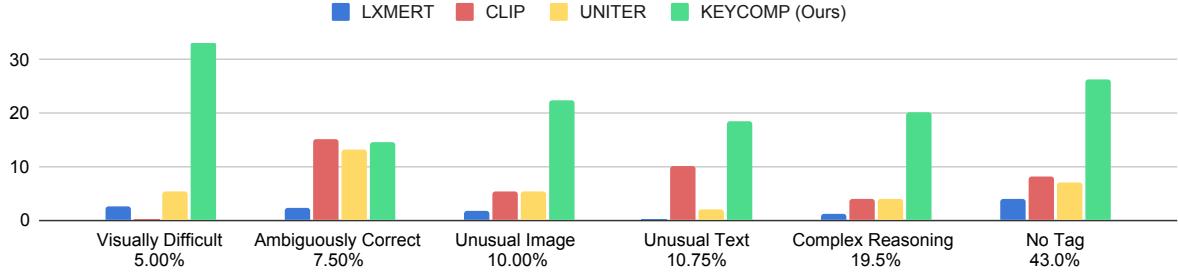


Figure 18: Fine-grained image score performance across different question categories. We give specific examples from each category in Appendix C. Percentages on the x-axis indicate each question type’s proportion of the dataset. To ensure representative results, question categories comprising less than 5% of the dataset are excluded.

F Effect of VLM Size and LLM Size

VLM	Text	Image	Group
MiniGPT-4 7b	23.3	25.5	0.1
MiniGPT-4 13b	30.3	24.6	12.4

Table 5: KEYCOMP performance while changing the size of the VLM image captioner.

LLM	Text	Image	Group
GPT-3.5	30.3	24.6	12.4
GPT-4	44.0	25.9	15.4

Table 6: KEYCOMP performance while changing the size of the LLM used for reasoning.

Tables 5 and 6 show the effect of changing the VLM and LLM sizes respectively. Our results suggest that changing the LLM size improves text score significantly and offers marginal improvement in image score. This intuitively makes sense because text score requires a deeper understanding of text to distinguish between similar captions. Upgrading the VLM also provides improvement to text score, but does not impact the image score.

G Error Analysis

In this section, we detail what examples KEYCOMP struggles with and suggest future directions for generative approaches. We categorize common errors as either VLM-based, LLM-based or both.

G.1 Spatial Reasoning (VLM)

For questions requiring spatial reasoning, we note that KEYCOMP sometimes produces inaccurate descriptions of the scene. We believe this stems from the image captioning model (VLM) and may be addressed by utilizing scene understanding models trained with object-relation level supervision.

