COGNITIVE RESILIENCE: UNRAVELING THE PROFICIENCY OF IMAGE-CAPTIONING MODELS TO INTERPRET MASKED VISUAL CONTENT

Zhicheng Du[1], **Zhaotian Xie**[1], **Huazhang Ying**[1], **Likun Zhang**[1], **Peiwu Qin**[1,*]
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
{duzc21, xzt21, yhz22, zlk21}@mails.tsinghua.edu.cn
pwqin@sz.tsinghua.edu.cn

ABSTRACT

This study explores the ability of Image Captioning (IC) models to decode masked visual content sourced from diverse datasets. Our findings reveal the IC model's capability to generate captions from masked images, closely resembling the original content. Notably, even in the presence of masks, the model adeptly crafts descriptive textual information that goes beyond what is observable in the original image-generated captions. While the decoding performance of the IC model experiences a decline with an increase in the masked region's area, the model still performs well when important regions of the image are not masked at high coverage.

1 INTRODUCTION

Image Captioning (IC), also known as the Image-to-Text task, is a pivotal multimodal research area whose goal is to generate natural language descriptions of visual content (Sharma & Padha, 2023; Ding, 2023). The IC model aims to furnish a holistic comprehension of the image, capturing object relationships, attributes, scene characteristics, and interactions among objects within the given scene (Ghandi et al., 2023). Recent years have borne witness to a pronounced proliferation of scholarly pursuits in the sphere of IC, culminating in a manifold of applications spanning diverse domains, including biomedical IC (Zheng & Yu, 2023) and bespoke assistive technology catering to the visually impaired (Guo et al., 2022). Concomitantly, the demonstrated efficacy of the masked autoencoder (MAE) (He et al., 2022) highlights the potential latent in visual self-supervised learning (SSL), which has attracted considerable attention in particular for its aptitude in predicting masked input attributes predicated upon unmasked input content (Zhang et al., 2022). Inspired by the success of MAE, we raise an incisive inquiry: Can IC models directly discern the masked visual content and yield accurate textual descriptions? While existing IC models have demonstrated proficiency in generating precise captions for clear images, limited insights exist into their capacity to comprehend masked visual content. Herein, we design several masking methodologies applied to the image, aiming to scrutinize the impact of diverse masking strategies on the captions generated by the model.

2 METHOD

The overall pipeline of this work is shown in Appendix A. We assess four IC models, including the Large Language and Vision Assistant (LLaVA) (Liu et al., 2023), ViT-GPT2-Image-Captioning model (Kumar, 2022), Generative Image-to-Text Transformer (GIT) (Wang et al., 2022), and Bootstrapping Language-Image Pre-Training (BLIP) (Li et al., 2022). We design three distinct masking methods (i.e., masked ratio, masked block size, and color) for image masking. The masked ratio represents the proportion of the covered image area to the total image size; masked block size denotes the area of the square used for random masking; and color means the hue of the masked block. A random selection of masked areas combines these ways to mask the image. We conduct both quantitative and qualitative analyses to scrutinize the disparities in textual descriptions generated from

[*]Corresponding author

the original and masked images. The pre-trained General Text Embeddings (GTE) model (Li et al., 2023) is used to compute semantic textual similarity scores as a criterion for quantitative evaluation.

3 Experiments

Dataset and Experimental setting. We construct a dataset comprising 60 images to deploy the performance evaluation of the IC model. To minimize the impact of chance and ensure a diverse representation, we randomly select ten images from each of the six distinct source datasets (i.e., the Pokemon[1], COCO2017 (Lin et al., 2014), LAION-COCO (Schuhmann et al., 2022), Impressions (Kruk et al., 2023), Midjourney-threads(Don-Yehiya et al., 2023), and Flickr8k (Hodosh et al., 2013) datasets). The distribution of these images is provided in Appendix B. The masked ratio incrementally ascends in 5% intervals, while the square masked block's side length takes on values of 4, 8, 16, 64, 128, and 256. The color of the masked block includes black, white, and grey.

