Polos: Multimodal Metric Learning from Human Feedback for Image Captioning

Yuiga Wada Kanta Kaneda Daichi Saito Komei Sugiura
Keio University
https://yuiga.dev/polos

Abstract

Establishing an automatic evaluation metric that closely aligns with human judgments is essential for effectively developing image captioning models. Recent data-driven metrics have demonstrated a stronger correlation with human judgments than classic metrics such as CIDEr; however they lack sufficient capabilities to handle hallucinations and generalize across diverse images and texts partially because they compute scalar similarities merely using embeddings learned from tasks unrelated to image captioning evaluation. In this study, we propose Polos, a supervised automatic evaluation metric for image captioning models. Polos computes scores from multimodal inputs, using a parallel feature extraction mechanism that leverages embeddings trained through large-scale contrastive learning. To train Polos, we introduce Multimodal Metric Learning from Human Feedback (M^{2}LHF), a framework for developing metrics based on human feedback. We constructed the Polaris dataset, which comprises 131K human judgments from 550 evaluators, which is approximately ten times larger than standard datasets. Our approach achieved state-of-the-art performance on Composite, Flickr8K-Expert, Flickr8K-CF, PASCAL-50S, FOIL, and the Polaris dataset, thereby demonstrating its effectiveness and robustness.

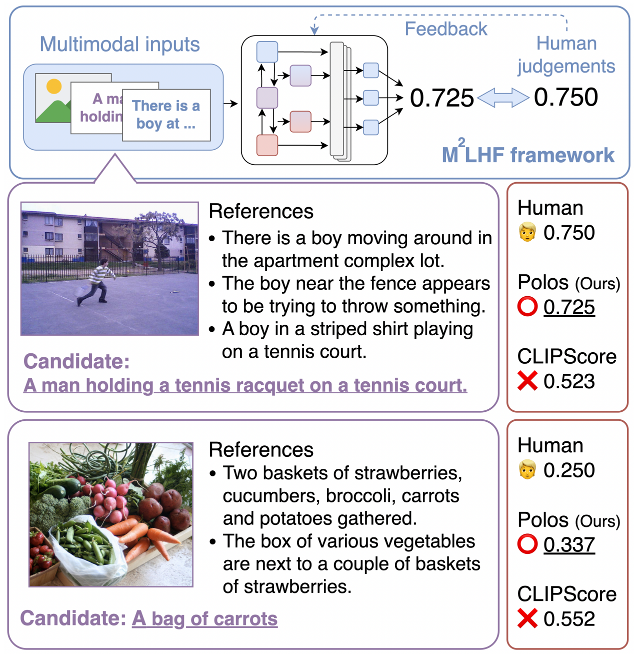

Figure 1. Our supervised metric Polos computes evaluation scores from multimodal inputs by integrating human feedback within the novel framework M^{2}LHF. Polos is capable of modeling intricate relationships within the vector space of text-image pairs as well as text-text pairs, thereby effectively evaluating the depicted samples.

1. Introduction

Extensive research on image captioning has yielded a wide array of practical applications in society, spanning from assisting people who are blind to facilitating dialog about images and answering questions derived from visual content (e.g., [21, 24, 68]). To effectively advance the development of image captioning models, it is crucial to establish an automatic evaluation metric that closely aligns with human judgment. Previous research has shown that classical automatic evaluation metrics [10, 12, 40, 51, 64] exhibit weak correlation with human judgments [10, 26]. This has prompted the introduction of data-driven automatic evaluation metrics [26, 34, 77, 78]. However, by merely measuring the similarity of embeddings learned from tasks unrelated to image captioning, these metrics potentially misjudge caption quality, which raises concerns about their accuracy in evaluating image captioning models. Furthermore, some experiments have identified the limitations of these metrics regarding handling *hallucinations* adequately.

Although recent similarity-based [26, 31, 34, 77, 78] and learning-based [35, 55] metrics have demonstrated performance superior to classic metrics, they are still not completely satisfactory. For instance, while the correlation between the state-of-the-art (SOTA) metric and human judgments on the Flickr8K dataset [27] is only about 0.56, the correlation coefficient among human judgments is approximately 0.73 [10]. This discrepancy may emerge because similarity-based metrics merely compute cosine similarity from well-known embeddings (e.g., [16, 19, 52]) learned from tasks unrelated to image captioning evaluation, which potentially leads to a misrepresentation of caption quality. Moreover, many learning-based metrics are specifically for limited settings (e.g., closed-vocabulary settings), and therefore fail to accommodate diverse images and texts.

In this paper, we propose Polos, a supervised automatic evaluation metric for image captioning models. Alongside this, we introduce Multimodal Metric Learning from Human Feedback (M^2LHF), a framework used to develop a practical supervised metric for image captioning. Fig.1 illustrates our proposed metric Polos and the M^2LHF framework. As a key contribution, our proposed metric fuses both similarity-based and learning-based approaches. Previous similarity-based approaches merely compute scalar similarities using classic methodologies (e.g., cosine similarity and optimal transport), whereas our metric models intricate relationships in the vector space of text-image pairs and text-text pairs. This is achieved through the *parallel feature extraction mechanism* that leverages SimCSE [22] and CLIP [52], which we provide details for in Section 3.3.

To train a metric that embodies the aforementioned characteristics through the M^2LHF approach, we have constructed the Polaris dataset, which contains a diverse range of human judgments. Compared with existing datasets, Polaris contains a greater diversity of captions and a more extensive range of evaluations. Notably, our dataset comprises 131K human judgments, whereas even the largest existing dataset [1] contains only approximately 12K in total. Regarding the total number of evaluators involved, even the CapEval1K dataset [35], which has the largest number of evaluators to the best of our knowledge, was assessed by only five evaluators. By contrast, our dataset is distinguished by its assessment by a much higher number of evaluators at 550 evaluators. Furthermore, our dataset is superior to existing datasets such as [1, 10, 27, 35] in terms of the inclusion of diverse captions, which were collected from humans and generated from ten image captioning models, including modern models.

The main contributions of this paper[1] are as follows :

- We propose Polos[2], a supervised automatic evaluation metric for image captioning models.
- We introduce M^2LHF, a novel framework used to develop a practical metric for image captioning.
- We introduce a *parallel feature extraction mechanism* that leverages text embeddings [43] pretrained with SimCSE and vision-language embeddings [52].
- We constructed the Polaris dataset, which contains 131,020 human judgments from 550 evaluators.
- We achieved SOTA performance on image captioning benchmarks including Composite, Flickr8K-Expert, and Flickr8K-CF, PASCAL-50S, FOIL, and Polaris.

[1]Project page: https://yuiga.dev/polos

[2]**P**ractical visi**O**n-and-**L**anguage evaluati**O**n metic for image captioning model**S**

2. Related Work

Data-driven metrics can be broadly divided into similarity-based metrics [26, 31, 34, 77] and learning-based metrics [35, 53, 55]. Similarity-based metrics compute similarities using classic approaches such as cosine similarity and optimal transport in an unsupervised manner, whereas learning-based metrics compute scores in a supervised manner.

