

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

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

Abstract. With the advancements of vision-language models, the growing demand for generating customized image descriptions under length, target regions, and other various control conditions brings new challenges for evaluation. Most existing metrics, designed primarily for single-sentence image captioning with an overall matching score, struggle to accommodate complex description requirements, resulting in insufficient accuracy and interpretability of evaluation. Therefore, we propose **HiFi-Score**, a **h**ierarchical parsing graph-based **f**ine-grained evaluation metric. Specifically, we model both text and images as parsing graphs, which organize multi-granular instances into a hierarchical structure according to their inclusion relationships, which provides a comprehensive scene analysis for both modalities from global to local. Based on the fine-grained matching between the graphs, we can evaluate the fidelity to ensure text contents are related to image and the adequacy to ensure the image is covered by text at multiple levels. Furthermore, we employ the large language model to evaluate fluency of the language expression. Human correlation experiments on four caption-level benchmarks show that the proposed metric outperforms existing metrics. At the paragraph-level, we construct a novel dataset ParaEval and demonstrate the accuracy of the HiFi-Score in evaluating long texts. We further show its superiority in assessing vision-language models and its flexibility when applied to various image description tasks.

Keywords: Image Description Evaluation · Hierarchical Parsing Graph · Vision-language Model

1 Introduction

The evaluation of image description generation has always been an important and challenging topic. In recent years, with the significant improvement in the generative abilities of vision-language models, such as Flamingo [2], Llava [31], InstructBLIP [9], etc., image description tasks have become more diversified. As shown in Fig. 1, compared to classic image captioning task, image paragraphing [23] requires models to generate longer, more detailed description. Dense captioning [19] focuses on generating distinctive descriptions for local regions.

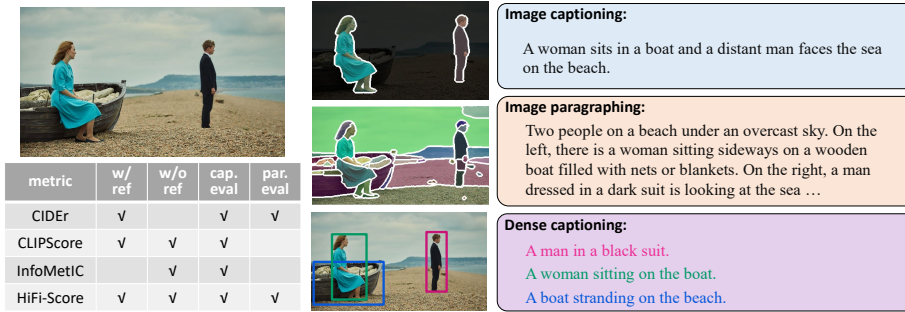


Fig. 1: Illustration of various image description tasks and evaluation metrics. Depending on different requirements, image description includes captioning, paragraphing, dense captioning, etc., requiring more general and comprehensive evaluation metrics.

Moreover, controllable captioning tasks are proposed to generate customized captions according to different requirements of length [10], format [11], and description coverage [7, 21]. Such various image description tasks pose higher demands on the evaluation metrics.

However, current evaluation metrics struggle to adapt to these complex descriptive scenarios. Early metrics (BLEU [35], CIDEr [41], etc.) evaluate captions only based on human-written references, which are not only limited by reference annotations but also hard to guarantee the accuracy of n-gram matching as the text grows longer. Recently, a number of methods based on image-text matching are proposed, such as UMIC [26] and CLIPScore [13]. Benefiting from the powerful alignment capability of vision-language models, these metrics can evaluate captions with greater accuracy and higher human consistency. However, they still encounter challenges when applied to diverse image description tasks because of the unitary modal representations and fixed evaluation methods. Besides, most metrics only give one overall matching score, which is less interpretable and readable, especially when users need to locate the incorrect expression in the candidate description.

Research on human cognition [34] suggests that humans perceive scenes in a pattern that progresses from coarse to fine and from global to local. Therefore, we propose to model both the image and description as Hierarchical Parsing Graphs (HPG) to provide a natural and structured representation. Based on the HPG, we introduce HiFi-Score to provide a fine-grained multidimensional evaluation of image descriptions. As in Fig. 2 (left), we first conduct fine-grained analysis of both modalities, and organize them into a tree-like structure, in which the root node represents the entire image or text and is recursively decomposed into child object nodes of different granularities, while the semantic relationships between objects are represented by edges between nodes. In addition, attributes are regarded as additional nodes attached to object nodes. Thus, the image-text matching for description evaluation is converted to the matching between the image HPG and the text HPG. This representation can adapt to image

descriptions of varying lengths and scopes. In addition to the top-level overall evaluation, instances at different levels can also receive independent evaluations, achieving more comprehensive and detailed results.

By progressively aggregating the fine-grained matching of various levels on the HPGs, multidimensional image description evaluation can be conducted. Inspired by the FAIEr metric [42], the evaluation of image descriptions can be decoupled into three dimensions: **fidelity**, **adequacy** and **fluency**. **Fidelity**, regarding the issue of hallucinations during generation, requires that all contents in the description should originate from the image without additions or distortions, measuring the precision of text. This implies that every element in the text HPG should have correspondence in the image HPG. **Adequacy** demands the text to capture the image gist and provide a comprehensive description within a constrained length, which can be seen as the recall of image content. The image HPG should be covered by the text HPG as much as possible, especially the areas considered important by humans. **Fluency** considers the language itself, indicating that the text is grammatically correct and linguistically smooth. We evaluate it through a prompt-based method employing the LLM. As depicted on the right side of Fig. 2, in practice, we convert the image and text HPGs into regions and phrases, respectively. We then find the most closely matching image region for each phrase and calculate the scores recursively from the bottom up on HPGs based on the similarity scores of local matching.

