

# Polos: Multimodal Metric Learning from Human Feedback for Image Captioning

## Supplementary Material

### A. Additional Related Work

**Image captioning** Numerous studies have been conducted in the domain of image captioning [17, 25, 39, 44, 69]. For instance, GRIT [62] adeptly leverages both grid and region-based features for enhanced image captioning, thereby eliminating the need for conventional CNN-based detectors. BLIP-2 [38] is a novel pre-training approach that efficiently uses LLMs that outperforms models including Flamingo [9] with notably fewer trainable parameters. This field includes applications in various domains, such as aiding persons with vision impairment [8, 24, 60] and robotics [30, 46]. Survey papers such as [48, 61] offer a comprehensive overview of image caption generation, including models, standard datasets, and evaluation metrics. Specifically, they provide a comprehensive summary of various automatic evaluation metrics, including similarity-based and learning-based metrics [33, 77, 78].

### B. Polaris Dataset

#### B.1. Meta-Analysis

As pointed out in [35], there are issues with utilizing existing datasets such as Flickr8K and Composite for training purposes. Fig. 4 shows the score distributions of human judgments in Composite, Flickr8K-Expert, Flickr8K-CF, and our proposed Polaris dataset. For the Flickr8K dataset, the majority of scores fall below 0.4, as the candidate captions were sourced from a reference caption pool through an image retrieval system. Moreover, the Flickr8K datasets do not contain captions generated by models, which presents an issue from the perspective of the domain mismatch because our aim is to build an automatic metric for image captioning. Consequently, we argue that Flickr8K-Expert and Flickr8K-CF are not suitable for training metrics. Furthermore, the human judgments in Flickr8K-CF were provided using a binary scheme, that only allowed responses categorized as “yes” or “no.” This method is problematic because of its lack of granularity and its propensity to force evaluators into making overly simplistic judgments. For instance, a disparity exists in quality for captions that describe content and our method may not be able to adequately evaluate captions of varying quality.

In the case of the Composite dataset, we note its exceptionally few human judgments. This paucity of data renders it inadequate for developing a practical metric. Moreover, as pointed out in [35], each sample’s score was determined by a single annotator, leading to potentially biased outputs. Upon manual examination of the captions, [35] also pointed out that these captions are often coarsely generated.

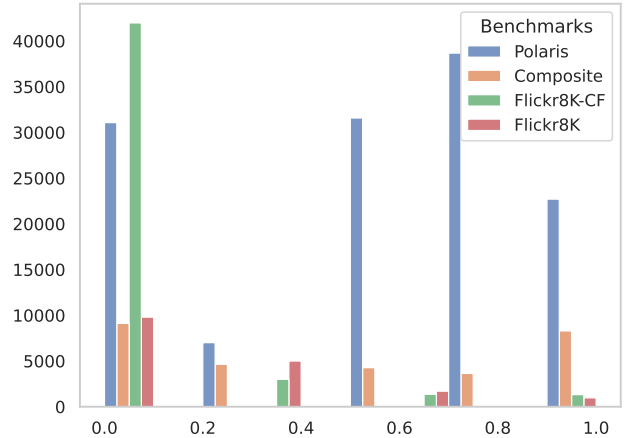


Figure 4. Score distributions of human judgments in **Composite**, **Flickr8K-Expert**, **Flickr8K-CF**, and our **Polaris** dataset. All scores were normalized from 0 to 1. Polaris distinguishes itself from other datasets by encompassing a vast collection of captions and integrating a broad spectrum of human judgments.

#### B.2. Statistics and Details

The Polaris dataset includes 13,691 images accompanied by 131,020 generated captions. Additionally, it contains 262,040 references. All sentences are in English. To minimize biases in evaluations and achieve more balanced judgments compared with other datasets, we engaged multiple human evaluators to evaluate each caption. Specifically, each generated caption was evaluated by approximately eight different evaluators. The generated captions encompass a vocabulary of 3,154 unique words, with a total of 1,177,512 words. On average, each caption is composed of 8.99 words. By contrast, the reference captions have a vocabulary of 22,275 unique words, with a word count of 8,309,300. Each reference caption, on average, consists of 10.7 words.

In the Polaris dataset, the training, validation, and test sets consist of 78,631, 26,269, and 26,123 samples, respectively. We used the training set to train the model, the validation set for hyperparameter tuning, and the test set to evaluate the performance of the model.

#### B.3. Image Captioning Models

The Polaris dataset comprises captions generated by the following 10 standard models. Table 5 summarizes these image captioning models. We selected these models as they are standard image captioning models. Additionally, we also chose older models to ensure diversity in the quality of their output sentences.