Standard and similarity-based metrics. Standard automatic metrics for evaluating image captioning models include BLEU [51], ROUGE [40], METEOR [12], CIDEr [64], and SPICE [10], which are primarily based on either n-grams or scene graphs. Extensions to these standard metrics, such as CIDEr-R [49] and JaSPICE [65], have also been proposed. Despite their widespread use, in several studies, researchers have highlighted the limitations of these metrics, indicating their suboptimal performance [26, 34, 53, 55, 56]. This has led to the emergence of data-driven metrics such as BERTScore [77] and MoverScore [78]. Additionally, there are similar metrics that leverage image features directly, such as [26, 29, 31, 34, 35].

CLIPScore [26] evaluates captions in an unsupervised manner by computing their similarity with embeddings derived from CLIP [52]. Its distinctive feature is its capacity not only to evaluate based on a reference-with-image manner but also in a reference-free context. Similarly, PAC-S [55] fine-tunes CLIP on *generated* image-text pairs and evaluates captions in the same manner. However, these dependences solely on the computation of cosine similarity between CLIP embeddings could potentially constrain its effectiveness and robustness across diverse scenarios.

MID[31] uses the negative Gaussian cross-mutual information using CLIP features. It was proposed as a bridge between reference-with-image and reference-free metrics and has demonstrated strong performance across multiple image captioning benchmarks. However, it is crucial to highlight that MID relies solely on the text embeddings provided by CLIP for processing text data, which could potentially lead to suboptimal evaluation. This concern stems from the fact that CLIP is predominantly trained on a dataset that comprises mainly short noun phrases, which means that it is possibly suboptimal for evaluating the longer sentences typically generated by image captioning models [55].

Learning-based metrics. Limited learning-based metrics exist for image captioning models; however multiple learning-based metrics exist in the field of evaluation of text generation [45, 56, 59, 70–72], including RUSE [59] and COMET [53]. COMET is a metric trained using human judgments that has demonstrated robust performance in evaluating machine translations. COMET comprises both an estimator model that directly predicts human judgments and a ranking model that predicts the quality order of the generated translations.

By contrast, a few learning-based metrics exist that were specifically designed for image captioning [32, 35, 36]. One such metric is UMIC [35], which is among the few learning-based metrics tailored for evaluating image captioning models. UMIC is a fine-tuned UNITER [14] model designed to rank captions against each other using CapEval1K. However, as we argue later, such *ranking* models have shortcomings when they process multiple potential references, such as varying focal points in captions and subjective variations in expression. Furthermore, UMIC is only a straightforward fine-tuned UNITER model and it cannot adequately handle diverse images and text, such as those associated with an open-vocabulary setting.

Datasets and benchmarks. Standard datasets commonly used for the evaluation of image captioning include Flickr8K-Expert, Flickr8K-CF [27], Composite [1] and PASCAL-50S [64]. The Flickr8K-Expert and Flickr8K-CF datasets comprise a significant amount of human judgments on captions provided by humans. However, these datasets do not contain any captions generated by models, which presents an issue from the perspective of the domain gap when using them for training metrics. The Composite dataset [1] encompasses 12K human judgments across images collected from MSCOCO [41], Flickr8k [27], and Flickr30k [73]. Although each image initially contains five references, only one reference was selected for human judgments within the dataset. The CapEval1k dataset was introduced by [35] for training automatic evaluation metrics. We note that the CapEval1K dataset has several limitations: it is a closed dataset, uses outdated models such as [11, 28, 54], and includes only 1K human judgments. In stark contrast, our Polaris dataset offers significant advantages: it is openly accessible, incorporates inferences from modern models [17, 37, 38, 62, 66, 67, 69, 75], and includes a substantial 131K human judgments. A meta-analysis of these datasets can be found in Appendix B.1.

Standard datasets for image captioning include MS-COCO, nocaps [7], Flickr30K, and CC3M [57]. The no-caps dataset contains a greater diversity of classes than MS-COCO, which enables a more comprehensive evaluation of image captioning models' ability to generate diverse captions. Our Polaris dataset is built on inferences from MS-COCO and nocaps to ensure caption diversity.

3. Methodology

3.1. Meta-Analysis

When designing automatic evaluation metrics, the decision to use an unsupervised or supervised approach is important. As detailed in previous research [74], this binary distinction can be further decomposed into four categories: *regression*, *ranking*, *matching*, and *generation*. We will discuss the above categories in the following subsections.

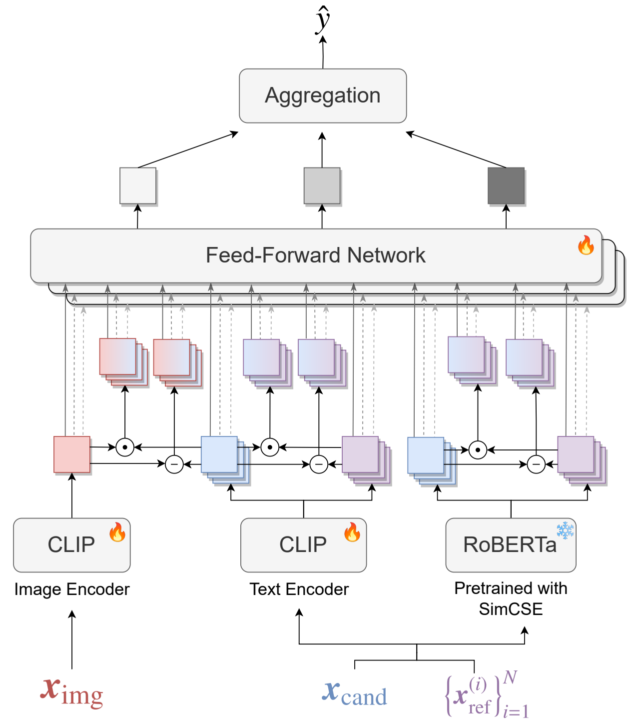

Figure 2. Overview of the proposed metric. In alignment with the principles of M^2LHF, Polos computes the evaluation $\hat{y}$ based on multimodal inputs and regresses the human evaluation. The proposed metric extracts effective features for caption evaluation using the difference and Hadamard product of features derived from both CLIP and RoBERTa.

Unsupervised approaches. In the domain of unsupervised approaches, metrics such as BERTScore [77], CLIP-Score [26], and BARTScore [74] are prominent. However, each of these metrics has its limitations for evaluating image captioning models. For instance, *matching* models such as BERTScore measure the similarity between tokens, which makes them ill-suited to explicitly handling images. Simultaneously, CLIPScore, which computes the similarity between sentence embeddings and image embeddings, has a major drawback. Because CLIP was designed to match an entire image to a text description, it has shortcomings regarding capturing the fine-grained alignment between specific image regions and text spans as indicated in [23, 79]. This limitation suggests that there is potential to improve performance, particularly by transitioning to a supervised approach. Similarly, unsupervised *generation* models such as BARTScore encounter challenges in the evaluation of image captioning models. This difficulty stems from the limited availability of large-scale, pretrained, and lightweight multimodal encoders.

