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 fluency 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 five 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 fluency, 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 τ correlation on Flickr8k-Expert and Composite Dataset, Pearson's ρ correlation on THumB1.0 and classification accuracy on Pascal-50S are used to measure the consistency between metrics and

human evaluations, respectively. Please refer to the main paper for the experimental results and analysis.

2 Paragraph-level Dataset ParaEval

In this section, we introduce the proposed paragraph-level dataset ParaEval and additional experiments on it.

2.1 Dataset Construction.

ParaEval dataset is proposed to assess the accuracy of existing metrics in evaluating long-context image descriptions. ParaEval collects more than 4,000 pairs of image and its paragraph description from the ImageParagraphs Dataset [8] and Localized Narratives Dataset [11]. The images source from Visual Genome [11] and OpenImages [9], respectively. We design four types of negative samples: plausible descriptions, descriptions with negative objects, negative attributes, and negative relationships.

Plausible descriptions are hard negative candidates which describe images very similar to the groundtruth image. We utilize CLIP [12] to calculate the similarity between the groundtruth image and other images, and the one with the highest similarity score is selected as a hard negative image and its description is chosen as the plausible description. On the ImageParagraphs Dataset, with 2,489 images in its test set, we obtain the hardest negative description for each image. On the test set of Localized Narratives Dataset, we first filter out text descriptions that fewer than 40 words, constructing a pool of over 5000 images. From the pool, we select hard negative samples with image similarity scores greater than 0.82 and finally generate 2,358 plausible samples. Since the candidate images of localized narratives are more than that of image paragraphs, the task on the localized narratives set is harder.

To construct negative samples containing negative objects, attributes and relationships, we replace the corresponding parts in the groundtruth, with the help of GPT-3.5. For negative objects, we generate multiple candidate negative descriptions by replacing different objects in a groundtruth description, and then require the LLM to identify the most reasonable one. To generate negative relationship samples, we ask the LLM to generate new relationship phrases that are directionally opposite or semantically different, and then integrate them back into the original sentence. For negative attributes, first existing attributes are categorized into several classes such as color, size, quantity, etc., and then negative attribute phrases are generated based on the attribute and object category, and finally integrated back to obtain descriptions with negative attributes.

2.2 Additional Experiments

Impact of amount of text replacement. Fig. 1 shows the accuracy of metrics as the number of replaced words in the negative samples increases. As the

difference between negative samples and the groundtruth one increases, it becomes easier to distinguish. Therefore, the overall accuracy is improving as more words are being replaced. The overall accuracy on negative relationships is significantly lower than the other two, indicating that metrics are less sensitive to relationships compared to objects and attributes. The proposed HiFi-Score performs relatively better.

Accuracy on different attribute categories. We classify the replaced attributes into 9 categories: color, size, etc., and then compare the accuracy of metrics on samples containing the corresponding category of attributes. As shown in Fig. 1(d), HiFi-Score achieves the highest accuracy in almost all categories, and comparable performance to BLIPScore in the spatial category. Overall, among these categories, the performance on action, material, and color is relatively higher, while spatial and environment categories are harder to distinguish. This reflects the varying sensitivity of current vision-language models to different attribute categories.