Year	Venue
BLIP-2 <sub>flan</sub> [38]	ICML'23
BLIP-2 <sub>opt</sub> [38]	ICML'23
GRIT [62]	ECCV'22
OFA [67]	ICML'22
GIT [66]	TMLR'22
BLIP <sub>large</sub> [37]	ICML'22
BLIP <sub>base</sub> [37]	ICML'22
VinVL [75]	CVPR'21
$\mathcal{M}^2$ -Transformer [17]	CVPR'20
SAT[69]	CVPR'15

Table 5. Image captioning models used for the Polaris dataset.

## B.4. Annotation Process

Prior to the evaluation, we provided the human evaluators with three sample images to familiarize them with the evaluation method. Fig. 5 shows the user interface of our annotation tool and example data with the instructions. 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.

## B.5. Training on Polaris

Table 6 shows the comparative results of the learning-based metrics trained on the Polaris dataset. The result demonstrates that Polos outperforms these metrics even when trained on the Polaris dataset.

Metrics	Trained on Polaris	Composite	Flickr8K (Expert)	Flickr8K (CF)	Polaris
PAC-S		55.7	54.3	36.0	52.5
PAC-S	✓	56.1	54.7	36.2	53.3
RefPAC-S		57.3	55.9	37.6	56.0
RefPAC-S	✓	57.4	56.0	37.4	56.9
Polos	✓	<b>57.6</b>	<b>56.4</b>	<b>37.8</b>	<b>57.8</b>

Table 6. Results of learning-based metrics trained on Polaris.

## C. Gameability in Image Captioning

Since some studies [62] have utilized CIDEr for reinforcement learning, we believe Polos could also be used for improving image caption models. However, as highlighted in [10, 64], reliance on a single metric carries the risk of ‘gaming’ the system. Therefore, when employing Polos for reinforcement learning, it is imperative to exercise caution to avoid such pitfalls.

## D. Error Analysis

To investigate the limitations of the proposed metric, we analyzed 100 instances where the method did not perform as expected. We define failed cases as samples that satisfy

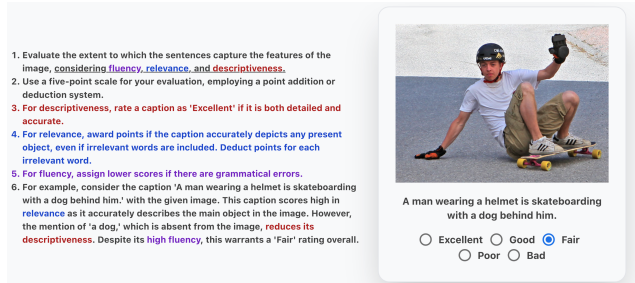


Figure 5. Annotation interface and instructions for captioning evaluation task. Human evaluators assessed the appropriateness of captions for a given image using a five-point scale, with criteria including fluency, relevance, and descriptiveness.

the condition  $|y - \hat{y}| \geq \theta$ . In this study, we held  $\theta$  at a value of 0.25, corresponding to the step size when normalizing a five-level evaluation. Table 7 categorizes the failed cases. The causes of failure can be grouped into seven categories:

- (i) Overestimation of captions lacking details (OCLD): This category pertains to instances where the proposed metric assigned higher scores to captions that lacked vital details, missing critical aspects of the images.
- (ii) Overestimation of captions with incorrect details (OCID): This category refers to instances where the proposed metric inaccurately assigned higher scores to captions containing incorrect or misleading details.
- (iii) Underestimation of captions where the focus areas differ from the references (UCFA): This category refers to instances where the proposed metric assigned lower scores to captions that, although accurate, focused on areas different from the references.
- (iv) Serious Errors (SE): This category encompasses instances where the evaluation deviated greatly from human judgments, being much higher or lower.
- (v) Overestimation of captions with grammatical inaccuracies (OSGI): This category refers to instances where the proposed metric erroneously assigned higher scores to captions that, while potentially accurate in content, contained grammatical errors.
- (vi) Annotation errors (AE): This category pertains to instances where human evaluations proved to be inaccurate. These evaluations were either higher or lower than what could be reasonably expected.
- (vii) Others: This category encompasses miscellaneous errors that do not fit into the aforementioned categories.

From Table 7, it can be inferred that the main bottleneck was the overestimation of captions that lacks detail. As mentioned in Section 4.4, a possible solution could be to enhance the fine-grained alignment [79].