Supervised approaches. Based on the above discussion, we believe that supervised metrics have distinct advantages. Both *regression* and *ranking* models are viable options;

however, we believe that *regression* models are better suited for evaluating image captioning models that involve multiple references because *ranking* models cannot adequately handle multiple potential references, such as varying areas of interest in captions and inherent subjectivity in expression. For instance, the comparison of captions with different focal points lacks substantive meaning and ranking subjective expressions is difficult. Although COMET [53] introduces both *ranking* and *regression* models, the task of evaluating machine translation with limited interpretation differs significantly from our task. This distinction is a factor that prevents the straightforward adoption of a *ranking* model. Given these considerations, in this study, we propose a metric based on *regression*.

3.2. Metric Design

Reference-free vs reference-with-image. In several studies, researchers have introduced reference-free metrics for image captioning [26, 35, 55]. However, these metrics have not achieved performance comparable with metrics that consider references. Additionally, the performance of reference-free metrics heavily relies on the alignment capabilities between image and language features. This is evident from the fact that, on the FOIL dataset, CLIP-S underperforms compared with traditional metrics such as CIDEr [26]. Such results suggest potential limitations in handling *hallucinations*, which poses a significant challenge in the field of image captioning. Given these considerations, we chose not to adopt the reference-free problem setting.

Limitations of CLIP text embeddings. As previously highlighted, many data-driven metrics do not have adequate generalization capabilities across diverse image and text types. This inadequacy partly stems from the sole use of text embeddings from the CLIP text encoder. CLIP provides versatile textual and visual features, as demonstrated by their performance in various tasks. However, the CLIP model, pretrained on web-collected image-caption pairs, is likely to be suboptimal for evaluation metrics because these annotations typically lack the richness and descriptiveness necessary for evaluating generated long captions, as indicated in [55]. Consequently, we posit that sentence embeddings pretrained with supervised SimCSE [22] may be a more advantageous approach than CLIP. This is partially supported by the observation that SimCSE outperformed previous sentence embedding techniques in the semantic text similarity task [2–6, 13, 47].

3.3. Proposed Method: Polos and M²LHF

We propose Polos, a supervised automatic evaluation metric tailored for image captioning models. Fig.2 shows the overview of the proposed Polos. To enhance robustness and practicality, we also present Multimodal Metric Learning from Human Feedback ($\mathrm{M^2LHF}$), a novel framework for developing metrics based on human feedback. Within $\mathrm{M^2LHF}$, a metric computes the evaluation $\hat{y}$ based on multimodal input $\boldsymbol{x}$ and directly regresses the human evaluation y. Our method is inspired by automatic evaluation metrics for machine translation, such as COMET and BLEURT [56], which were explicitly designed to predict human judgments. However, our framework differs from previous works [53, 56, 59] in that it handles both images and text, and learns directly from human judgments based on multimodal inputs. We consider that our approach, which applies regression using both image and language features, is considered to be broadly applicable to automatic evaluation metrics that have learnable parameters.

We define the input $\boldsymbol{x}$ to the model as follows:

$$\boldsymbol{x} = \left\{\boldsymbol{x}_{\mathrm{cand}}, \left\{\boldsymbol{x}_{\mathrm{ref}}^{(i)}\right\}_{i=1}^{N}, \boldsymbol{x}_{\mathrm{img}}\right\}, \tag{1}$$

where $\boldsymbol{x}_{\mathrm{cand}} \in \{1,0\}^{V\times L}$ represents the candidate, $\{\boldsymbol{x}_{\mathrm{ref}}^{(i)}\} \in \{1,0\}^{N\times V\times L}$ represents the i-th reference out of N, and $\boldsymbol{x}_{\mathrm{img}} \in \mathbb{R}^{3\times H\times W}$ represents the image. Here, V, L, N, H and W denote vocabulary size, maximum token length, number of reference sentences in one sample, and height and width of the image, respectively.

As previously emphasized, selecting an appropriate method for sentence embedding necessitates careful consideration. In this study, we use both the CLIP text encoder and RoBERTa trained with supervised SimCSE to obtain sentence embeddings. Initially, using RoBERTa pretrained with supervised SimCSE, we extract sentence embedding $\boldsymbol{c}_{\mathrm{rb}} \in \mathbb{R}^{L\times d_R}$ and $\{\boldsymbol{r}_{\mathrm{rb}}^{(i)}\}_{i=1}^{N} \in \mathbb{R}^{N\times L\times d_R}$ from $\boldsymbol{x}_{\mathrm{cand}}$ and $\{\boldsymbol{x}_{\mathrm{ref}}^{(i)}\}_{i=1}^{N}$, respectively. Note that d_R represents the output dimension of RoBERTa and the sentence embeddings are derived from the `[CLS]` token of inputs into RoBERTa. Following this, using the CLIP text encoder, we derive language feature $\boldsymbol{c}_{\mathrm{clip}} \in \mathbb{R}^{d_{\mathrm{CLIP}}}$ and $\boldsymbol{r}_{\mathrm{clip}}^{(i)} \in \mathbb{R}^{d_{\mathrm{CLIP}}}$ from $\boldsymbol{x}_{\mathrm{cand}}$ and $\boldsymbol{x}_{\mathrm{ref}}^{(i)}$, respectively, where d_{CLIP} denotes the output dimension of the CLIP encoders. Additionally, utilizing the pretrained CLIP image encoder (ViT-B/16), we obtain image features $\boldsymbol{v} \in \mathbb{R}^{d_{\mathrm{CLIP}}}$ from $\boldsymbol{x}_{\mathrm{img}}$.

3.3.1 Parallel Feature Extraction Mechanism

In alignment with the principles of $\mathrm{M^2LHF}$, we also propose a *parallel feature extraction mechanism*. This mechanism serves as a multimodal extension of the RUSE method[53, 59], employing both difference and Hadamard products. It extracts effective features for caption evaluation by utilizing the difference and Hadamard product of features derived from both CLIP and RoBERTa. Given that CLIP is designed to minimize the cosine similarity between corresponding language and image features, the Hadamard product applied to CLIP features is considered to be effective. Additionally, the difference and Hadamard product operations generate vectors that encapsulate similarity because each element of the vector can be amplified or attenuated in relation to the others.

Initially, given the following inputs:

$$\left\{\boldsymbol{c}_{\text{clip}},\ \boldsymbol{r}^{(i)}_{\text{clip}},\ \boldsymbol{c}_{\text{rb}},\ \boldsymbol{r}^{(i)}_{\text{rb}},\ \boldsymbol{v}\right\}, \tag{2}$$

the proposed framework first computes $\boldsymbol{h}^{(i)}_{\text{inter}}$ as:

$$\boldsymbol{h}^{(i)}_{\text{inter}} = [F(\boldsymbol{c}_{\text{clip}}, \boldsymbol{r}^{(i)}_{\text{clip}});\ F(\boldsymbol{c}_{\text{clip}}, \boldsymbol{v});\ F(\boldsymbol{c}_{\text{rb}}, \boldsymbol{r}^{(i)}_{\text{rb}})]. \tag{3}$$

Here, F is the following function:

$$F(\boldsymbol{c}, \boldsymbol{r}) = [\boldsymbol{c};\ \boldsymbol{r};\ |\boldsymbol{c} - \boldsymbol{r}|;\ \boldsymbol{c} \odot \boldsymbol{r}], \tag{4}$$

where $\odot$ denotes the Hadamard product.