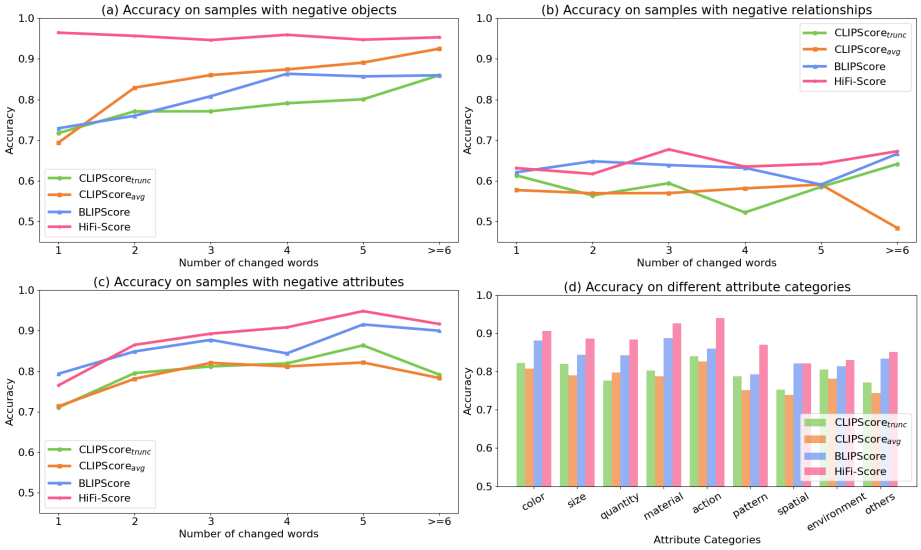

Fig. 1: (a-c) Accuracy as the number of changed words increases on samples with negative objects, relationships, and attributes. (d) Accuracy comparison on different attribute categories.

3 LLM Prompts

In this section, we discuss the prompting strategies used for Text HPG Generation (Sec. 3.1 in the main paper) and Fluency Evaluation (Sec. 3.3 in the main paper).

3.1 Text HPG Generation

Given an image description of arbitrary length, we first parse objects, attributes corresponding to each object, and relationships between objects to phrases, and then convert them into a hierarchical parsing graph. The text parsing process can be implemented using GPT-3.5 based on the prompt in Fig. 2. Compared to previous rule-based grammar analysis([2], [15]), LLMs have stronger capabilities in understanding complex sentences and referent disambiguation, laying a solid foundation for fine-grained evaluation of long image descriptions.

Given an image description, please help me parse and extract all objects and assign a number to each object. For each object, find all attributes and relations related to it, and summarize each as a phrase. Each phrase should contain the object. If there are multiple objects of the same kind, assign a number to each of them.

Description: A group of people are walking on the beach. One of them is a girl carrying a pink and white umbrella with a black logo waving her hands. Another girl is looking at her.
Output:
Objects:
1. Object: people_1
Phrases: a group of people_1, people_1 walking on beach_2
2. Object: beach_2
Phrases: beach_2
.....

Fig. 2: Prompt to generate textual hierarchical parsing graphs

3.2 Fluency Evaluation

We empirically decouple fluency into three dimensions: grammatical correctness, clarity of expression, and logical rationality. We utilize GPT-3.5 to render binary assessments across the mentioned criteria, provide specific reasons, and give the possible revised text. The prompt and evaluation examples are displayed in Fig. 3 and Fig. 4, respectively.

In the main paper, we consider fluency as an additional dimension independent from fidelity and adequacy. This is because, unlike the latter two which are

Evaluate the acceptability of the following text. Provide your assessment and reasons based on specific criteria, and correct it.
"Description to evaluate."

Assess the text using the following criteria:
1. Grammatical Correctness: Does the text adhere to the rules of grammar? Only consider grammar issues such as agreement errors and article errors.
2. Clarity: Is the text clear and comprehensible?
3. Logical Coherence: Is the information and argumentation within the text logically consistent?

Please format your assessment of the text's acceptability as follows:
Assessment:
- Grammatical Correctness: [Yes or no]
- Clarity: [Yes or no]
- Logical Coherence: [Yes or no]

Summary: [Overall acceptability of the text and areas for improvement]

Corrected Text: [1-2 corrected text or "No need to correct"]

Fig. 3: Prompt to evaluate fluency of candidate description

Fig. 4: Examples of fluency evaluation.

closely related to images, fluency is a pure linguistic dimension. Furthermore, fidelity and adequacy are continuous scores, while fluency consists of three binary scores. Focusing on obtaining fine-grained multi-dimensional evaluations rather than merely overall scores, we do not forcefully integrate fluency to the

other two. If a total score is needed, we recognize fluency as a factor to the sum of fidelity and adequacy. For every grammar, clarity, or logic error, the fluency factor decreases by 0.05. On 500 random samples from Composite, Kendall's τ: **Fid+Ade-64.78**, **Flu*(Fid+Ade)-64.82**, demonstrating that incorporating fluency maintains a high human consistency.