Errors	Description	#Error
OCLD	Overestimation of captions lacking details	29
OCID	Overestimation of captions with incorrect details	22
UCFA	Underestimation of captions where the focus area differs from the reference	15
SE	Serious errors (e.g., assigning a higher score to captions with major mistakes)	11
OGI	Overestimation of captions with grammatical inaccuracies	11
AE	Annotation errors in human judgments	9
Others	Miscellaneous or less common errors	3
Total	—	100

Table 7. Categorization of failed samples.

## E. Implementation Details

Table 8 shows the experimental settings for the proposed method. We trained our model on a Tesla A100 GPU, and the training time was approximately 4.6 hours. To measure the inference time, we tested our metric on a system equipped with a GeForce RTX 3090 and an Intel Core i9-10900KF. The inference times per sample for SPICE [10], RefPAC-S [55], and Polos were 16.6 ms, 4.45 ms, and 6.91 ms, respectively. Notably, Polos operates at a speed 2.4 times faster than SPICE. Furthermore, in the 6.91 ms processing time of Polos, CLIP took 4.90 ms, RoBERTa 1.96 ms, and MLP just 0.054 ms. We used early stopping in our model to optimize for the highest Kendall’s  $\tau$ . At each epoch, we evaluated Kendall’s  $\tau$  on the validation set. If no improvement was observed over five consecutive epochs, we stopped training. Subsequently, we evaluated the model’s performance using the test set, referring to the epoch where the validation set achieved its best  $\tau$  value.

Table 8. Experimental settings for the proposed method.

Batch size	64
Optimizer	Adam ( $\beta_1 = 0.9, \beta_2 = 0.98$ )
Learning Rate	$3.0 \times 10^{-5}$

## F. Evaluation of LLMs

To showcase the ability to evaluate LLaVA [42] and GPT-4V [50] captions, we used both to generate captions for Fig.3 (a). Fig.6 shows LLaVA’s output, which is fluent and descriptive but includes *hallucinations* and less relevant texts (underlined texts), and received a human rating of 0.75. The scores from PAC-S, RefPAC-S, and Polos were 0.93, 0.62, and 0.69, respectively, with Polos most closely matching human judgment. This result suggests the efficacy of Polos for long captions generated by LLaVA.

Fig.7 shows the GPT-4V generated caption. Similarly, GPT-4V’s caption, which also contained *hallucinations* and

This image shows a delicious-looking sandwich accompanied by a side of fries. The sandwich appears to be a burger, featuring a beef patty topped with a fried egg, tomato, lettuce, and cheese. The egg on top of the burger is a popular choice for adding a unique twist to the traditional burger, providing a creamy texture and a burst of flavor. The fries are served in a basket, which is a common way to present them in a casual dining setting. The basket helps to keep the fries contained and prevents them from spilling onto the table.

Figure 6. Caption generated by LLaVA. Underlined words indicate *hallucinations* or less relevant texts. This caption received a human rating of 0.75.

The image displays a gourmet burger accompanied by a side of sweet potato fries. The burger is constructed with multiple layers, starting with a soft-looking, toasted sesame seed bun on the bottom. It is followed by a layer of crisp green lettuce and a well-seared beef patty that has visible grill marks, hinting at a smoky flavor. The burger is crowned with a fried egg, which appears to have a soft yolk. A wooden skewer holds the burger together. In the blurred background, there’s a cold glass of water, indicating a casual dining atmosphere.

Figure 7. Caption generated by GPT-4V. This sample contains *hallucinations* and less relevant texts and was assigned a human rating of 0.625.

less relevant texts, was assigned a human rating of 0.625. The scores from PAC-S, RefPAC-S, and Polos were 0.89, 0.84, and 0.65, respectively. In this sample, Polos also demonstrated a close alignment with human judgment, suggesting its efficacy for evaluating GPT-4V generated captions.

## G. Additional Qualitative Results

Fig.8 and Fig.9 provide additional comparisons between the CLIPScore family and Polos. We observed that the CLIPScore family tends to overestimate scores. Specifically, RefCLIP-S and RefPAC-S may not effectively compare references and a candidate. Although CLIP-S could not show overestimation, this does not imply adequacy in caption evaluation. Rather, it may signal estimation deficiencies, especially for longer captions, stemming from poor alignment between words and images, as CLIP relies heavily on the alignment between image and language features.