Subsequently, we apply the MLP to $\boldsymbol{h}^{(i)}_{\text{inter}}$ to compute $\boldsymbol{h}^{(i)}$, which effectively captures the similarity among the multimodal features for the i-th reference. Note that we chose MLP because pilot experiments demonstrated its superior performance compared with Transformer [63].

Finally, we compute the evaluation score $\hat{y}$ as:

$$\hat{y} = \underset{i}{\text{Aggregate}}(\sigma(\text{MLP}(\boldsymbol{h}^{(i)}))), \tag{5}$$

where σ denotes the sigmoid function, which scales the output to the range $[0, 1]$. In this context, Aggregate denotes an aggregation function. This function can encompass various operations, such as calculating the maximum or average value. Notably, in our experimental setup, we opted for the max function.

For the loss function, we adopted the mean squared error, which is a standard choice in regression problems because of its effectiveness in quantifying the variance between predicted and human judgments. Our implementation details can be found in Appendix E.

4. Experimental Evaluation

4.1. Setups

Polaris dataset. In this study, we introduce the Polaris dataset, which consists of image-caption pairs and human judgments on the appropriateness of the captions. Training supervised models to predict human judgments benefits significantly from a large-scale corpus that contains diverse captions. However, to the best of our knowledge, there are few open datasets with diverse captions. Therefore, we constructed the Polaris dataset, which contains a total of 131,020 human judgments collected from 550 evaluators.

In comparison with existing datasets, which comprise the inference results of image captioning models and are suitable for metric training, Polaris is distinguished by its exceptional diversity of captions and broader spectrum of evaluations. Our dataset, which comprises over 131K human judgments, significantly surpasses the largest dataset suitable for metric training, which contains only around 12K judgments [1]. Regarding the diversity of evaluations, the Polaris dataset had an average of eight evaluators per caption, which provides a more comprehensive judgment than the Flickr8K dataset, which had an average of only three evaluators per caption. Even the CapEval1K dataset, one of the largest in terms of evaluator participation, only involved five evaluators, which underscores the extensive nature of Polaris with its inclusion of 550 evaluators. Moreover, Polaris differentiates itself from existing datasets such as [1, 10, 27, 35] by being openly accessible, constructed with modern models, and incorporating a wide range of diverse captions.

The Polaris dataset comprises captions generated by the ten standard models: SAT [69], $\mathcal{M}^2$-Transformer [17], VinVL [75], GRIT [62], $\text{BLIP}_{\text{base}}$, $\text{BLIP}_{\text{large}}$ [37], GIT [66], OFA [67], $\text{BLIP-2}_{\text{flan}}$, and $\text{BLIP-2}_{\text{opt}}$ [38]. Within this set, $\text{BLIP}_{\text{base}}$ and $\text{BLIP}_{\text{large}}$ refer to versions of BLIP that employ ViT-B and ViT-L as their image encoders. Similarly, $\text{BLIP-2}_{\text{flan}}$ and $\text{BLIP-2}_{\text{opt}}$ denote versions of BLIP-2 that adopt Flan-T5 [15] and OPT [76] as their Large Language Models (LLMs), respectively. We selected these models because they are standard image captioning models. Additionally, we also chose older models to ensure diversity in the quality of their output sentences. We included inference results for each model, as performed on the MS-COCO [41] and nocaps [7] datasets in the Polaris dataset. We selected MS-COCO because it is the standard dataset for image captioning, whereas we chose nocaps because of its greater diversity of classes compared with MS-COCO.

For a given image, human evaluators assessed the appropriateness of its caption using a five-point scale, taking into account factors such as fluency, relevance, and descriptiveness. We used a crowdsourcing service to collect these evaluations. In the Polaris dataset, we transformed the human judgments, which were rated on a five-point scale, to values in the range $[0, 1]$ using min-max normalization. To eliminate unreliable data, we excluded data from evaluators who exhibited suspicious behavior, such as extremely short response times or consistently providing identical values. The statistical information and details of the Polaris dataset can be found in Appendix B.

Baseline metrics. We adopted BLEU [51], ROUGE [40], METEOR [12], CIDEr [64] and SPICE [10] because they are standard metrics for image captioning tasks. Additionally, we included MoverScore [78], BERTScore [77], BARTScore [74], ViLBERTScore [34], TIGEr [29], LEIC [18], CLIPScore [26], MID [31], UMIC [35] and PAC-S [55] as baseline metrics because they are representative metrics for image captioning.

	Composite	Flickr8K (Expert)	Flickr8K (CF)	Polaris
Classic metrics				
BLEU [51]	30.6	30.8	16.4	46.3
ROUGE [40]	32.4	32.3	19.9	46.3
CIDEr [64]	37.7	43.9	24.6	52.1
METEOR [12]	38.9	41.8	22.2	51.2
SPICE [10]	40.3	44.9	24.4	51.0
SPARCS [20]	43.1	48.1	10.4	43.3
Similarity-based metrics				
MoverScore [78]	30.1	46.7	22.8	46.4
BERTScore [77]	30.1	46.7	22.8	51.6
BARTScore [74]	43.5	37.8	24.3	47.3
LEIC [18]	–	–	29.5	–
TIGEr [29]	45.4	–	–	–
ViLBERTScore [34]	52.4	50.1	–	–
CLIP-S [26]	53.8	51.2	34.4	52.3
RefCLIP-S [26]	55.4	53.0	36.4	52.3
MID [31]	55.7	54.9	37.3	51.3
Learning-based metrics				
PAC-S [55]	55.7	54.3	36.0	52.5
UMIC [35]	56.1	46.8	30.1	49.8
RefPAC-S [55]	<u>57.3</u>	<u>55.9</u>	<u>37.6</u>	<u>56.0</u>
Polos (Ours)	**57.6** (+0.3)	**56.4** (+0.5)	**37.8** (+0.2)	**57.8** (+1.8)

Table 1. Correlation coefficients between various metrics and human judgments. The symbol '–' indicates non-executable code or unavailable data. **Bold** font indicates the highest recorded value and <u>underlining</u> indicates the second-highest value.

Benchmarks. To assess the practicality of a supervised metric, it is essential to evaluate the metric using both in-domain and out-of-domain datasets. Particularly in the context of supervised automatic evaluation, cases exist in which supervised metrics seemingly outperform unsupervised metrics on test sets (in-domain). However, as we demonstrate in the following section, this does not inherently imply better performance on out-of-domain data. Given that supervised metrics are frequently applied to out-of-domain data, those that lack robustness can be impractical. Therefore, evaluating zero-shot performance in supervised models is paramount. In this study, in addition to the Polaris dataset, we used Composite, Flickr8K, PASCAL-50S and FOIL to evaluate zero-shot performance.