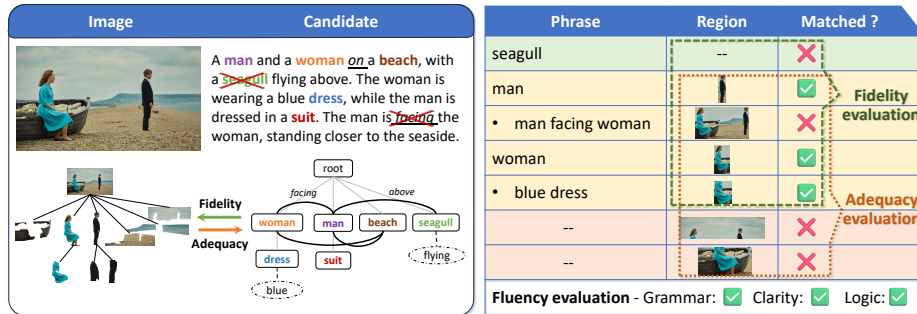


Fig. 2: Overview of image description evaluation by HiFi-Score. Left: Generation and matching of hierarchical parsing graphs. Right: Local matching between regions and phrases for score calculation.

In summary, we propose HiFi-Score, a fine-grained image description evaluation metric based on the Hierarchical Parsing Graph (HPG). Specifically, we model both the image and the text hierarchically to obtain a general and extendable structured representation and then assess the fidelity and adequacy based on the local matching of graphs. Furthermore, we make use of the LLM to evaluate fluency of the candidate descriptions, further refining the evaluation. We conduct experiments on caption-level datasets Flickr8k-Expert [14], Compos-

ite [1], Pascal-50S [41], and THumB1.0 [20], demonstrating the proposed metric exhibits higher human consistency. In addition, we construct a paragraph-level evaluation dataset ParaEval, and verify that the metric is more accurate in evaluating long texts. HiFi-Score demonstrates its potential for evaluating VLMs’ generative capabilities and the diverse image description tasks.

2 Related Works

As shown in Fig. 3, according to the main evaluation criteria type, caption evaluation metrics can be categorized into two types: reference-based and image-based. Each type can be further divided into coarse-grained and fine-grained metrics. The coarse-grained metrics mainly focus on the overall matching between the candidate and image or references, while the fine-grained metrics analyze the local semantics, providing a more comprehensive and elaborated evaluation.

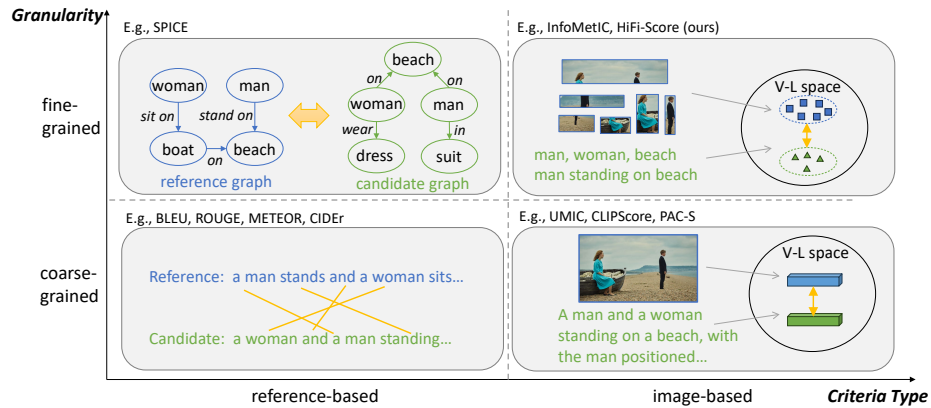


Fig. 3: Caption evaluation metrics can be classified into two types: reference-based and image-based. Each type can be further divided into coarse-grained and fine-grained based on the granularity of matching.

2.1 Reference-based metrics

These methods utilize human-written references in place of the source image as the basis for evaluating candidate captions. Inspired by machine translation, early metrics such as BLUE [35], ROUGE [30], METEOR [4], and CIDEr [41] assess the similarity between the references and candidate caption by calculating n-gram overlap based on certain rules. SPICE [3] decomposes both sentences into semantic scene graphs to obtain a more structured representation and a more detailed matching, which can be seen as a fine-grained metric.

Compared to these entirely rule-based methods, recent studies that incorporate learning-based metrics show greater resilience to linguistic ambiguity. For example, caption evaluation can be improved by utilizing a learning-based framework to combine four rule-based metrics [40]. BERTScore [43] calculates the similarity between the text embeddings acquired from the pre-trained BERT model [12]. With the development of LLMs, CLAIR [5] further improves the consistency with human, relying solely on the powerful language capability of GPT-3.5. However, all reference-based metrics ignore the original image, resulting in limitations imposed by the reference annotations and inevitable information loss from the image to the reference.

2.2 Image-based metrics

A small number of references cannot cover all the image content, and reference annotating is time-consuming and labor-intensive. Therefore, many evaluation metrics based on image-text matching have been proposed. LEIC [8] proposes a neural network that is trained to determine whether a given caption is generated by humans. TIGER [18] and REO [17] utilize the image-text matching model SCAN [27] to generate grounding vectors for the image, candidate caption, and references, and then evaluate candidate captions based on the similarity of these vectors. VIFIDEL [33] focuses on fidelity in image captioning, using weighted object labels from images and comparing them to captions via word mover’s distance [25]. FAIER [42] proposes to evaluate captions from the aspects of fidelity, adequacy and fluency, and measure the former two through cross-modal scene graph matching.