Results. Figure 1 illustrates the quantitative and qualitative outcomes of the LLaVA. The upper and lower sections of the image showcase examples of qualitative analysis, depicting high and low semantic similarity between the captions of the masked image and the subtitles of the original image, respectively. In the middle, we observe the quantitative change in semantic similarity as the masked ratio increases. Our findings reveal a non-linear correlation between the captions generated by the IC model and the original image as the random masking rate increases. Surprisingly, the accuracy of captions corresponding to images with a low masking rate is higher than that of those with a high masking rate. Our study elucidates that the model exhibits heightened accuracy in caption generation when pivotal and contextually relevant information within the image remains unmasked, contingent upon the positioning of the masking block. Notably, larger masking blocks in most instances increase the likelihood of concealing pertinent information within the image region, leading the model to produce inaccurate descriptions even when the masking rate is low. Interestingly, we observe that the color of the masked block can erroneously influence the IC model to output information not present in the image (see Appendix C). Moreover, the masking process induces the IC model to output information absent in the original caption (see Appendix D). These insights highlight the nuanced impact of masking strategies on the IC model's output.

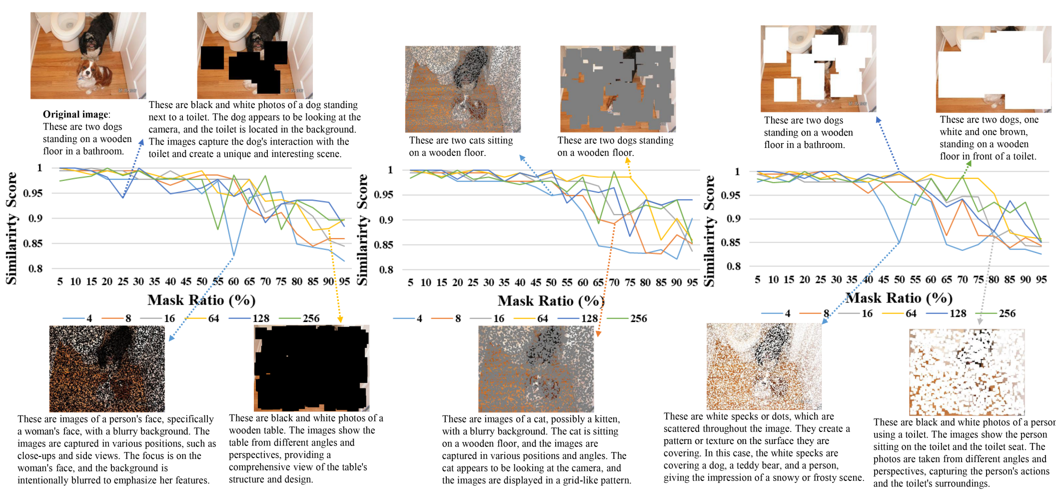

Figure 1: The quantitative and qualitative outcome results of LLaVA. Three line charts represent the colors of masked block, where the left one is black, middle one is gray, and the right one is white.

4 Discussion and Conclusion

Our study comprehensively analyzes the ability of IC models to understand masked visual content across various conditions and degrees. The relevance of three distinct approaches to mask processing and the subsequent elaboration of textual descriptions generated by IC models is highlighted.

[1]https://huggingface.co/datasets/diffusers/pokemon-gpt4-captions

In future research endeavors, we plan to delve deeper into mining the relationships and importance ranking between different regions within an image and exploring innovative image masking processing methods to further advance the development of visual SSL and multimodal models.

URM Statement

The authors acknowledge that all key authors of this work meet the URM criteria of ICLR 2024 Tiny Papers Track.

Reproducibility Statement

The source code, dataset used in this study, and experimental results will be available at https://github.com/dodoxxb/cognitive-resilience.

Acknowledgements

This research is supported by the National Natural Science Foundation of China 31970752,32350410397;Science,Technology,Innovation Commission of Shenzhen Municipality,JCYJ20220530143014032,JCYJ20230807113017035,WDZC20200820173710001,Shenzhen Medical Research Funds,D2301002;Department of Chemical Engineering-iBHE special cooperation joint fund project, DCE-iBHE-2022-3; Tsinghua Shenzhen Interna-tional Graduate School Cross-disciplinary Research and Innovation Fund Research Plan, JC2022009; and Bureau of Planning, Land and Resources of Shenzhen Municipality (2022) 207.