4.2. Correlation with Human Judgments

4.2.1 Caption-level Likert judgments

We conducted comparative experiments with baselines across Composite, Flickr8K-Expert, Flickr8K-CF, and the Polaris dataset. The results are presented in Table 1, which details the comparison of Kendall's τ with the baselines across the aforementioned datasets. Following the methodology of previous research [10, 26, 31], we used τ_b (Kendall-B) for the Flickr8K-CF dataset and τ_c (Kendall-C) for all other datasets. Notably, our proposed metric achieved SOTA results with scores of 57.6, 56.4, 37.8, and 57.8 for Composite, Flickr8K-Expert, Flickr8K-CF and the Polaris dataset, respectively. Specifically, our metric outperformed RefPAC-S by margins of 0.3, 0.5, 0.2, and 1.8 points on Composite, Flickr8K-Expert, Flickr8K-CF, and the Polaris dataset, respectively. This indicates that our supervised metric is superior to other supervised metrics such as RefPAC-S. Moreover, we achieved better performance than reference-free metrics such as PAC-S and intermediary metrics (metrics positioned between reference-with-image and reference-free metrics) such as MID. This highlights the critical role of incorporating references. See B.5 for a comparison of Polos and other metrics trained on Polaris.

Fig.3 shows various examples of the proposed metric for the Polaris dataset (further results are provided in Appendix G). First, Fig.3 (a) shows the sample that human evaluators deemed high quality. In this sample, $\boldsymbol{x}_{\mathrm{cand}}$ captured the image content suitably, which resulted in a high human judgment of 0.750. CLIP-S assessed it at 0.402, whereas our proposed method aligned more closely with the human evaluation with a score of 0.745. Fig.3 (b) illustrates samples rated as mediocre by human evaluators because of partial accuracy. For instance, the veracity of the phrase "hanging on a tree" could not be confirmed by the image content, which led to a moderate human score of 0.450. RefPAC-S and RefCLIP-S overestimated the caption with scores of 0.825 and 0.746, respectively, whereas our proposed metric provided a more judicious score of 0.513, thereby reflecting its effectiveness in recognizing "normal" quality captions. Fig.3 (c) shows examples labeled as poor by human evaluators. A discrepancy, such as misidentifying the animal in the reference, resulted in a low human score of 0.071. Again, RefPAC-S and PAC-S assessments overestimated the caption, with scores at 0.903 and 0.856, respectively, whereas Polos assigned a more realistic score of 0.173, effectively capturing the *hallucination* error. These results show that our proposed metric appropriately handled multimodalities and produced evaluation scores close to human judgments.

Fig.3 (d) illustrates a sample for which the proposed metric did not perform as expected. In Fig.3 (d), $\boldsymbol{x}_{\mathrm{ref}}^{(1)}$ and $\boldsymbol{x}_{\mathrm{ref}}^{(2)}$ were described as "A cardboard box with a brown teddy bear and items in it outside on a curb" and "a box full of stuffed animals and other children's items," respectively. Conversely, $\boldsymbol{x}_{\mathrm{cand}}$ was described as "a brown teddy bear sitting in a cardboard box." For this sample, the average of human judgments was $\boldsymbol{x}_{\mathrm{cand}}$ as 0.50. We believe that this was because the box contained various items in addition to the teddy bear. By contrast, our proposed metric evaluated this sample with a score of 0.790, which indicates a notable disparity from human judgment. Similarly, RefPAC-S and RefCLIP-S output scores of 0.847 and 0.751, respectively, which also demonstrates a discrepancy between them and human judgment. Given the failure of our metric and the

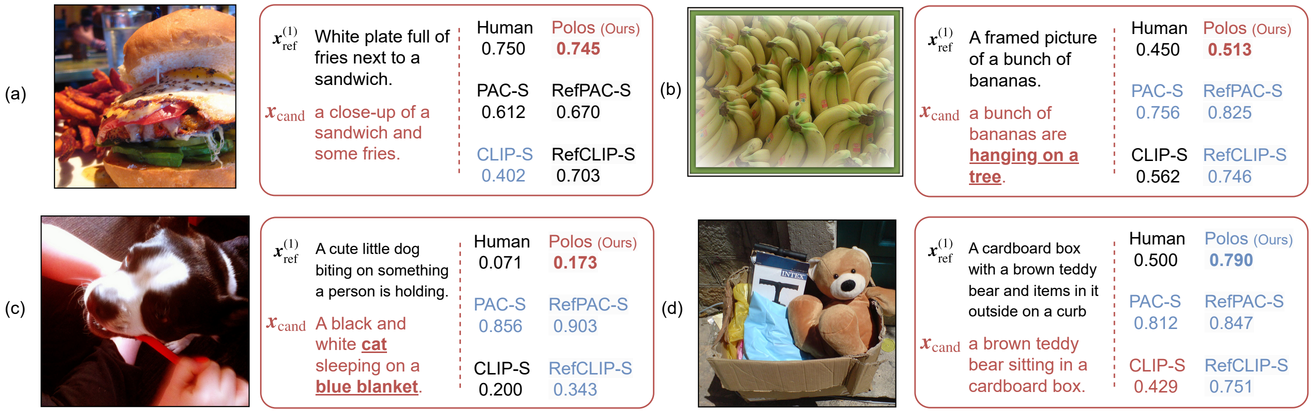

Figure 3. Examples of successful and failed cases from the Polaris dataset. Values in **blue** indicate critical errors, and values in **red** represent those closest to the human judgments. Underlined words indicate significant inaccuracies in the candidate captions. These results demonstrate that our proposed metric effectively handled multimodal inputs and yielded evaluation scores that aligned closely with human judgment. Note that $\boldsymbol{x}_{\text{ref}}^{(1)}$ and $\boldsymbol{x}_{\text{cand}}$ denote one of the reference captions and the candidate caption, respectively.

	HC	HI	HM	MM	Mean
Classic metrics					
BLEU [51]	60.4	90.6	84.9	54.7	72.7
METEOR [12]	63.8	97.7	93.7	65.4	80.2
ROUGE [40]	63.7	95.3	92.3	61.2	78.1
SPICE [10]	63.6	96.3	86.7	68.3	78.7
CIDEr [64]	65.1	98.1	90.5	64.8	79.6
Similarity-based metrics					
ViLBERTScore [34]	49.9	99.6	93.1	75.8	79.6
BERTScore [77]	65.4	98.1	96.4	60.3	80.1
MoverScore [78]	65.1	97.1	93.2	65.6	80.3
TIGEr [29]	56.0	**99.8**	92.8	74.2	80.7
CLIP-S [26]	56.5	99.3	96.4	70.4	80.7
RefCLIP-S [26]	64.5	99.6	95.4	72.8	83.1
MID [31]	67.0	99.7	97.4	76.8	85.2
Learning-based metrics					
PAC-S [55]	60.6	99.3	96.9	72.9	82.4
RefPAC-S [55]	67.7	99.6	96.0	75.6	84.7
UMIC [35]	66.1	**99.8**	**98.1**	76.2	85.1
Polos (Ours)	**70.0** (+3.0)	99.6	97.4	**79.0** (+1.2)	**86.5** (+1.3)

Table 2. Pascal50-S accuracy results (five references).