In recent years, with the development of vision-language pre-training models, the cross-modal understanding ability has been significantly improved, and evaluation metrics based on pre-training models rather than training from scratch have been widely explored. UMIC [26] refines UNITER [6] through contrastive learning, in which the model is trained to distinguish and differentiate between the ground-truth captions and various synthetic negative samples. CLIP-Score [13] leverages the ability of CLIP [38] for aligning visual and textual information by converting all candidate, image, and references into CLIP features and then computing cosine similarity. PAC-S [39] is also a CLIP-based contrastive model, which can benefit from both web-collected and cleaned data sources. These metrics provide only an overall score, thus lacking in informativeness and interpretability. To address this weakness, InfoMetIC [15] utilizes CLIP features and conducts contrastive learning at both the coarse-grained and fine-grained levels, allowing it to report the matching results of local elements for each modality. However, focusing on the single-sentence caption task, it still lacks the capability to handle longer and more flexible image description tasks. Therefore, we propose HiFi-Score based on structured HPGs for multidimensional evaluation, which is suitable for a wider range of image description scenarios.

3 Method

The overall evaluation pipeline of our HiFi-Score is shown in Fig. 4. The framework comprises three modules: hierarchical parsing graph generation (3.1), text-image grounding and matching (3.2), and multi-dimension scoring (3.3).

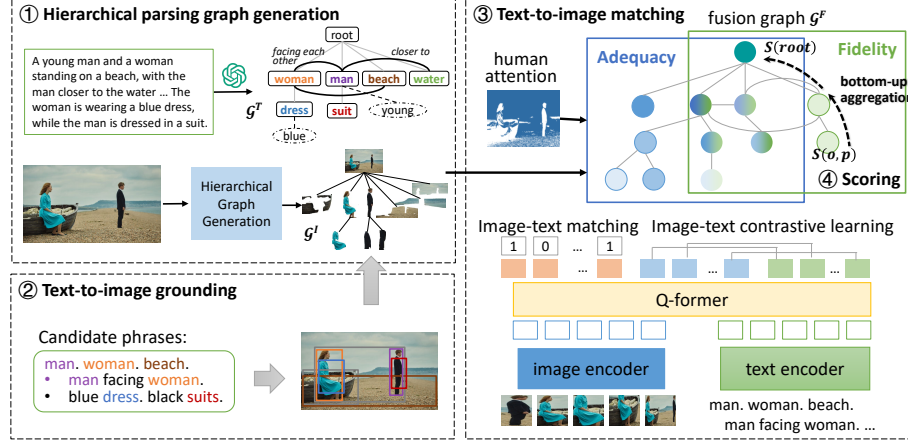


Fig. 4: Framework of the proposed HiFi-Score. We first decompose both image and text into hierarchical parsing graphs (HPG), then ground the text phrases to corresponding image regions to get a fusion graph, and calculate their matching score. Finally, the fidelity and adequacy scores are calculated recursively from the bottom up.

3.1 Hierarchical parsing graph

Definition. The Hierarchical Parsing Graph (HPG), represents both images and corresponding textual descriptions based on their fine-grained analysis, organizing global and local semantic regions or phrases into a hierarchical structure. An HPG is denoted as $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$, where $\mathcal{N} = \{n_i \mid i = 1, 2, \dots, N\}$ represents the set of N nodes, and $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{R} \times \mathcal{N}$ is the set of edges.

The root node n_0 represents the entire image (in the image HPG) or full text (in the text HPG), and other nodes represent local semantic components, that is, local regions or phrases, respectively. Each node $n_i = (o_i, A_i)$ is made of o_i , an instance of local region or phrase, and $A_i = \{a_k\}$ represents the attributes attached to the instance such as color, shape, material, etc. The edges in $\mathcal{E}$ can be represented as a binary adjacent matrix $\mathbf{M} = \{m_{ij}\}$.

In the HPG, the relationship set $\mathcal{R}$ is divided into the subordinate relationship $\mathcal{R}_{sub} = \{r_{sub}\}$ and k semantic relationships $\mathcal{R}_{sem} = \{r_i \mid i = 1, 2, \dots, k\}$. Subordination forms the main structure of the hierarchical layers, indicating that the child node is part of the parent node, such as “head - part of - dog”,

and semantic relationships can exist between any two nodes, like “person - ride - bike”. Therefore, the matrix $\mathbf{M}$ is reduced by bitwise OR from a group of adjacent matrices, i.e. $\mathbf{M} = \bigvee_{r \in \mathcal{R}} \mathbf{M}^{(r)}$, where $\mathbf{M}^{(r)}$ refers to the adjacent matrix of relationship r .

Compared to explicit linguistic expression, the attributes and relationships within the image are implicitly integrated into the instances. For simplicity, the image HPG does not contain attributes and semantic edges. Since a description may not fully capture image contents, the text HPG always contains fewer nodes and subordinate edges in contrast to the image HPG of the scenario.