4 Extended quantitative experiments

In this section, we conduct extended experiments to explore the impact of various human attention sources for adequacy evaluation, the effect of different weight values, and compare HiFi-Score to baseline models that use only entity-matching and other metrics that employ larger backbones.

Impact of human attention sources. We validate the impact of human attention from different sources through experiments. As shown in Table 1, HiFi-Score$_{none}$ without the introduction of human attention has lower consistency with human, proving the necessity of incorporating human priors. Furthermore, the performance of HiFi-Score$_{ref}$ and HiFi-Score is similar, indicating that machine-generated saliency maps can to some extent replace human-written references as a basis for evaluation. This can effectively reduce the dependency on annotated references.

Table 1: Impact of different kinds of human attention. HiFi-Score$_{none}$ does not consider human attention and set all attention weights as 1. HiFi-Score$_{ref}$ assigns weights to instances based on the frequency of that region mentioned in the reference. HiFi-Score is the default full model that uses saliency map as human attention.

		F-Ex	Com	THumB1.0		
Model	Attn.	τ_c	τ_c	P	R	Total
HiFi-Score$_{none}$	None	56.7	61.3	0.40	0.18	0.40
HiFi-Score$_{ref}$	reference-based	58.3	65.6	0.42	0.22	0.45
HiFi-Score	saliency-based	**58.4**	**65.7**	**0.43**	**0.22**	**0.45**

Impact of weight values. Fig. 5 shows the trade-off of the weight values γ, δ and α.

(1) Fidelity v.s. Adequacy (γ). Table 2 in the main paper indicates Fidelity's impact is slightly higher than Adequacy's, as description's correctness is more fundamental, and combining both enhances performance. The same can be observed in Fig. 5. As **purple lines**, Fidelity's weight should be slightly higher for optimal performance.

(2) ITM v.s. ITC (δ). As in Table 2 in the main paper and **green lines** in Fig. 5, Flickr8K-Exp and Pascal-50S prefer ITC, while Composite and

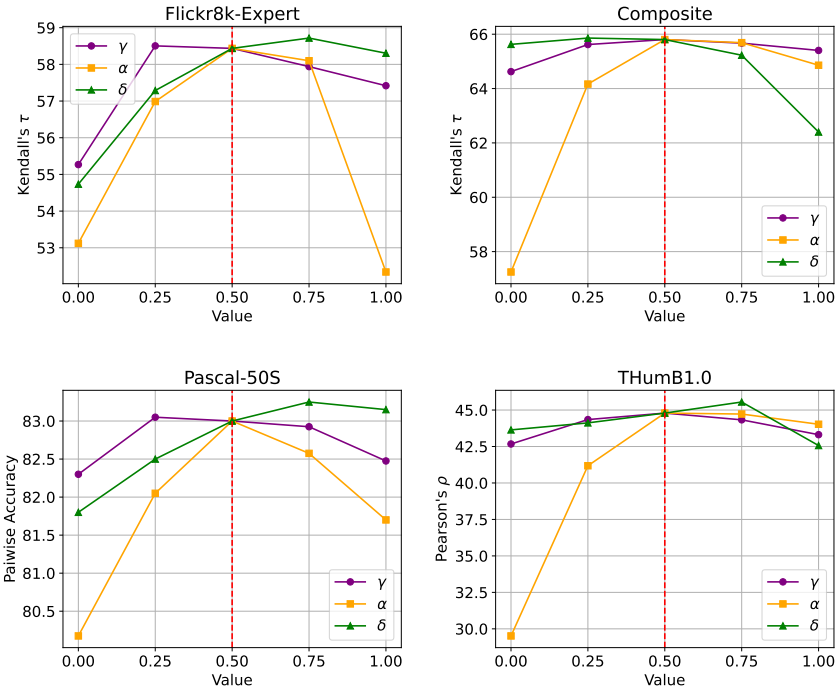

Fig. 5: Impact of different weight values on the metric performance.

