Image Captioning Evaluation in the Age of Multimodal LLMs: Challenges and Future Perspectives

Sara Sarto[1] , **Marcella Cornia**[1] , **Rita Cucchiara**[1,2]
[1]University of Modena and Reggio Emilia, Italy
[2]IIT-CNR, Italy
{sara.sarto, marcella.cornia, rita.cucchiara}@unimore.it

Abstract

The evaluation of machine-generated captions is a complex and evolving challenge. With the advent of Multimodal Large Language Models (MLLMs), image captioning has become a core task, increasing the need for robust and reliable evaluation metrics. This survey provides a comprehensive overview of advancements in image captioning evaluation, analyzing the evolution, strengths, and limitations of existing metrics. We assess these metrics across multiple dimensions, including correlation with human judgment, ranking accuracy, and sensitivity to hallucinations. Additionally, we explore the challenges posed by the longer and more detailed captions generated by MLLMs and examine the adaptability of current metrics to these stylistic variations. Our analysis highlights some limitations of standard evaluation approaches and suggests promising directions for future research in image captioning assessment. For a comprehensive overview of captioning evaluation refer to our project page available at https://github.com/aimagelab/awesome-captioning-evaluation.

1 Introduction

Image captioning, the task of generating natural language descriptions of visual content [Stefanini *et al.*, 2022], is a fundamental component of modern Artificial Intelligence systems such as Multimodal Large Language Models (MLLMs) and autonomous agents. These models, indeed, need to understand and interact with visual data effectively and be capable of generating high-quality descriptions. As MLLMs continue to advance, the development of robust and reliable evaluation metrics for image descriptions becomes essential.

The history of image description has been marked by continuous innovation. Early captioning models evolved from template-based and recurrent neural networks [Vinyals *et al.*, 2015] to Transformer-based architectures [Cornia *et al.*, 2020; Pan *et al.*, 2020]. Typically, an image captioning model consists of a visual encoder and a language model. Visual encoders range from CNNs [Anderson *et al.*, 2018] to Vision Transformers [Dosovitskiy *et al.*, 2021], while language models have transitioned from LSTMs to Transformers [Vaswani *et al.*, 2017]. More recently, MLLMs, such as LLaVA [Liu *et al.*, 2024], IDEFICS [Laurençon *et al.*, 2023], and Llama 3.2 Vision [Dubey *et al.*, 2024], have integrated these capabilities, transforming how captions are generated and evaluated.

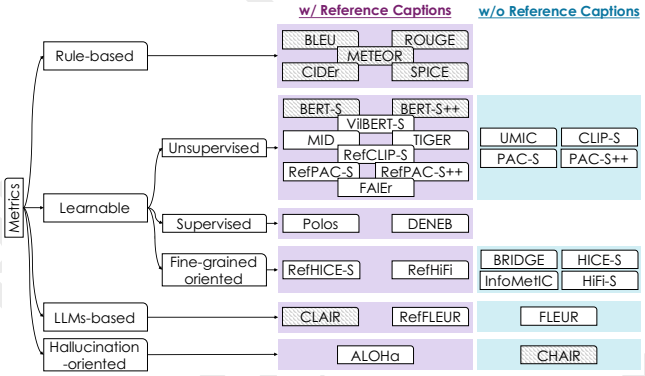

Figure 1: Taxonomy of image captioning metrics, distinguishing between reference-based and reference-free approaches. Grey dashed lines indicate the absence of the input image during evaluation.

Training strategies have also evolved, incorporating supervised learning integrating both cross-entropy loss and reinforcement learning using self-critical sequence training [Rennie *et al.*, 2017]. Several benchmark datasets have facilitated the development of image captioning models, including COCO [Lin *et al.*, 2014], Conceptual Captions [Sharma *et al.*, 2018], and nocaps [Agrawal *et al.*, 2019], each providing varying levels of diversity and complexity.

As captioning models improve, the role of evaluation metrics becomes increasingly critical. A good metric must assess captions across multiple dimensions to ensure alignment with human expectations both linguistically and semantically. A strong captioning metric should reward captions that are **fluent**, **accurate**, and **faithful** to the image content, while also ensuring they focus on the **most salient aspects** and strike a balance between **detail** and **conciseness**. Furthermore, it should penalize hallucinated or misleading information.