	$x_{\text{ref}}^{(1)}$ : A crying woman looking at herself in a mirror. $x_{\text{cand}}$ : <u>a woman holding a cup with a candle in a</u>	CLIP-S 0.39 PAC-S 0.81	RefCLIP-S 0.69 RefPAC-S 0.87	<b>Polos</b> <b>0.18</b>	<b>Human</b> <b>0.18</b>
	$x_{\text{ref}}^{(1)}$ : a upside down boat is on top of a big hil $x_{\text{cand}}$ : a couple of boats that are sitting in the snow	CLIP-S 0.5 PAC-S 0.83	RefCLIP-S 0.87 RefPAC-S 0.86	<b>Polos</b> <b>0.64</b>	<b>Human</b> <b>0.6</b>
	$x_{\text{ref}}^{(1)}$ : Pink donut with white sprinkles on the top of it. $x_{\text{cand}}$ : a blue umbrella with a blue and white design	CLIP-S 0.26 PAC-S 0.7	RefCLIP-S 0.46 RefPAC-S 0.82	<b>Polos</b> <b>0.083</b>	<b>Human</b> <b>0.031</b>
	$x_{\text{ref}}^{(1)}$ : A white car is parked in the street at night time. $x_{\text{cand}}$ : a blurry image of a city street at night	CLIP-S 0.49 PAC-S 0.89	RefCLIP-S 0.86 RefPAC-S 0.88	<b>Polos</b> <b>0.77</b>	<b>Human</b> <b>0.61</b>
	$x_{\text{ref}}^{(1)}$ : a bicycle with a basket and a blue and pink umbrella $x_{\text{cand}}$ : a person holding an umbrella in a <u>room</u>	CLIP-S 0.42 PAC-S 0.9	RefCLIP-S 0.73 RefPAC-S 0.92	<b>Polos</b> <b>0.35</b>	<b>Human</b> <b>0.34</b>
	$x_{\text{ref}}^{(1)}$ : A cat is jumping off of a stack of suitcases. $x_{\text{cand}}$ : a <u>blue</u> suitcase is sitting on a <u>blue</u> couch	CLIP-S 0.35 PAC-S 0.87	RefCLIP-S 0.61 RefPAC-S 0.92	<b>Polos</b> <b>0.2</b>	<b>Human</b> <b>0.25</b>

Figure 8. Additional examples from the Polaris dataset (the blue blocks indicate critical errors and the underlined words represent noteworthy features.) The CLIPScore family tends to overestimate scores. Specifically, reference-with-image metrics such as RefCLIP-S and RefPAC-S may not effectively compare references and a candidate. CLIP-S does not exhibit a tendency to overestimate; however, this does not necessarily imply that it estimates captions adequately. Rather, it may indicate a deficiency in its estimation capabilities, particularly for longer captions. This limitation likely stems from poor alignment between words and images in extended captions, as CLIP heavily relies on the alignment between image and language features.



	$x_{\text{ref}}^{(1)}$ : A cat standing on top of a car trunk next to a parked motorcycle. $x_{\text{cand}}$ : a <u>dog</u> is sitting on a motorcycle seat	CLIP-S 0.42	RefCLIP-S 0.74	Polos 0.36	Human 0.31
		PAC-S 0.88	RefPAC-S 0.93		
	$x_{\text{ref}}^{(1)}$ : A man standing next to a dog on the ground. $x_{\text{cand}}$ : a dog is standing in a <u>wooden bench</u>	CLIP-S 0.39	RefCLIP-S 0.68	Polos 0.31	Human 0.25
		PAC-S 0.8	RefPAC-S 0.85		
	$x_{\text{ref}}^{(1)}$ : A young boy reaching up to grab a red apple. $x_{\text{cand}}$ : a box of apples	CLIP-S 0.41	RefCLIP-S 0.72	Polos 0.51	Human 0.47
		PAC-S 0.73	RefPAC-S 0.78		
	$x_{\text{ref}}^{(1)}$ : A banana peel made to look like a <u>clothing zipper</u> . $x_{\text{cand}}$ : a couple of bananas sitting on top of a napkin.	CLIP-S 0.41	RefCLIP-S 0.72	Polos 0.48	Human 0.42
		PAC-S 0.78	RefPAC-S 0.88		
	$x_{\text{ref}}^{(1)}$ : a white stove with a pizza on a pan, a <u>coffee pot</u> and a <u>hot water pot</u> $x_{\text{cand}}$ : a stove top oven with a pizza sitting on top of it.	CLIP-S 0.5	RefCLIP-S 0.88	Polos 0.76	Human 0.61
		PAC-S 0.81	RefPAC-S 0.86		
	$x_{\text{ref}}^{(1)}$ : A girl in white shirt painting a black umbrella. $x_{\text{cand}}$ : a girl in a room with a chair and a book shelf.	CLIP-S 0.36	RefCLIP-S 0.63	Polos 0.44	Human 0.38
		PAC-S 0.81	RefPAC-S 0.83		

Figure 9. Additional examples from the Polaris dataset. These are visualized in the same manner as in Fig.8.