THumB1.0 prefer ITM. Combining the two in a certain proportion can enhance performance. $\delta = 0.5$ is selected as a balanced trade-off.

(3) Impact of α (Eq. 5/6). As yellow lines in Fig. 5, $\alpha = 0.5$ performs best, as a balance between parent and child nodes.

Comparison to entity-matching. HPG-based matching emphasizes relationships between regions and layers. To verify the role of structuring, we remove the hierarchy from HPGs and simply calculate the average similarities of the matched parts across all regions. Results are **Flickr8k-Expert - 55.2; Composite - 56.5; Pascal-50S - 80.9; THumB1.0 - 33.7**, clearly inferior to HiFi-Score, highlighting structured modeling's benefits.

Comparison to the metric with larger parameters. We additionally test the CLIPScore [3] with larger ViT backbones. As in the Table 2, the performance of CLIPScore improves as the capacity of the backbone increases. HiFi-Score still outperforms CLIPScore (ViT/G-14), which has comparable parameter number.

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)

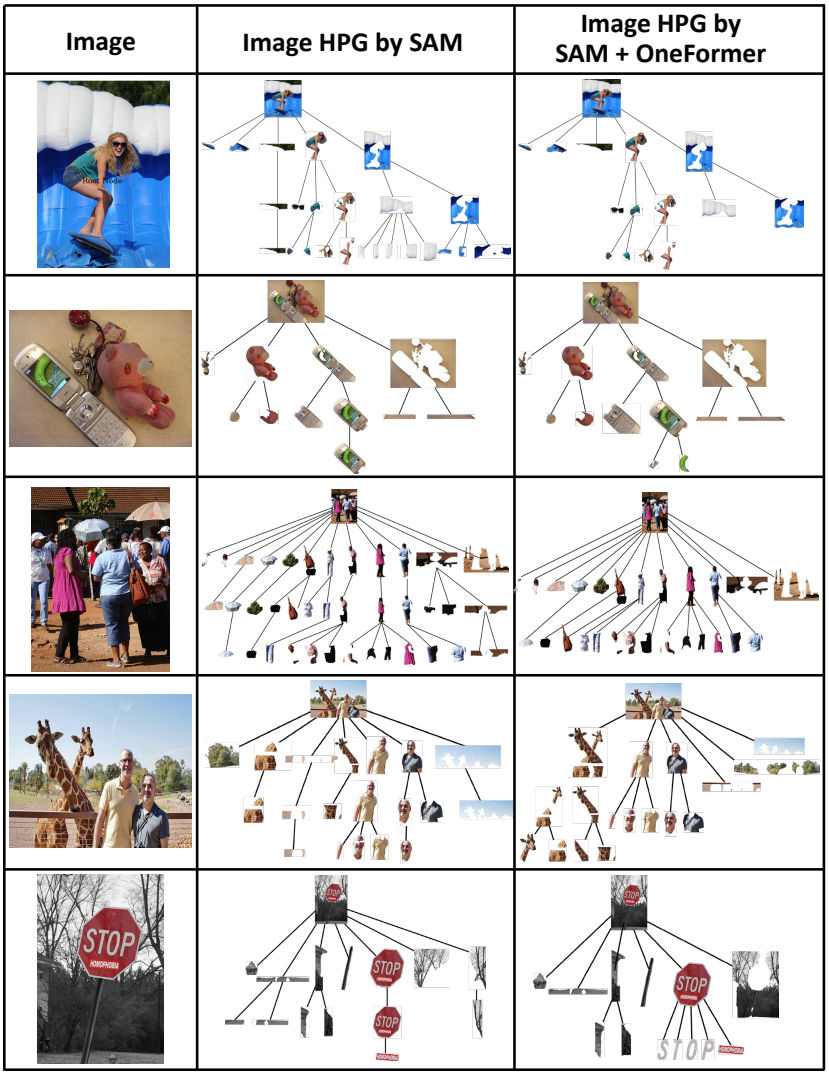

Fig. 6: Examples of image HPGs.

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)