Early metrics such as BLEU [Papineni *et al.*, 2002], ROUGE [Lin, 2004], and METEOR [Banerjee and Lavie, 2005] were borrowed from machine translation and focused on n-gram overlap with reference captions. Later, specialized metrics like CIDEr [Vedantam *et al.*, 2015] and SPICE [An-

derson *et al.*, 2016] aimed to improve correlation with human judgment by incorporating consensus-based scoring and scene graph analysis, respectively. While these metrics introduced refinements, they remained constrained by their reliance on reference captions and linguistic overlap.

Recent advances have shifted towards learnable metrics leveraging vision-and-language pre-trained models. CLIP-S [Hessel *et al.*, 2021] introduces a contrastive approach to measure image-caption alignment based on CLIP embeddings. Fine-grained metrics like BRIDGE [Sarto *et al.*, 2024a] and HiFi-S [Yao *et al.*, 2024] capture localized semantics and hierarchical structures. LLM-based metrics such as CLAIR [Chan *et al.*, 2023] and FLEUR [Lee *et al.*, 2024] integrate reasoning and explanation into caption evaluation.

With the rise of MLLMs as image description generators, the nature of captions has changed, as these models generate captions that differ in length and specificity. Conventional evaluation metrics may no longer reliably assess caption quality in this new paradigm, posing new challenges for the task. Following this insight, this survey provides a comprehensive analysis of existing captioning metrics, from traditional rule-based approaches to modern learnable methods and LLM-driven techniques. We explore their evolution, strengths, and limitations while emphasizing the growing need for robust evaluation frameworks that can adapt to the stylistic diversity introduced by MLLMs. Additionally, we present an in-depth experimental study assessing metrics across multiple dimensions, including correlation with human judgment, ranking accuracy, and sensitivity to hallucinations. Our analysis highlights the performance gaps of conventional metrics and showcases promising directions for future research.

2 The Evolution of Captioning Metrics

In this section, we present the evolution of captioning metrics, outlining their development from traditional non-learning-based evaluation strategies to recent advancements incorporating LLMs. Fig. 1 presents a taxonomy that highlights the key characteristics of the most commonly used metrics, including their reliance on reference captions or image input.

2.1 Rule-based Metrics

Many widely used evaluation metrics for image captioning originate from NLP tasks and rely heavily on n-gram matching techniques. Popular metrics in this domain include BLEU [Papineni *et al.*, 2002], METEOR [Banerjee and Lavie, 2005], ROUGE [Lin, 2004], CIDEr [Vedantam *et al.*, 2015], and SPICE [Anderson *et al.*, 2016]. Despite their extensive use in image captioning, these traditional metrics exhibit several limitations. Notably, they do not consider the input image during evaluation and often demonstrate weak correlations with human judgment. Also, their reliance on reference captions restricts their ability to comprehensively evaluate generated captions, as their effectiveness is inherently influenced by the style and quality of the reference dataset.

BLEU [Papineni *et al.*, 2002] and METEOR [Banerjee and Lavie, 2005] were originally designed for machine translation tasks. BLEU evaluates n-gram precision and introduces the concept of modified unigram precision, which adjusts word frequency counts by considering the maximum occurrence of each word across all reference translations. By design, it penalizes overly long translations through precision-based scoring and mitigates excessively short ones with a brevity penalty, promoting balance in word choice, order, and length. However, BLEU lacks a direct measure of recall, which quantifies the proportion of matched n-grams in the reference. Recall is particularly crucial for evaluating translation quality, as it reflects the extent to which a candidate text captures the content of the reference. Introduced to mitigate this limitation, METEOR prioritizes recall by considering unigram matches between candidate and reference sentences and incorporating exact matches, stemmed forms, and semantic equivalences. This enables METEOR to provide a more detailed and holistic assessment of translation quality. Similarly, ROUGE [Lin, 2004], initially developed for summarization tasks, has been adapted for evaluating visual descriptions and emphasizes n-gram recall by analyzing the overlap of n-grams between candidate and reference texts.