CLIPScore family, the primary cause of these failures is likely to be an overemphasis on objects that are prominently visible, which results in overlooking the broader context of the image. This is likely to be caused by CLIP's shortcomings in capturing the fine-grained alignment between specific image regions and text, as indicated in [23, 79].

4.2.2 Pairwise ranking on Pascal-50S

Pascal-50S [64] introduced an alternative evaluation framework for accuracy, which comprised 4K pairwise preference judgments between two captions. These judgments encompass four distinct scenarios: pairs of HC (human correct) captions, HI pairs (both human-written, with one incorrect), HM pairs (one from a human and the other generated by a machine), and MM pairs (both generated by machines). The caption that received the majority vote was deemed the preferred choice, with ties resolved randomly. In our study, following the procedure in [26], we calculated the average over five evaluations randomly selected from 48 candidates. Table 2 shows the accuracy results for PASCAL-50S. In our experiments, we achieved SOTA results with accuracies of 70.0%, 79.0%, and 86.5% for HC, MM, and Mean, respectively. The results in Tables 1 and 2 demonstrate that our proposed metric outperformed the baselines in terms of zero-shot performance. These results provide strong evidence that our metric is both robust and practical, thereby serving as a potent automatic evaluation metric for image captioning.

4.2.3 Sensitivity to hallucination

Previous studies [26, 31] measured how evaluation metrics handle *hallucinations* in captions using the FOIL (Find One mismatch between Image and Language caption) dataset [58]. Following the procedure used in [26], we evaluated 32K test images with either one or four references. Subsequently, we computed the accuracy of various metrics to evaluate their effectiveness in consistently awarding higher scores to the true candidate compared with the FOIL dataset.

	1-ref	4-ref
BLEU [51]	66.5	82.6
ROUGE [40]	71.7	79.3
METEOR [12]	78.8	82.6
CIDEr [64]	82.5	90.6
SPICE [10]	75.5	86.1
BARTScore [74]	85.3	91.1
MoverScore [78]	88.4	88.4
BERTScore [77]	88.6	92.1
CLIP-S [26]	87.2	87.2
MID [31]	90.5	90.5
PAC-S [55]	89.9	89.9
RefCLIP-S [26]	91.0	92.6
RefPAC-S [55]	**93.7**	94.9
Polos (Ours)	93.3	**95.4**

Table 3. FOIL hallucination pairwise detection accuracy results.

Table 3 presents the accuracy results for the FOIL dataset. Our proposed method outperformed previous metrics, achieving SOTA results in the 4-ref setting. Specifically, it achieved an accuracy of 93.3% in the 1-ref setting and 95.4% in the 4-ref setting. As previously mentioned, CLIP-S lags behind the traditional metric CIDEr, whereas our method outperformed it in both the 1-ref and 4-ref settings. Moreover, compared with reference-with-image metric RefPAC-S as well as the intermediary metric MID, our approach led by 0.6 and 4.9 points in the 4-ref setting.

4.3. Ablation Study

We conducted three ablation studies to demonstrate the effectiveness of our proposed method. Table 4 presents the results of the ablation studies.

Parallel feature extraction ablation. We investigated the performance of our *parallel feature extraction mechanism* by excluding the Hadamard product and difference, specifically by modifying $F(\mathbf{c}, \mathbf{r})$ (eq.4) to the function $F'(\mathbf{c}, \mathbf{r}) = [\mathbf{c};\ \mathbf{r}]$. The correlation coefficients between Metric (i) and human judgments were lower by 18.3, 15.4, and 6.4 points compared to Metric (vi) for the Composite, Flickr8K, and Polaris, respectively. These indicate that our *parallel feature extraction mechanism* provided superior performance.

M^2LHF ablation. We investigated the performance of $\mathrm{M}^2\mathrm{LHF}$ by modifying our model to make predictions based solely on text, excluding any of the following: $\boldsymbol{x}_{\mathrm{img}}$, $(\boldsymbol{c}_{\mathrm{clip}}, \{\boldsymbol{r}_{\mathrm{clip}}^{(i)}\}_{i=1}^{N})$, or $(\boldsymbol{c}_{\mathrm{rb}}, \{\boldsymbol{r}_{\mathrm{rb}}^{(i)}\}_{i=1}^{N})$. Initially, by omitting $\boldsymbol{x}_{\mathrm{img}}$, we assessed the significance of the image feature. Compared to Metric (vi), the correlation coefficients between Metric (ii) and human judgments were lower by 0.4, 1.4, and 0.7 points on Composite, Flickr8K, and Polaris respectively. These results suggest that the image feature played a pivotal role in enhancing the performance of our proposed metric. Subsequently, we evaluated the contribution of each module by excluding either CLIP or RoBERTa. Relative to Metric (vi), the correlation coefficients between Metric (iii) and human judgments decreased by 2.6, 3.2, and 2.4 points. Similarly, the exclusion of RoBERTa led to a decrease in performance. Although we observed that RoBERTa pretrained by SimCSE was found to enhance performance, these results highlighted CLIP as the most influential feature extractor. Overall, these results validated the efficacy of the image feature and each module, underscoring that the introduction of $\mathrm{M}^2\mathrm{LHF}$ contributed to the performance improvement of the proposed metric.

Aggregation mechanism ablation. We investigated the impact on performance by setting the $\mathrm{Aggregate}$ function to either Max or Mean. The correlation coefficients between Metric (v) and human judgments were lower by 2.5, 1.0, and 5.7 points compared with Metric (vi) for the Composite, Flickr8K, and the Polaris dataset, respectively.

Metric	P	$\boldsymbol{x}_{\mathrm{img}}$	CLIP	RoBERTa	Aggregate	Composite	Flickr8K	Polaris
(i)		✓	✓	✓	Max	39.3	41.0	51.4
(ii)	✓		✓	✓	Max	56.8	55.5	57.1
(iii)	✓			✓	Max	55.0	53.2	55.4
(iv)	✓	✓	✓		Max	56.0	55.0	57.4
(v)	✓	✓	✓	✓	Mean	55.1	55.4	52.1
(vi)	✓	✓	✓	✓	Max	**57.6**	**56.4**	**57.8**

Table 4. Ablation study results: performance of $\mathrm{M}^2\mathrm{LHF}$ and the *parallel feature extraction* and comparison of the effects of using either the Max or Mean within the $\mathrm{Aggregate}$ function. 'P' indicates the use of the *parallel feature extraction mechanism*

These results indicate that using the Max function as the $\mathrm{Aggregate}$ function provided superior performance.