References

Yi Ding. A systematic literature review on image captioning. In Constantine Stephanidis, Margherita Antona, Stavroula Ntoa, and Gavriel Salvendy (eds.), *HCI International 2023 Posters - 25th International Conference on Human-Computer Interaction, HCII 2023, Copenhagen, Denmark, July 23-28, 2023, Proceedings, Part V*, volume 1836 of *Communications in Computer and Information Science*, pp. 396–404. Springer, 2023. doi: 10.1007/978-3-031-36004-6_54. URL https://doi.org/10.1007/978-3-031-36004-6_54.

Shachar Don-Yehiya, Leshem Choshen, and Omri Abend. Human learning by model feedback: The dynamics of iterative prompting with midjourney. *CoRR*, abs/2311.12131, 2023. doi: 10.48550/ARXIV.2311.12131. URL https://doi.org/10.48550/arXiv.2311.12131.

Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar. Deep learning approaches on image captioning: A review. *ACM Comput. Surv.*, 56(3), oct 2023. ISSN 0360-0300. doi: 10.1145/3617592. URL https://doi.org/10.1145/3617592.

Yu Guo, Yue Chen, Yuanyan Xie, Xiaojuan Ban, and Mohammad S. Obaidat. An offline assistance tool for visually impaired people based on image captioning. In Donald A. Adjeroh, Qi Long, Xinghua Mindy Shi, Fei Guo, Xiaohua Hu, Srinivas Aluru, Giri Narasimhan, Jianxin Wang, Mingon Kang, Ananda Mondal, and Jin Liu (eds.), *IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2022, Las Vegas, NV, USA, December 6-8, 2022*, pp. 969–976. IEEE, 2022. doi: 10.1109/BIBM55620.2022.9994947. URL https://doi.org/10.1109/BIBM55620.2022.9994947.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 15979–15988. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01553. URL https://doi.org/10.1109/CVPR52688.2022.01553.

Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Int. Res.*, 47(1):853–899, may 2013. ISSN 1076-9757.

Julia Kruk, Caleb Ziems, and Diyi Yang. Impressions: Understanding visual semiotics and aesthetic impact, 2023.

Ankur Kumar. The illustrated image captioning using transformers. *ankur3107.github.io*, 2022. URL https://ankur3107.github.io/blogs/the-illustrated-image-captioning-using-transformers/.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 2022. URL https://proceedings.mlr.press/v162/li22n.html.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning, 2023.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/a1859debfb3b59d094f3504d5ebb6c25-Abstract-Datasets_and_Benchmarks.html.

Himanshu Sharma and Devanand Padha. A comprehensive survey on image captioning: from handcrafted to deep learning-based techniques, a taxonomy and open research issues. *Artif. Intell. Rev.*, 56(11):13619–13661, 2023. doi: 10.1007/S10462-023-10488-2. URL https://doi.org/10.1007/s10462-023-10488-2.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. GIT: A generative image-to-text transformer for vision and language. *Trans. Mach. Learn. Res.*, 2022, 2022. URL https://openreview.net/forum?id=b4tMhpN0JC.

Chaoning Zhang, Chenshuang Zhang, Junha Song, John Seon Keun Yi, Kang Zhang, and In So Kweon. A survey on masked autoencoder for self-supervised learning in vision and beyond. *CoRR*, abs/2208.00173, 2022. doi: 10.48550/ARXIV.2208.00173. URL https://doi.org/10.48550/arXiv.2208.00173.

Ervine Zheng and Qi Yu. Evidential interactive learning for medical image captioning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42478–42491. PMLR, 2023. URL https://proceedings.mlr.press/v202/zheng23g.html.