Recognizing the limitations of these general-purpose metrics, CIDEr [Vedantam *et al.*, 2015] and SPICE [Anderson *et al.*, 2016] were designed for the captioning task. CIDEr employs a TF-IDF weighting scheme to reweight the significance of different n-grams, capturing both precision and recall while accounting for their contextual importance. SPICE takes a structured approach by converting captions into scene graphs and measuring their similarity, aligning the evaluation with the semantic and structural elements of the captions.

2.2 Learnable Metrics

All the aforementioned metrics primarily rely on textual-level comparisons and struggle to adequately account for synonym matches from a linguistic perspective. Moreover, they assume that human-written reference captions perfectly reflect the content of the image, which may not always be the case. To overcome these limitations, recent metrics leverage pre-trained models to compare textual-only and visual-textual content, using both supervised and unsupervised approaches.

Unsupervised Metrics. BERT-S [Zhang *et al.*, 2020], for instance, utilizes learned embeddings from the pre-trained language model BERT [Devlin *et al.*, 2018] to more effectively measure semantic similarities between candidate and reference captions. Building on this, BERT-S++ [Yi *et al.*, 2020] enhances the approach by incorporating the variance across multiple reference captions, improving robustness. However, these metrics still do not integrate the visual modality, which can limit their ability to fully evaluate the performance of captioning models. To overcome this, metrics that explicitly include the image as input have been introduced. These methods leverage pre-trained vision-and-language models, such as UNITER [Chen *et al.*, 2020], ViLBERT [Lu *et al.*, 2019], and CLIP [Radford *et al.*, 2021], to evaluate captions in a way that aligns more closely with both textual and visual content.

Among these metrics, UMIC [Lee *et al.*, 2021] stands out as it effectively addresses a key limitation of BERT-S: the difficulty of evaluating image captions in the absence of sufficient reference captions. To overcome this, it introduces a reference-free approach built on a fine-tuned version of a pre-trained multimodal Transformer (*i.e.*, UNITER). By leverag-

ing contrastive learning, the model is trained to differentiate between ground-truth captions and diverse synthetic negative samples. These negative samples are meticulously designed to cover various undesirable captioning cases, such as those that are grammatically incorrect, irrelevant to the image, or relevant but containing incorrect keywords. This approach ensures a more robust evaluation of image captioning models without relying on reference captions.

TIGEr [Jiang *et al.*, 2019], on the other hand, focuses on the integration of the image as one of the inputs for computing an evaluation score. Building upon a text-image grounding model [Lee *et al.*, 2018], it allows evaluating caption quality not only based on how well a caption represents image content but also on how well machine-generated captions match human-generated ones. Following this line and inspired by BERT-S, ViLBERT-S [Lee *et al.*, 2020] computes cosine similarity between token embeddings for reference and candidate sentences. However, different from BERT-S, the token embedding is computed with the consideration of image contexts. On a different line, FAIEr [Wang *et al.*, 2021] employs the scene graph as a bridge between the image and textual modality. Specifically, it fuses scene graphs of the image and references as a union scene graph and compares it with the scene graph of generated captions.

Leveraging the success of employing pre-trained vision-and-language models, numerous metrics started exploiting the CLIP extensive pre-training to compute image captioning evaluation scores. Among these, CLIP-S [Hessel *et al.*, 2021] was the first metric to utilize a modified cosine similarity between image and candidate caption representations derived from the visual and textual encoders of CLIP. In a similar way, MID [Kim *et al.*, 2022] leverages CLIP-based visual-textual features to compute negative Gaussian cross-mutual information, yielding a more effective evaluation metric.

While these advancements highlight the potential of contrastive-based embedding spaces for caption evaluation, large-scale models pre-trained on web-sourced data have inherent limitations. In particular, CLIP, which is trained on web-collected image-caption pairs, may be suboptimal for image evaluation tasks, as such annotations often lack the richness, grammatical correctness, and detail required to assess generated, typically longer captions. To address this, methods like PAC-S [Sarto *et al.*, 2023] and its improved version PAC-S++ [Sarto *et al.*, 2024b] aim to refine the CLIP embedding space by incorporating a fine-tuning phase. This process incorporates curated image-caption pairs along with synthetically generated positive examples, achieving improved performance in image captioning evaluation.

