HiFi-Score: Fine-grained Image Description Evaluation with Hierarchical Parsing Graphs Supplementary Material

Ziwei Yao[1;2], Ruiping Wang[1;2], and Xilin Chen[1;2]

[1] Key Laboratory of AI Safety of CAS, Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing 100190, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
ziwei.yao@vipl.ict.ac.cn , {wangruiping,xlchen}@ict.ac.cn

In the supplementary material, we introduce in detail caption-level datasets and the proposed ParaEval dataset, prompt strategies for the text HPG generation and uency evaluation, extended quantitative experiments, evaluation cost, and visualization of image HPG examples.

1 Caption-level Datasets

We conduct quantitative experiments on four caption-level benchmarks (Flickr8k-Expert [4], Composite Dataset [1], THumB1.0 [6], and Pascal-50S dataset [14]) to compare our metric with the state-of-the-art metrics, as in Sec. 4.2 in the main paper. Below is the detailed introduction to these datasets.

Flickr8k-Expert [4] consists of 5,644 image-caption pairs, with 5 references for each image. Each candidate caption is annotated with 3 scores, scoring from 1 (unrelated to the image) to 4 (describing the image correctly).

Composite Dataset [1] contains 3,995 test images from MS COCO [10], Flickr30k [16], and Flickr8k [4]. Each image has one human-written caption and two machine-generated captions. All candidate captions are scored from 1 (not related to the image) to 5 (perfectly related to the image).

THumB1.0 [6] samples 500 images from MS COCO [10] and provides human assessments for ve corresponding captions, including one written by human and four generated by captioning models. Annotators evaluate each image-caption pair with precision to measure the correctness of caption, recall to assess the coverage of salient information, and a total score combining precision, recall as well as uency, conciseness and inclusive language.

Pascal-50S dataset [14] comprises 1,000 images from the UIUC PASCAL Sentence Dataset [13], along with 50 references for each image. The dataset includes 4,000 human-assesed caption pairs, forming 4 groups with 1,000 pairs in each: HC (correct human-written pairs), HI (correct and incorrect human-written pairs), HM (human-written and machine-generated pairs) and MM (machine-generated pairs).

Following previous works, Kendall's correlation on Flickr8k-Expert and Composite Dataset, Pearson's correlation on THumB1.0 and classi cation accuracy on Pascal-50S are used to measure the consistency between metrics and

Table 2: Comparison of HiFi-Score and CLIPScore with larger parameters.

Metric	Params	F-Ex	Com.	Pas.	Thu.
Official ClipScore (ViT-B/32)	151M	51.4	53.8	80.9	31.9
ClipScore (ViT-L/14)	427M	53.0	55.4	81.5	36.3
ClipScore (ViT-G/14)	1.37B	54.5	56.7	81.6	37.0
HiFi-Score (Ours)	1.60B	**58.4**	**65.8**	**83.0**	**44.8**

5 Evaluation Cost

Here we summarize the computational and time costs of HiFi-Score. The parameters for each module are as follows: SAM-641M, OneFormer-223M, GLIP-429M, BLIP-2 (stage1)-1.17B. Time cost for evaluating a single image-text pair: Image HPG-1.9s; Text HPG 1.5s; scoring-0.7s. The bottleneck lies in off-the-shelf HPG generation, which can be optimized by pre-extracting for common benchmarks. GPT-3.5 pricing is about $0.7 for 1000 candidates, much cheaper than human.

We prioritize the frameworks' completeness and performance rather than the cost when designing HiFi-Score, but we believe that with the development of lightweight open-source models (such as MobileSam, Gemini-Pro), the cost will significantly decrease in the near future.

6 Image HPG examples

In Fig. 6, we present more visual examples of image HPGs, which are generated by SAM [7] alone or by the combination of SAM and OneFormer [5]. As mentioned in Sec. 3.1 of the main paper, SAM is a non-semantic segmentation method, thus it may break down complete objects or large background areas into multiple segments. Additionally, it is sometimes overly sensitive to the details of masks, resulting in fragmented and ineffective HPGs. On the other hand, the panoptic segmentation method like OneFormer simultaneously segments foreground things and background stuff, allowing control over the granularity of hierarchy based on categories. For instance, backgrounds like sky and lawns do not need further subdivision, while important instances like humans and animals can retain more hierarchical levels. By combining SAM and OneFormer, image HPGs become more accurate, clearer, and more focused noticeably.