A THE OVERALL PIPELINE OF THIS STUDY

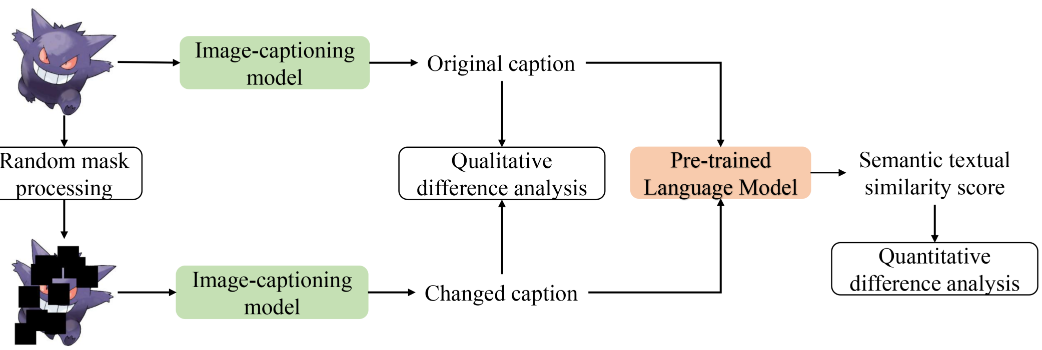

Figure 2: Overall pipeline of this work, where the random mask processing includes three ways (i.e., masked ratio, masked block size, and color), and the pre-trained GTE model are used to deploy the experiment of semantic textual similarity.

B THE T-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING (T-SNE) VISUALIZATION OF THE DATASET

We employ the ResNet-50 model (He et al., 2016) to extract features from the dataset used in this study. Subsequently, we utilize the t-SNE algorithm (van der Maaten & Hinton, 2008) to visualize the distribution of the dataset.

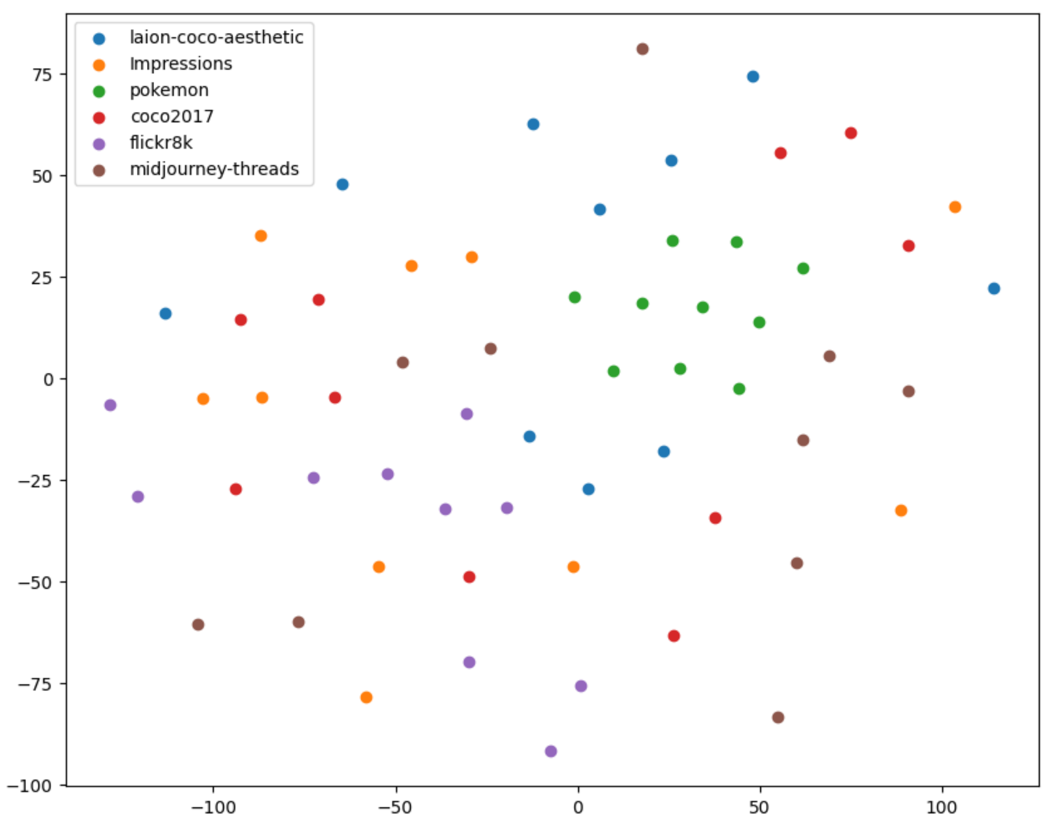

Figure 3: The distribution of the dataset used in this work

C THE EFFECT OF THE MASKED BLOCK'S COLOR

The qualitative analysis demonstrates that incorporating the masked operation with diverse colors has a noticeable impact on the IC model's output, resulting in more and varied information under equivalent conditions. The strategic use of different colors proves advantageous, enriching the model's ability to provide a more detailed image description. Even in instances where the IC model encounters difficulties in accurately portraying the original image, the meticulous application of the masking process can play a crucial role in addressing these challenges and facilitating improvements.