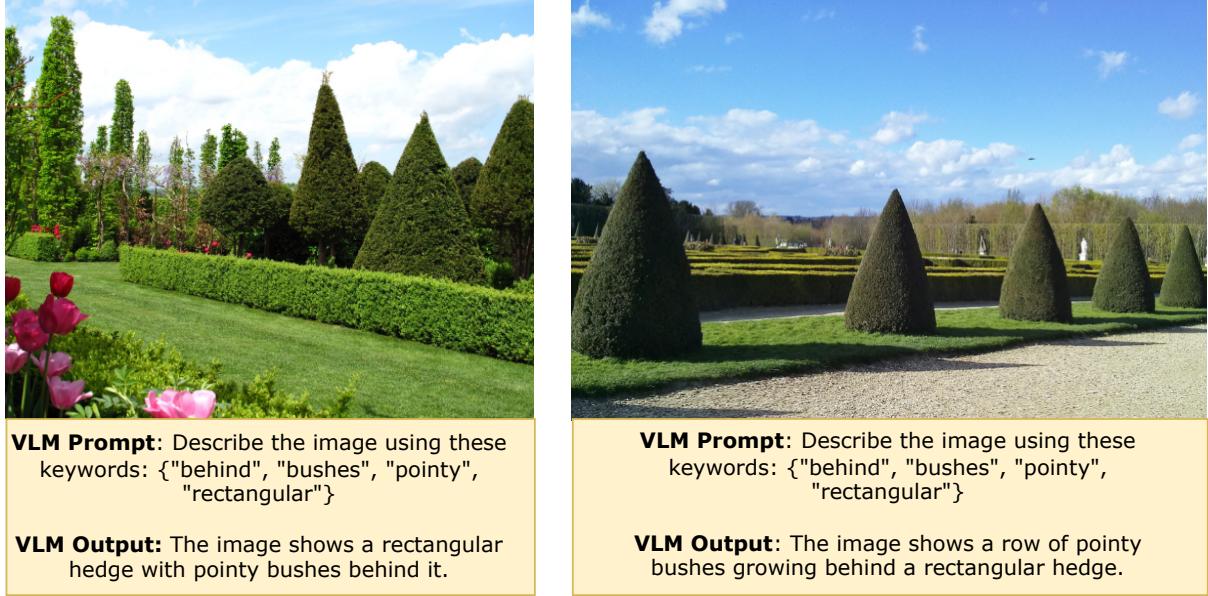


Figure 19: Spatial Reasoning Error. As the VLM does not explicitly model the spatial information of objects in an image, the VLM mistakenly thinks the bushes are behind the hedges in the right image. This is particularly difficult when two objects have similar colors.

G.2 Incorrect LLM Reasoning

Occasionally the LLM is presented with VLM descriptions that should contain enough information to answer the question, but still selects the wrong answer. This often occurs when the VLM description has a complicated syntax structure that involves multiple objects.

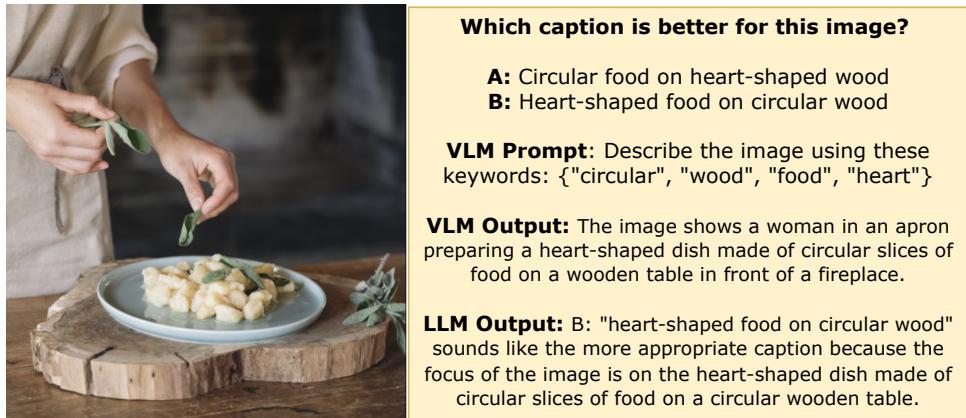


Figure 20: LLM Interpretation Error. Although VLMs provide a sufficiently detailed description, the description has a complicated syntax structure that involves multiple objects. The LLM response still does not match the VLM description.

G.3 Describing Out of Focus or Occluded Objects (VLM)

Sometimes the VLM produces inaccurate descriptions of scene elements which are out of focus or missing parts of objects. In these cases, even though the prompt requires the LLM to choose an option, the LLM may sometimes refuse to choose, saying it is uncertain and offers a better caption. This makes it hard to automate the evaluation and may be addressed with further fine-tuning in this instruction format or additional prompt engineering.



Which caption is better for this image?

A: The person closer to the camera weightlifts and the person farther from the camera runs

B: The person closer to the camera runs and the person farther from the camera weightlifts

VLM Prompt: Describe the image using these keywords: {"person", "farther", "camera", "weightlift" "run"}

VLM Output: The image shows a man and a woman running on a treadmill in a gym. The man is wearing a blue shirt and black shorts, while the woman is wearing a black shirt and blue shorts. They are both wearing athletic shoes and headphones.

LLM Output: Neither A nor B would be an accurate caption for the described image. Both individuals described are running on treadmills, not weightlifting. A more accurate caption could be, "A man and woman exercising on treadmills in a busy gym".

Figure 21: LLM does not give an answer. In this case, the VLM description also does not describe the scene accurately, likely because the man is out of focus and his head is not visible.