Supervised Metrics. All the aforementioned metrics leverage the CLIP embedding space, either pre-trained or fine-tuned on image-caption pairs. However, CLIP was not originally designed specifically for computing evaluation scores, which can limit its effectiveness. To address this challenge, a less explored research direction focuses on refining the CLIP embedding space through supervised training approaches. For instance, the Polos metric [Wada *et al.*, 2024] is based on Polaris, a dataset specifically designed for predicting evaluation scores under direct supervision from human-annotated judgments. As part of this work, the authors propose a multimodal metric learning framework based on human feedback, which handles both image and text inputs and learns directly from human evaluations based on multimodal inputs.

Building on this foundation, the DENEB metric [Matsuda *et al.*, 2024] introduces a supervised evaluation approach specifically designed to be robust against hallucinations. DENEB processes multiple reference captions simultaneously to effectively capture similarities between an image, a candidate caption, and reference captions. Furthermore, to improve the visual diversity of the Polaris dataset, DENEB introduces Nebula, an extended dataset containing approximately three times the number of images, further enhancing its robustness and scalability for evaluation tasks.

Fine-grained Oriented Metrics. More recent metrics still utilize multimodal models (*e.g.* CLIP) while integrating additional components to enhance performance, particularly by improving their ability to reward fine-grained details.

Some metrics adopt a reference-free approach, arguing that reliance on reference captions introduces challenges and additional costs, thereby increasing the complexity of the evaluation process. For instance, BRIDGE [Sarto *et al.*, 2024a] is a reference-free metric that employs a dual-encoder architecture combined with a mapping module responsible for generating multimodal pseudo-captions, which integrate textual and dense visual features. Similarly, InfoMetIC [Hu *et al.*, 2023] is designed to provide fine-grained feedback on caption quality. Beyond assessing precision and recall, it identifies specific errors, such as incorrect words and unmentioned image regions. To achieve this, InfoMetIC incorporates a fusion module to model intra- and inter-modality relationships, along with a fine-grained module that enhances the accuracy of error localization in both text and image content.

Available in both reference-free and reference-based versions, HICE-S [Zeng *et al.*, 2024] highlights the limitations of CLIP-based metrics, which primarily assess global image-text compatibility but often struggle with detecting local textual hallucinations and maintaining sensitivity to small visual objects. To overcome these challenges, HICE-S introduces a hierarchical scoring mechanism that leverages the SAM model [Kirillov *et al.*, 2023] to generate masks for localized visual regions and corresponding textual phrases. These are then processed through a modified version of the CLIP architecture [Sun *et al.*, 2024], enhancing the model's ability to capture fine-grained visual-semantic relationships.

Following a similar hierarchical approach, HiFi-S [Yao *et al.*, 2024] is a fine-grained image description evaluation metric that represents both text and images as parsing graphs. These graphs organize multi-granular instances into a hierarchical structure based on their inclusion relationships, enabling a comprehensive scene analysis across modalities from global to local levels. Additionally, HiFi-S incorporates an LLM to evaluate the fluency of candidate descriptions, further enhancing the evaluation process.

2.3 LLMs-based Metrics

In recent years, the integration of LLMs into the captioning evaluation pipeline has gained popularity, as demonstrated by HiFi-S. While first attempts compared generated sentences with reference captions using BERT-based embeddings, more