Image HPG generation. In order to obtain a detailed analysis for the entire image, and generate an image HPG $\mathcal{G}^I = \{\mathcal{N}^I, \mathcal{E}^I\}$, we first utilize SAM [22] to segment images and generate masks for as many regions as possible. However, as SAM is a type of non-semantic segmentation, it may generate small fragmented masks or split complete semantic instances, affecting the integration of the main objects and background. Therefore, we further use Oneformer [16] as a panoptic segmentator to obtain masks for both objects and background. Combining the results of SAM and Oneformer, the content of the image is resolved into a comprehensive, multi-granularity region set $\mathcal{N}^I = \{o_i\}$. Each region o_i is a binary mask the same shape as the input image, and $|o_i|$ refers to the area of o_i , equivalent to the count of 1 in the mask.

To obtain the edges $\mathcal{E}^I$, we organize all regions into an image HPG based on the inclusion relationship between them. We take the entire image as the root node n_0 , and then sort the segmentation masks by area size from the smallest to the largest, that is, $|o_1| \leq |o_2| \leq \dots \leq |o_N|$. For mask o_i , considering each region $o_j \in \{o_k | k > i\}$ larger than it, the subordinate edge is assigned by thresholding the overlap ratio between o_j and o_i as

$$m_{ij}^{I, (r_{sub})} = \mathbb{1} \left(\frac{|o_i \wedge o_j|}{|o_i|} \geq \tau \right), \quad (1)$$

where $\mathbb{1}$ refers to the indicator function, $\wedge$ refers to bitwise AND, and τ is a threshold hyperparameter. If o_i is not contained in any other area, it directly becomes a child node of the root node, which means $m_{i,0}^{I, (r_{sub})} = 1$.

Text HPG Generation. Given an image description of arbitrary length, we use the LLM to parse objects, attributes corresponding to each object, and relationships between objects to phrases, and then convert them into an HPG $\mathcal{G}^T = \{\mathcal{N}^T, \mathcal{E}^T\}$. Semantic relationships can be directly obtained by analyzing the textual context, whereas subordinate relationships denote the hierarchical structure between instances, which may not be expressed in the language and require additional inference. Typical subordinate relationships include “part of” and “member of”. The former indicates that a part is a component of a whole, such as body parts and clothing, while the latter indicates that an object is part of a group of similar entities, such as “person” and “people”. For more details, please refer to the supplementary material.

3.2 Text-image grounding and matching

Text-to-image grounding. A single image description can usually be considered as a subset of the image contents. Since the image HPG is a comprehensive and hierarchical representation of the image, all elements from the text HPG $\mathcal{G}^T$ should be able to be grounded to one or more regions in the image HPG $\mathcal{G}^I$.

First, we ground the elements in $\mathcal{G}^T$ to the input image. For each phrase p_i in $\mathcal{G}^T$ (including all the objects o^T , attributes a^T , and relationships r^T) analyzed from the candidate description, all noun chunks are extracted as target entities $E_i = \{e_m \mid m = 1, 2, \dots, M\}$, where M is the number of entities extracted from p_i . Object and attribute phrases usually have only one target entity ($M = 1$), while relationship phrases have multiple entities. Then we locate each e_m in the image by GLIP [29] to obtain the probable bounding boxes $B_m = \{b_{mk} \mid k = 1, 2, \dots, K\}$, where K refers to the number of the probable bounding boxes. The bounding boxes are input as a prompt into SAM [22] to obtain more accurate segmentation masks, and we sort them by the confidence score in descent order, as $S_m = \{s_{mk} \mid k = 1, 2, \dots, K\}$. Next, according to the masks S_m , we ground the entity e_m to the nodes o_j on image HPG $\mathcal{G}^I$. The grounded nodes for entity e_m are obtained by

$$\text{grounding}(e_m, \mathcal{G}^I) = \bigcup_{k=1}^{K'} \underset{o_j \in \mathcal{G}^I}{\operatorname{argmax}} \operatorname{IoU}(s_{mk}, o_j^I), \quad (2)$$

where K' denotes the number of candidate nodes selected for e_m , according to whether the entity is singular or plural. If the entity is singular, only the first segment with the highest confidence score is used ($K' = 1$). Otherwise all the segments are considered ($K' = K$). The matching results of phrase p_i is finally obtained through aggregating the grounded segments of all entities, as follows:

$$\text{grounding}(p_i, \mathcal{G}^I) = \bigcup_{m=1}^M \text{grounding}(e_m, \mathcal{G}^I). \quad (3)$$

Through the above process, the objects o^T , attributes a^T , and relationships r^T in the text HPG are matched with as many nodes as possible in the image HPG. Thus, the fusion graph $\mathcal{G}^F = \{\mathcal{N}^F, \mathcal{E}^F\}$ that integrates the image HPG $\mathcal{G}^I$ and text HPG $\mathcal{G}^T$ can be obtained. Nodes in the text HPG but not in the image HPG represent content that does not exist in the image, violating fidelity. Nodes in the image HPG but not in the text HPG represent areas not covered by the candidate descriptions, which are considered negative for adequacy.