4.4. Discussion and Limitations

Although our metric generated compelling results, our approach has some limitations. As shown in Fig.3, our metric demonstrated a tendency to overestimate captions that lacked intricate details. This phenomenon is likely to be attributed to its excessive focus on prominently visible objects; hence it consequently overlooked the broader context within the image. As discussed in Section 4.2.1, this limitation may be attributed to CLIP's inherent limitations in capturing the fine-grained alignment between specific image regions and their corresponding textual descriptions. Despite this, we firmly believe that this study represents a significant stride toward the development of a more practical metric for image captioning models. In future work, we plan to extend our metric by enhancing the fine-grained alignment, drawing inspiration from methods such as RegionCLIP [79].

5. Conclusion

In this paper, we introduced Polos, an automatic evaluation metric for image captioning. The contribution of this paper is fourfold: i) the introduction of $\mathrm{M}^2\mathrm{LHF}$, a novel framework used to develop a practical metric for image captioning; ii) the introduction of a *parallel feature extraction mechanism* that leverages CLIP and RoBERTa pretrained with SimCSE; iii) the construction of the Polaris dataset, containing a total of 131,020 human judgments collected from 550 evaluators; and iv) achieving SOTA performance on image captioning benchmarks including Composite, Flickr8K-Expert, and Flickr8K-CF, PASCAL-50S, FOIL, and the Polaris dataset.

Acknowledgments

This work was supported by a grant from Apple, Inc. Any views, opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and should not be interpreted as reflecting the views, policies or position, either expressed or implied, of Apple, Inc. This work was also partially supported by JSPS KAKENHI Grant Number 23H03478, JST CREST, and NEDO.

References

[1] Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, et al. From Images to Sentences through Scene Description Graphs using Commonsense Reasoning and Knowledge. *arXiv preprint arXiv:1511.03292*, 2015. 2, 3, 5

[2] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, and Mona Diab. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In *SemEval*, pages 81–91, 2014. 4

[3] Eneko Agirre, Carmen Banea, et al. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In *SemEval*, pages 252–263, 2015.

[4] Eneko Agirre, Carmen Banea, et al. SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. In *SemEval*, pages 497–511, 2016.

[5] Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Agirre. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In *SemEval*, pages 385–393, 2012.

[6] Eneko Agirre, Daniel Cer, Mona Diab, Aitor Agirre, and Weiwei Guo. SEM 2013 Shared Task: Semantic Textual Similarity. In *SEM*, pages 32–43, 2013. 4

[7] Harsh Agrawal, Karan Desai, et al. nocaps: Novel Object Captioning at Scale. In *ICCV*, pages 8948–8957, 2019. 3, 5

[8] Hiba Ahsan, Daivat Bhatt, Kaivan Shah, and Nikita Bhalla. Multi-Modal Image Captioning for the Visually Impaired. In *NAACL-HLT*, pages 53–60, 2021. 1

[9] Jean Alayrac, Jeff Donahue, Pauline Luc, et al. Flamingo: A Visual Language Model for Few-shot Learning. In *NeurIPS*, volume 35, pages 23716–23736, 2022. 1

[10] Peter Anderson, Basura Fernando, Mark Johnson, et al. SPICE: Semantic Propositional Image Caption Evaluation. In *ECCV*, pages 382–398, 2016. 1, 2, 5, 6, 7, 3

[11] Peter Anderson, Xiaodong He, et al. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *CVPR*, pages 6077–6086, 2018. 3

[12] Satanjeev Banerjee et al. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *ACL*, pages 65–72, 2005. 1, 2, 5, 6, 7

[13] Daniel Cer, Mona Diab, et al. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *SemEval*, pages 1–14, 2017. 4

[14] Yen Chen, Linjie Li, Licheng Yu, Ahmed Kholy, Faisal, Zhe Gan, Yu Cheng, et al. UNITER: Universal Image-text Representation Learning. In *ECCV*, pages 104–120, 2020. 3

[15] Hyung Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, et al. Scaling Instruction-finetuned Language Models. *arXiv preprint arXiv:2210.11416*, 2022. 5

[16] Alexis Conneau, Kartikay Khandelwal, et al. Unsupervised Cross-lingual Representation Learning at Scale. In *ACL*, pages 8440–8451, 2020. 1

[17] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, et al. Meshed-Memory Transformer for Image Captioning. In *CVPR*, pages 10578–10587, 2020. 3, 5, 1, 2

[18] Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. Learning to Evaluate Image Captioning. In *CVPR*, pages 5804–5812, 2018. 5, 6

[19] Jacob Devlin, Ming-Wei Chang, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, pages 4171–4186, 2019. 1

[20] Joshua Feinglass and Yezhou Yang. SMURF: SeMantic and linguistic UndeRstanding Fusion for Caption Evaluation via Typicality Analysis. In *IJCNLP*, pages 2250–2260, 2021. 6

[21] Adam Fisch, Kenton Lee, Ming Chang, Jonathan Clark, and Regina Barzilay. CapWAP: Image Captioning with a Purpose. In *EMNLP*, pages 8755–8768, 2020. 1

[22] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *EMNLP*, pages 6894–6910, 2021. 2, 4

[23] Xiuye Gu, Tsung Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. In *ICLR*, 2022. 3, 7

[24] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning Images Taken by People Who Are Blind. In *ECCV*, pages 417–434, 2020. 1

[25] Simao Herdade, Armin Kappeler, Kofi Boakye, et al. Image Captioning: Transforming Objects into Words. In *NeurIPS*, volume 32, pages 11137–11147, 2019. 1

[26] Jack Hessel, Ari Holtzman, et al. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *EMNLP*, pages 7514–7528, 2021. 1, 2, 3, 4, 5, 6, 7

[27] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *JAIR*, 47:853–899, 2013. 1, 2, 3, 5

[28] Lun Huang, Wenmin Wang, Jie Chen, and Xiao Wei. Attention on Attention for Image Captioning. In *ICCV*, pages 4634–4643, 2019. 3

[29] Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, et al. TIGEr: Text-to-image Grounding For Image Caption Evaluation. In *EMNLP*, 2019. 2, 5, 6, 7

[30] Motonari Kambara et al. Case Relation Transformer: A Crossmodal Language Generation Model for Fetching Instructions. *IEEE RAL*, 6:8371–8378, 2021. 1

[31] Jin Kim et al. Mutual Information Divergence: A Unified Metric for Multimodal Generative Models. In *NeurIPS*, volume 35, pages 35072–35086, 2022. 1, 2, 5, 6, 7

[32] Yongil Kim, Yerin Hwang, Hyeongu Yun, Seunghyun Yoon, et al. PR-MCS: Perturbation Robust Metric for MultiLingual Image Captioning. In *EMNLP*, pages 12237–12258, 2023. 3

[33] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From Word Embeddings To Document Distances. *PMLR*, 37:957–966, 2015. 1

[34] Hwanhee Lee, Seunghyun Yoon, et al. ViLBERTScore: Evaluating Image Caption Using Vision-and-Language BERT. In *Eval4NLP*, pages 34–39. 1, 2, 5, 6, 7

[35] Hwanhee Lee, Seunghyun Yoon, et al. UMIC: An Unreferenced Metric for Image Captioning via Contrastive Learning. In *ACL*, pages 220–226, 2021. 1, 2, 3, 4, 5, 6, 7