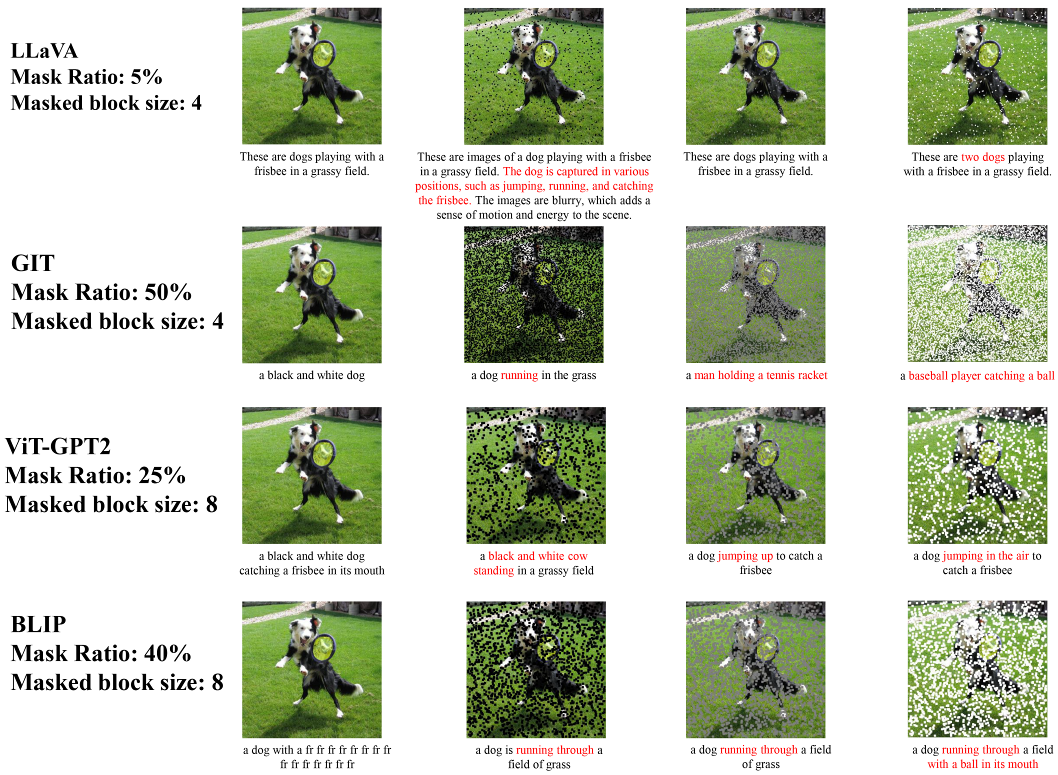

Figure 4: The results of the interpretation of images processed with different mask block colors, where the text in red means the differences with the text from the original image. The first column is the original image, and the second through fourth columns are the images processed with black, gray and white masks respectively (image example from the flickr8k dataset and the used IC model is LLaVA).

D THE SUPPLEMENTARY EFFECT OF THE MASK PROCESSING FOR THE IC MODEL

Through qualitative analysis, we observe that a thoughtfully designed masked operation enhances the ability of the IC model to capture additional hidden information in images. It's worth noting that this newfound information may occasionally be incorrect, representing the model's speculative understanding of the image. Nevertheless, the positive aspect is that the masked operation enables the IC model to delve into more nuanced details of the image, showcasing an improved understanding.

Figure 5: The results of showing mask processing help the IC model output more informative results, where the text in red means it supplements the text from the original image (image example from the flickr8k dataset and the used IC model is LLaVA).