Text-image matching. Through grounding, we identify the matching and non-matching parts in the image and text HPGs. Then for the matching parts, we need to further calculate their matching degree. Current VLMs (CLIP [38], BLIP-2 [28], etc.) have learned a joint vision-language representation space through pre-training on large amount of image-text pairs, which enables them to effectively perform cross-modal matching between images and text. We use

BLIP-2’s ITC (image-text contrastive learning) and ITM (image-text matching) to calculate the matching score between the phrase and its corresponding region. To improve region-phrase local matching, we fine-tune BLIP-2 using dense caption annotations from Visual Genome. For phrase p in $\mathcal{G}^T$ and the matching region o in $\mathcal{G}^I$, we use the Q-former with first-stage pretraining to extract features and calculate similarity scores, and the final local matching score $S(o, p)$ is the weighted sum of ITC and ITM scores. δ denotes the weight.

$$S(o, p) = \delta \cdot \text{ITC}(o, p) + (1 - \delta) \cdot \text{ITM}(o, p) \quad (4)$$

3.3 Multidimensional Scoring

Based on the fusion parsing graph and the local similarity of the matched parts, we evaluate the candidate description from two aspects, fidelity and adequacy. **Fidelity** is calculated based on the matched parts and the text HPG, indicating the extent to which the text description conforms to the image. For each node $p \in \mathcal{G}^T$ with its matching score s_p , its fidelity score is a combination of its own score and the average score of its child nodes $C(p)$. The child nodes of p include not only the subordinate sub-nodes belonging to it, but also the attribute nodes attached to it and the relationship edges with it as the subject. The fidelity score for the entire description is recursively calculated from the bottom up, starting from the local matching score s_p of leaf nodes, until it reaches the root node.

$$S_F(p) = \alpha \cdot s_p + \frac{1 - \alpha}{|C(p)|} \sum_{c \in C(p)} s_c, \quad S_F(I, T) = S_F(\text{root}) \quad (5)$$

in which the hyperparameter α is used to adjust the weights of the node itself and its child nodes. The overall fidelity score of the entire image and the candidate description $S_F(I, T)$ is the root score $S_F(\text{root})$.

Adequacy is calculated based on the matched parts and the image HPG $\mathcal{G}^I$, combined with human attention, indicating how much of the content in the image is covered by the text description. Similar to fidelity, adequacy is also obtained through a bottom-up recursive calculation along the tree structure based on each node $o \in \mathcal{G}^I$. By introducing human attention, we can further determine the importance of different regions in the image HPG, which are then incorporated as weights ω_o into the adequacy calculation.

$$S_A(o) = \omega_o \left(\alpha \cdot s_o + \frac{1 - \alpha}{|C(o)|} \sum_{c \in C(o)} s_c \right), \quad S_A(I, T) = S_A(\text{root}) \quad (6)$$

The overall adequacy score of the entire image and the candidate description $S_A(I, T)$ is the root score $S_A(\text{root})$.

There are various ways to obtain human attention depending on the tasks. For general image captioning and image paragraphing, we introduce a saliency map as the human attention map. If there are human-written references, the areas mentioned by more references should have higher weights. For tasks that

require the description for specific areas, such as dense captioning and object-controllable captioning, the weight of the designated areas can be set to 1, while the weights of other regions are set to 0.

Finally, by simply adding the fidelity score and the adequacy score, the total score for the candidate description is

$$S(I, T) = \gamma \cdot S_F(I, T) + (1 - \gamma) \cdot S_A(I, T), \quad (7)$$

in which γ denotes the weight of fidelity score and adequacy score.

Fluency Evaluation. The evaluation of fluency presents ongoing challenges in natural language processing and linguistics, with inherent subjectivity and complex evaluation criteria. Compared with fidelity and adequacy that rely on image-text matching, fluency is purely linguistic unrelated to the image, making it to some extent orthogonal to the other two. 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, and provide specific reasons. More details (including prompts) can be found in the supplementary material.

4 Experiments

4.1 Implementation details

To obtain image HPGs, we use the *vit_h* model of SAM [22] to generate non-semantic masks and conduct panoptic segmentation by OneFormer [16] with the backbone of DiNAT-L. The gpt-3.5-turbo-1106 API is utilized for text HPGs and fluency evaluation based on prompts. We use GLIP-L [29] to grounding each text phrase to corresponding image region. We use the provided first-stage-pretrained BLIP-2 [28] model for global image-text matching and fine-tune with dense captioning annotations of Visual Genome [24] for local matching. The hyperparameters α , δ and γ in score calculation are set to 0.5. The default human attention weights are obtained by the PyTorch implementation of saliency detection method [44]. Please refer to our project page¹ for more details.

4.2 Caption-level judgement

Datasets. We conduct quantitative experiments on four caption-level benchmarks (Flickr8k-Expert [14], Composite Dataset [1], THumB1.0 [20], and Pascal-50S dataset [41]) to compare our metric with the state-of-the-art metrics. Following previous works, Kendall’s τ correlation on both 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 these metrics and human evaluations, respectively. More details about the datasets can be found in the supplementary material.

¹ <https://github.com/lydia7632/HiFi-Score>

Table 1: Caption-level comparison on Flickr8k-Expert(F-Ex), Composite(Com), THumB1.0 and Pascal-50S. ‘w/ ref’ indicates that the metric incorporates groundtruth references when evaluating captions.