[36] Tomer Levinboim, Ashish V. Thapliyal, Piyush Sharma, and Radu Soricut. Quality Estimation for Image Captions Based on Large-scale Human Evaluations. In *NAACL*, pages 3157–3166, 2021. 3

[37] Junnan Li et al. BLIP: Bootstrapping Language-image Pre-training for Unified Vision-language Understanding and Generation. In *ICML*, pages 12888–12900, 2022. 3, 5, 2

[38] Junnan Li, Dongxu Li, et al. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *ICML*, 2023. 3, 5, 1, 2

[39] Jingyu Li, Zhendong Mao, Shancheng Fang, et al. ER-SAN: Enhanced-Adaptive Relation Self-Attention Network for Image Captioning. In *IJCAI*, pages 1081–1087, 2022. 1

[40] Chin Lin. ROUGE: A Package For Automatic Evaluation Of Summaries. In *ACL*, pages 74–81, 2004. 1, 2, 5, 6, 7
[41] Tsung Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, et al. Microsoft COCO: Common Objects in Context. In *ECCV*, pages 740–755, 2014. 3, 5
[42] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *NeurIPS*, 2023. 3
[43] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019. 2
[44] Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, et al. Dual-Level Collaborative Transformer for Image Captioning. In *AAAI*, volume 35, pages 2286–2293, 2021. 1
[45] Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. LENS: A learnable evaluation metric for text simplification. In *ACL*, pages 16383–16408, July 2023. 2
[46] Aly Magassouba, Komei Sugiura, and Hisashi Kawai. Multimodal Attention Branch Network for Perspective-Free Sentence Generation. In *CoRL*, pages 76–85, 2019. 1
[47] Marco Marelli, Stefano Menini, Marco Baroni, et al. A SICK Cure for the Evaluation of Compositional Distributional Semantic Models. In *LREC*, pages 216–223, 2014. 4
[48] Yue Ming, Nannan Hu, Chunxiao Fan, Jiangwan Zhou, and Hui Yu. Visuals to Text: A Comprehensive Review on Automatic Image Captioning. *JAS*, 9(8):1339–1365, 2022. 1
[49] Gabriel Oliveira, Esther Colombini, and Sandra Avila. CIDEr-R: Robust Consensus-based Image Description Evaluation. In *W-NUT*, pages 351–360, 2021. 2
[50] OpenAI. GPT-4V (ision) System Card. 2023. 3
[51] Kishore Papineni, Salim Roukos, Todd Ward, and Wei Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *ACL*, pages 311–318, 2002. 1, 2, 5, 6, 7
[52] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning Transferable Visual Models from Natural Language Supervision. In *ICML*, pages 8748–8763, 2021. 1, 2
[53] Ricardo Rei, Craig Stewart, Ana Farinha, and Alon Lavie. COMET: A Neural Framework for MT Evaluation. In *EMNLP*, pages 2685–2702, 2020. 2, 4
[54] Steven Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical Sequence Training for Image Captioning. In *CVPR*, pages 7008–7024, 2017. 3
[55] Sara Sarto et al. Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation. In *CVPR*, pages 6914–6924, 2023. 1, 2, 4, 5, 6, 7, 3
[56] Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning Robust Metrics for Text Generation. In *ACL*, pages 7881–7892, 2020. 2, 4
[57] Piyush Sharma et al. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset for Automatic Image Captioning. In *ACL*, pages 2556–2565, 2018. 3
[58] Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, et al. FOIL it! Find One Mismatch Between Image and Language caption. In *ACL*, pages 255–265, 2017. 7
[59] Hiroki Shimanaka et al. RUSE: Regressor Using Sentence Embeddings for Automatic Machine Translation Evaluation. In *WMT18*, pages 751–758, 2018. 2, 4
[60] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, et al. TextCaps: A Dataset for Image Captioning with Reading Comprehension. In *ECCV*, pages 742–758, 2020. 1
[61] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, et al. From Show to Tell: A Survey on Deep Learning-based Image Captioning. *PAMI*, 45(1):539–559, 2022. 1
[62] Masanori Suganuma, Takayuki Okatani, et al. GRIT: Faster and Better Image Captioning Transformer Using Dual Visual Features. In *ECCV*, pages 167–184, 2022. 3, 5, 1, 2
[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, et al. Attention Is All You Need. In *NIPS*, volume 30, pages 5998–6008, 2017. 5
[64] Ramakrishna Vedantam, Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based Image Description Evaluation. In *CVPR*, pages 4566–4575, 2015. 1, 2, 3, 5, 6, 7
[65] Yuiga Wada, Kanta Kaneda, et al. JaSPICE: Automatic Evaluation Metric Using Predicate-Argument Structures for Image Captioning Models. In *CoNLL*, 2023. 2
[66] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, et al. GIT: A Generative Image-to-text Transformer for Vision and Language. *TMLR*, 2022. 3, 5, 2
[67] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, et al. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-sequence Learning Framework. In *ICML*, pages 23318–23340, 2022. 3, 5, 2
[68] Julia White, Gabriel Poesia, Robert Hawkins, et al. Open-domain Clarification Question Generation Without Question Examples. In *EMNLP*, pages 563–570, 2021. 1
[69] Kelvin Xu, Jimmy Ba, Ryan Kiros, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *ICML*, pages 2048–2057, 2015. 3, 5, 1, 2
[70] Wenda Xu, Xian Qian, Mingxuan Wang, et al. SESCORE2: Learning Text Generation Evaluation via Synthesizing Realistic Mistakes. In *ACL*, pages 5166–5183, 2023. 2
[71] Wenda Xu, Yi-Lin Tuan, Yujie Lu, et al. Not All Errors are Equal: Learning Text Generation Metrics using Stratified Error Synthesis. In *EMNLP*, pages 6559–6574, 2022.
[72] Wenda Xu, Danqing Wang, et al. INSTRUCTSCORE: Towards Explainable Text Generation Evaluation with Automatic Feedback. In *EMNLP*, pages 5967–5994, 2023. 2
[73] Peter Young et al. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions. *TACL*, 2:67–78, 2014. 3
[74] Weizhe Yuan, Graham Neubig, et al. BARTScore: Evaluating Generated Text as Text Generation. In *NeurIPS*, volume 34, pages 27263–27277, 2021. 3, 5, 6, 7
[75] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, et al. VinVL: Revisiting Visual Representations in Vision-language Models. In *CVPR*, pages 5579–5588, 2021. 3, 5, 2
[76] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, et al. OPT: Open Pre-trained Transformer Language Models. *arXiv preprint arXiv:2205.01068*, 2022. 5
[77] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In *ICLR*, 2020. 1, 2, 3, 5, 6, 7
[78] Wei Zhao et al. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. In *EMNLP-IJCNLP*, pages 563–578, 2019. 1, 2, 5, 6, 7
[79] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, et al. RegionCLIP: Region-based Language-image Pretraining. In *CVPR*, pages 16793–16803, 2022. 3, 7, 8, 2