References

1. Aditya, S., Yang, Y., Baral, C., Aloimonos, Y., Fermüller, C.: Image understanding using vision and reasoning through scene description graph. Computer Vision and Image Understanding **173**, 33–45 (2018)
2. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. In: Proceedings of the European conference on computer vision (ECCV). pp. 382–398 (2016)

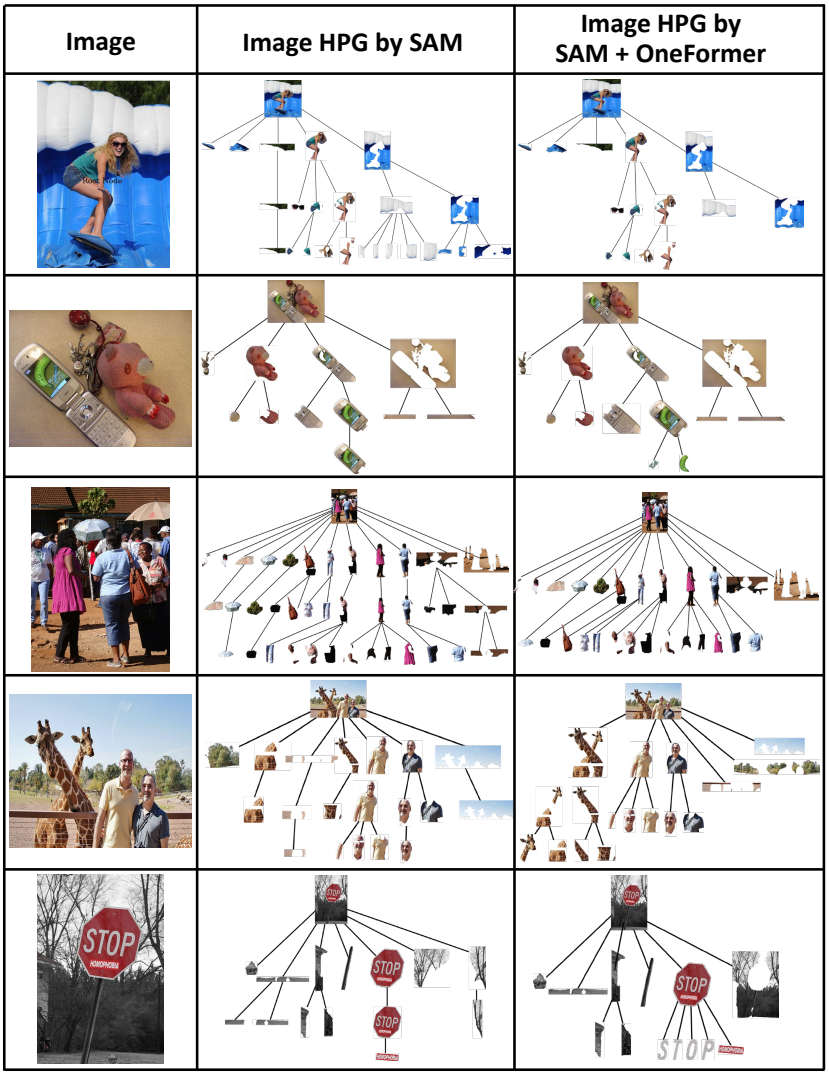

Fig. 6: Examples of image HPGs.

3. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: CLIPScore: A reference-free evaluation metric for image captioning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 7514–7528 (2021)
4. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research **47**, 853–899 (2013)
5. Jain, J., Li, J., Chiu, M.T., Hassani, A., Orlov, N., Shi, H.: Oneformer: One transformer to rule universal image segmentation. In: Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2989–2998 (2023)
6. Kasai, J., Sakaguchi, K., Dunagan, L., Morrison, J., Bras, R.L., Choi, Y., Smith, N.A.: Transparent human evaluation for image captioning. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). pp. 3464–3478 (2022)
7. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
8. Krause, J., Johnson, J., Krishna, R., Fei-Fei, L.: A hierarchical approach for generating descriptive image paragraphs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 317–325 (2017)
9. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision **128**(7), 1956–1981 (2020)
10. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proceedings of the European conference on computer vision (ECCV). pp. 740–755 (2014)
11. Pont-Tuset, J., Uijlings, J., Changpinyo, S., Soricut, R., Ferrari, V.: Connecting vision and language with localized narratives. In: Proceedings of the European conference on computer vision (ECCV). pp. 647–664 (2020)
12. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning. pp. 8748–8763 (2021)
13. Rashtchian, C., Young, P., Hodosh, M., Hockenmaier, J.: Collecting image annotations using amazon's mechanical turk. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. pp. 139–147 (2010)
14. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4566–4575 (2015)
15. Wu, H., Mao, J., Zhang, Y., Jiang, Y., Li, L., Sun, W., Ma, W.Y.: Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6609–6618 (2019)
16. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics **2**, 67–78 (2014)