Metric	w/ ref	F-Ex	Comp.	Pascal-50S					THumB1.0		
		τ_c	τ_c	HC	HI	HM	MM	Mean	P	R	Total
BLEU-4 [35]	✓	30.8	30.6	52.5	90.4	63.0	42.3	55.8	0.12	0.03	0.10
ROUGE-L [30]	✓	32.3	32.4	55.0	95.3	93.1	58.7	75.5	0.18	0.08	0.19
METEOR [4]	✓	41.8	38.9	59.0	97.7	93.9	62.0	78.2	0.18	0.08	0.18
CIDEr [41]	✓	43.9	37.7	53.7	98.1	90.8	63.7	76.6	0.21	0.10	0.22
SPICE [3]	✓	44.9	40.3	56.9	96.3	87.1	66.4	76.7	0.20	0.09	0.21
BERT-Score [†] [43]	✓	39.2	30.1	54.4	96.1	94.3	56.4	75.3	0.20	0.10	0.21
FAIER-4 [†] [42]	✓	52.6	57.7	56.7	99.6	92.7	73.4	81.4	-	-	-
RefCLIPScore [13]	✓	53.0	55.2	64.5	99.6	95.4	72.8	83.1	0.31	0.26	0.41
UMIC [†] [26]		46.8	56.1	66.1	99.8	98.1	76.2	85.1	-	-	-
CLIPScore [13]		51.4	53.8	55.4	99.3	96.4	71.5	80.7	0.17	0.28	0.32
InfoMetIC [†] [15]		54.2	59.2	69.0	99.8	94.0	78.3	86.5	0.21	0.32	0.38
PAC-S [†] [39]		53.9	55.7	60.6	99.3	96.9	72.9	82.4	-	-	-
BLIPScore [32]		52.5	61.6	60.8	99.8	96.5	71.5	82.2	0.41	0.23	0.45
HiFi-Score _{ref}	✓	58.4	65.7	59.1	99.8	96.9	76.3	83.0	0.42	0.22	0.45
HiFi-Score		58.4	65.8	59.5	99.8	96.9	75.8	83.0	0.43	0.22	0.45

[†] indicates the result is cited directly from their original paper.

Comparison with SOTA. Table 1 presents the results for caption-level judgement compared with the state-of-the-art metrics both with and without references. BLIPScore (originally used for evaluating text-to-image generation in [32]) utilizes BLIP-2 as the base image-text model, which can be considered a global matching baseline. Similar to CLIPScore, BLIPScore calculates the cosine similarity between image and text feature. HiFi-Score is our reference-free model, utilizing the saliency detection [44] output as human attention weights for calculating adequacy score by default, while HiFi-Score_{ref} assigns weights based on the frequency of regions mentioned in human references.

As in Table 1, our HiFi-Score outperforms SOTA metrics on Flickr8k-Expert [14], Composite [1] and THumB1.0 [20] and gains comparable performance on Pascal-50S [41]. Compared to BLIPScore with the same base model, the performance improvement of HiFi-Score on most datasets indicates that incorporating fine-grained analysis can enhance consistency with human evaluation. The improvement on THumB1.0 is slight, possibly because model-generated captions in this dataset are shorter and less detailed, limiting HiFi-Score’s capability. Compared to InfoMetIC, a fine-grained metric that matches object regions with words, HiFi-Score’s superior performance highlights the effectiveness of structured modeling of local details within both images and texts.

For Pascal-50S, UMIC and InfoMetIC outperform other metrics, mainly due to their contrastive learning to differentiate positive and carefully-designed negative captions, which aligns well with Pascal-50S’s test setting of choosing the

better caption from a pair. Without targeted training, HiFi-Score’s good performance verifies its generalization.

Table 2: Ablation study of modules and settings. ‘Finetuned’ denotes whether the BLIP-2 local matching model is finetuned. ‘Fid.’ and ‘Ade.’ mean fidelity and adequacy. ITC and ITM are two sub-scores of image-text matching. ‘VG’ is Visual Genome.

Finetuned	Fid. / Ade.	ITC / ITM	Flicker8k-Exp Composite Pascal-50S THumB1.0			
			τ_c	τ_c	Mean	Total
None	Fid.+Ade.	ITC+ITM	55.0	64.9	82.9	44.7
VG	Fid.	ITC	58.0	62.4	82.7	39.0
		ITM	54.1	65.2	81.7	42.7
		ITC+ITM	57.4	65.4	82.5	43.3
VG	Ade.	ITC	54.4	59.5	82.4	36.9
		ITM	52.7	64.4	81.4	41.9
		ITC+ITM	55.3	64.6	82.3	42.7
VG	Fid.+Ade.	ITC	58.3	62.4	83.1	42.6
		ITM	54.7	65.6	81.8	43.6
		ITC+ITM	58.4	65.8	83.0	44.8

Ablation Study. Here we conduct ablation experiments to validate the effectiveness of modules and settings. As shown in Table 2, the full model finetuned on Visual Genome (last row) yields the overall best performance among the variant models. Compared to the model without fine-tuning (first row), the performance improvement demonstrates that fine-tuning is beneficial for enhancing the model’s local matching capability. When comparing fidelity and adequacy, fidelity’s performance is a bit higher than adequacy’s, as correctness (mainly reflected in fidelity) is the most fundamental criterion for image descriptions. The best outcomes are achieved when both are combined. As for ITC and ITM, it is observed that Flicker8k-Exp and Pascal-50S prefer ITC, whereas Composite and THumB1.0 prefer ITM. This difference likely arises because ITC excels in global alignment between modals, while ITM has stronger ability to discern fine-grained differences. The combination of ITC and ITM can bring performance gain by taking the advantages of both. More experimental analyses (trade-off of weights, impact of human attention, etc.) are in the supplementary materials.

4.3 Paragraph-level judgement

ParaEval Dataset Construction. Unlike image captioning evaluation, the evaluation for image paragraphing remains underdeveloped and lacks benchmarks to measure the rationality of existing metrics. Therefore, we constructed the **ParaEval** dataset to assess the accuracy of existing metrics in evaluating long-context image descriptions. ParaEval collects 4,000 pairs of image and its

paragraph description from the ImageParagraphs Dataset [23] and Localized Narratives Dataset [37]. We design four types of negative samples: plausible descriptions, negative objects, negative attributes, and negative relationships as displayed in Table 3, and ask metrics to distinguish between groundtruth and wrong descriptions. More details can be found in the supplementary material.

Comparison with Other Metrics. We compare HiFi-Score with CLIPScore and BLIPScore, which can to some extent accommodate the growth of text length. The maximum input text length of CLIP is 77 tokens, less than most paragraphs, so we designed CLIPScore_{trunc} that directly truncates the part that exceeds the limit and CLIPScore_{avg} that calculates the similarity between each sentence in the paragraph and the image and then takes the average value as the total score. As shown in Table 4, the classification accuracy of HiFi-Score surpasses that of other methods across all four types of negatives, proving that fine-grained matching enhances the ability to identify detailed errors. For all metrics, distinguishing plausible descriptions is easiest because of the sufficient details in long texts. Among the other three, performance of distinguishing incorrect objects is best, and the worst is relationships, proving that understanding relationships between objects is still a key weakness of learning-based methods.

Table 3: Example of the ParaEval dataset. Four different negative samples are constructed for one groundtruth paragraph.

[groundtruth] There is a person doing a jump on a snowboard. The rider is wearing a blue jump suit. The snowboard is white in color.	
[plausible] A person jumping on a yellow snowboard above the snow. They are wearing a white jacket and black pants. The snow is below the man.	[attribute] There is a person doing a jump on a snowboard. The rider is wearing a black working suit. The snowboard is red in color.
[object] There is a person doing a jump on a trampoline . The rider is wearing a blue jump suit. The trampoline is white in color.	[relation] There is a person sitting on a snowboard. The rider is wearing a blue jump suit. The snowboard is white in color.

Table 4: Comparison for pairwise judgment on four types of negative samples on the image paragraph set.

	Plau.	Obj.	Attr.	Rel.
ClipScore_{trunc}	87.5	76.4	77.7	53.2
ClipScore_{avg}	76.4	85.6	78.7	56.1
BLIPScore	92.5	76.6	82.2	58.1
HiFi-Score	95.4	93.2	87.7	63.8

Visualization. The evaluation process of HiFi-Score is visualized in Fig. 5, which demonstrates the flexibility and interpretability, and Fig. 6, which shows the evaluation of multiple captions and paragraphs generated by vision-language models KOSMOS-2 [36] and InstructBLIP [9], respectively. The error localization ability of HiFi-Score in long text evaluations is beneficial for model diagnostics.

Limitations. As a learning-based metric, the limitations of HiFi-Score mainly stem from the pre-trained modules, including the out-of-distribution (OOD) problem, fairness that originates from training data bias, and domain generalization for specific fields. These may lead to mismatching and inaccurate scoring. Nevertheless, it is believed such issues will be improved with the development of visual, language, and multi-modal models. Since our framework is open, new superior models can be easily plugged in, further enhancing the evaluation.

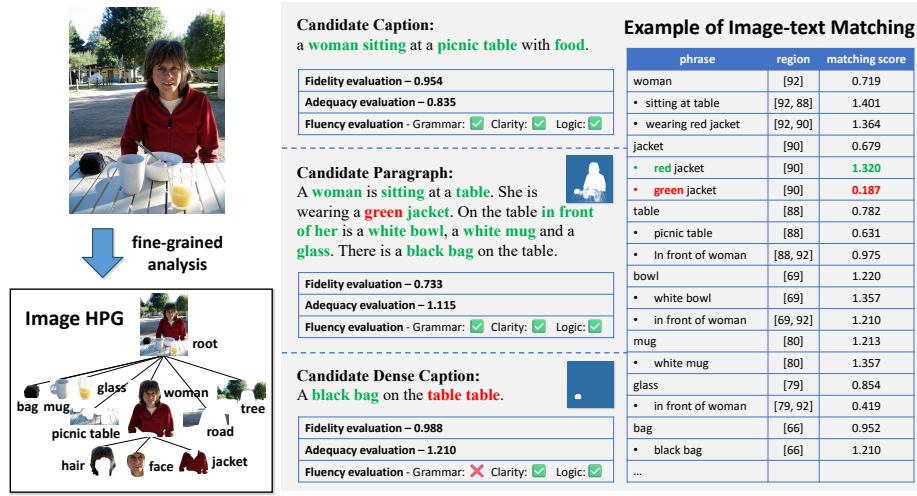


Fig. 5: Visualization of HiFi-Score evaluating different image description tasks.

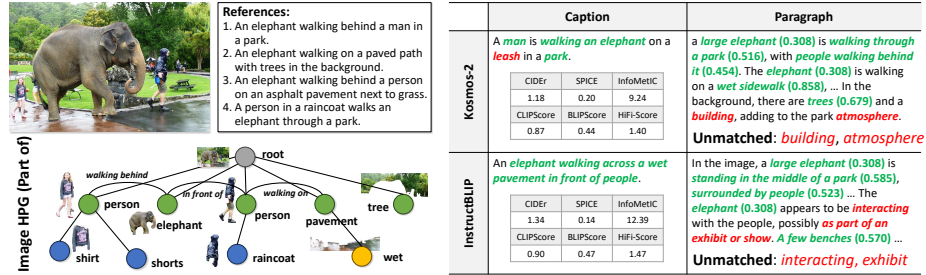


Fig. 6: Visualization of HiFi-Score evaluating image descriptions generated by different vision-language models.

5 Conclusion

In conclusion, we propose a novel image description evaluation metric, HiFi-Score, which provides fine-grained multi-dimensional evaluation based on hierarchical parsing graphs. Specifically, we model both images and texts as HPGs, analyzing all semantic instances from global to local and incorporating attributes and semantic relationships. Then, the scores of fidelity and adequacy are calculated gradually from the bottom up. We further evaluate the fluency of language expression taking the advantage of the LLMs. Experimental results on caption-level datasets demonstrate HiFi-Score’s higher human consistency. Moreover, we construct paragraph-level dataset ParaEval to evaluate current metrics with long-context descriptions, and validate the superiority of HiFi-Score. Extensive evaluations demonstrate the potential of HiFi-Score for evaluating VLMs’ generative capabilities and the diverse image description tasks.

Acknowledgements

This work is partially supported by Natural Science Foundation of China under contract No. U21B2025, and National Key R&D Program of China No. 2021ZD0111901, 2023YFF1105104.

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. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*. pp. 23716–23736 (2022)
3. 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)
4. Banerjee, S., Lavie, A.: METEOR: An automatic metric for mt evaluation with improved correlation with human judgments. In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. pp. 65–72 (2005)
5. Chan, D., Petryk, S., Gonzalez, J., Darrell, T., Canny, J.: CLAIR: Evaluating image captions with large language models. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 13638–13646 (2023)
6. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: UNITER: Universal image-text representation learning. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 104–120 (2020)
7. Cornia, M., Baraldi, L., Cucchiara, R.: Show, control and tell: A framework for generating controllable and grounded captions. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 8307–8316 (2019)
8. Cui, Y., Yang, G., Veit, A., Huang, X., Belongie, S.: Learning to evaluate image captioning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 5804–5812 (2018)
9. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS)*. pp. 49250–49267 (2023)
10. Deng, C., Ding, N., Tan, M., Wu, Q.: Length-controllable image captioning. In: *Proceedings of the European conference on computer vision (ECCV)*. pp. 712–729 (2020)
11. Deshpande, A., Aneja, J., Wang, L., Schwing, A.G., Forsyth, D.: Fast, diverse and accurate image captioning guided by part-of-speech. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 10695–10704 (2019)

12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). pp. 4171–4186 (2019)
13. 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)
14. 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)
15. Hu, A., Chen, S., Zhang, L., Jin, Q.: InfoMetIC: An informative metric for reference-free image caption evaluation. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL). pp. 3171–3185 (2023)
16. 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)
17. Jiang, M., Hu, J., Huang, Q., Zhang, L., Diesner, J., Gao, J.: REO-relevance, extraneous, omission: A fine-grained evaluation for image captioning. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 1475–1480 (2019)
18. Jiang, M., Huang, Q., Zhang, L., Wang, X., Zhang, P., Gan, Z., Diesner, J., Gao, J.: TIGer: Text-to-image grounding for image caption evaluation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 2141–2152 (2019)
19. Johnson, J., Karpathy, A., Fei-Fei, L.: DenseCap: Fully convolutional localization networks for dense captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4565–4574 (2016)
20. 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)
21. Kim, D.J., Choi, J., Oh, T.H., Kweon, I.S.: Dense relational captioning: Triple-stream networks for relationship-based captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6271–6280 (2019)
22. 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)
23. 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)
24. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* **123**, 32–73 (2017)

25. Kusner, M., Sun, Y., Kolkin, N., Weinberger, K.: From word embeddings to document distances. In: Proceedings of the 32nd International Conference on Machine Learning (ICML). pp. 957–966 (2015)
26. Lee, H., Yoon, S., Deroncourt, F., Bui, T., Jung, K.: UMIC: An unreference metric for image captioning via contrastive learning. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP). pp. 220–226 (2021)
27. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: Proceedings of the European conference on computer vision (ECCV). pp. 201–216 (2018)
28. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of the 40th International Conference on Machine Learning (ICML). pp. 19730–19742 (2023)
29. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10965–10975 (2022)
30. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81 (2004)
31. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS). pp. 34892–34916 (2023)
32. Lu, Y., Yang, X., Li, X., Wang, X.E., Wang, W.Y.: LLMscore: Unveiling the power of large language models in text-to-image synthesis evaluation. In: Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS). pp. 23075–23093 (2023)
33. Madhyastha, P., Wang, J., Specia, L.: VIFIDEL: Evaluating the visual fidelity of image descriptions. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 6539–6550 (2019)
34. Navon, D.: Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology* **9**(3), 353–383 (1977)
35. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 311–318 (2002)
36. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Ye, Q., Wei, F.: Grounding multimodal large language models to the world. In: Proceedings of the 12th International Conference on Learning Representations (ICLR) (2024)
37. 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)
38. 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)
39. Sarto, S., Barraco, M., Cornia, M., Baraldi, L., Cucchiara, R.: Positive-augmented contrastive learning for image and video captioning evaluation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6914–6924 (2023)

40. Sharif, N., Bennamoun, M., White, L.R., Shah, S.A.A.: Learning-based composite metrics for improved caption evaluation. In: Proceedings of ACL 2018, Student Research Workshop. pp. 14–20 (2018)
41. 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)
42. Wang, S., Yao, Z., Wang, R., Wu, Z., Chen, X.: FAIER: Fidelity and adequacy ensured image caption evaluation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14050–14059 (2021)
43. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: Evaluating text generation with bert. In: Proceedings of the 8th International Conference on Learning Representations (ICLR) (2020)
44. Zhao, T., Wu, X.: Pyramid feature attention network for saliency detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3085–3094 (2019)