	External Models	Inputs		Flickr8k-Expert	Flickr8k-CF	Polaris	Nebula	Composite	Pascal-50S	FOIL
		Image	Refs	Kendall τ_c	Kendall τ_b	Kendall τ_c	Kendall τ_c	Kendall τ_c	Accuracy	Accuracy
Rule-based										
BLEU-4 [Papineni *et al.*, 2002]	-		✓	30.8	16.9	46.3	40.4	30.6	74.0	66.2
ROUGE [Lin, 2004]	-		✓	32.3	19.9	46.3	42.6	32.4	78.0	54.6
METEOR [Banerjee and Lavie, 2005]	-		✓	41.8	22.2	51.2	46.8	38.9	**81.1**	70.1
CIDEr [Vedantam *et al.*, 2015]	-		✓	43.9	**24.6**	**52.1**	**48.1**	37.7	80.1	**85.7**
SPICE [Anderson *et al.*, 2016]	-		✓	**44.9**	24.4	51.0	44.0	**40.3**	76.7	75.5
Learnable Unsupervised										
FAIEr [Wang *et al.*, 2021]	-	✓	✓	-	-	-	-	-	81.4	-
TIGEr [Jiang *et al.*, 2019]	SCAN	✓	✓	49.3	-	-	-	45.4	80.7	-
UMIC [Lee *et al.*, 2021]	$\text{UNITER}_{\text{BASE}}$	✓		46.8	-	49.8	-	56.1	85.1	-
BERT-S [Zhang *et al.*, 2020]	$\text{BERT}_{\text{BASE}}$		✓	39.2	22.8	51.6	47.0	30.1	80.1	88.6
BERT-S++ [Yi *et al.*, 2020]	$\text{BERT}_{\text{BASE}}$		✓	46.7	-	-	-	44.9	80.1	-
ViLBERT-S [Lee *et al.*, 2020]	$\text{ViLBERT}_{\text{BASE}}$	✓	✓	50.1	-	-	-	52.4	79.6	-
MID [Kim *et al.*, 2022]	CLIP ViT-B	✓	✓	54.9	37.3	51.3	**51.3**	-	85.2	90.5
CLIP-S [Hessel *et al.*, 2021]	CLIP ViT-B	✓		51.2	34.4	52.3	46.9	53.8	80.9	87.2
RefCLIP-S [Hessel *et al.*, 2021]	CLIP ViT-B	✓	✓	53.0	36.4	52.3	46.9	55.4	83.3	91.0
PAC-S [Sarto *et al.*, 2023]	CLIP ViT-B	✓		54.3	36.0	52.3	47.2	55.7	82.4	89.9
RefPAC-S [Sarto *et al.*, 2023]	CLIP ViT-B	✓	✓	55.9	37.6	55.2	50.6	57.3	84.7	93.7
PAC-S++ [Sarto *et al.*, 2024b]	CLIP ViT-B	✓		54.5	37.0	52.4	-	58.3	82.3	90.2
RefPAC-S++ [Sarto *et al.*, 2024b]	CLIP ViT-B	✓	✓	55.7	37.9	54.8	-	59.1	84.5	93.5
InfoMetIC [Hu *et al.*, 2023]	CLIP ViT-B	✓		55.5	36.6	-	-	59.3	**86.5**	-
CLIP-S [Hessel *et al.*, 2021]	CLIP ViT-L	✓		52.6	35.2	53.2	-	55.4	81.7	90.9
RefCLIP-S [Hessel *et al.*, 2021]	CLIP ViT-L	✓	✓	54.4	36.5	55.5	-	-	85.0	94.9
PAC-S [Sarto *et al.*, 2023]	CLIP ViT-L	✓		55.5	36.8	52.4	47.9	56.5	82.2	91.9
RefPAC-S [Sarto *et al.*, 2023]	CLIP ViT-L	✓	✓	57.1	37.7	55.5	50.4	-	85.0	**95.3**
PAC-S++ [Sarto *et al.*, 2024b]	CLIP ViT-L	✓		57.4	38.5	53.6	-	**62.0**	82.4	-
RefPAC-S++ [Sarto *et al.*, 2024b]	CLIP ViT-L	✓	✓	**57.9**	**38.8**	**55.6**	-	61.6	84.7	-
Learnable Supervised										
Polos [Wada *et al.*, 2024]	CLIP ViT-B	✓	✓	56.4	37.8	**57.8**	53.9	57.6	86.5	93.3
DENEB [Matsuda *et al.*, 2024]	CLIP ViT-B	✓	✓	56.5	38.0	-	54.1	57.9	87.0	95.1
DENEB [Matsuda *et al.*, 2024]	CLIP ViT-L	✓	✓	**56.8**	**38.3**	-	**54.3**	**58.2**	**87.8**	**95.4**
Learnable Fine-grained Oriented										
BRIDGE [Sarto *et al.*, 2024a]	CLIP ViT-B	✓		54.8	36.1	-	-	55.0	82.6	91.5
BRIDGE [Sarto *et al.*, 2024a]	CLIP ViT-L	✓		55.8	36.3	-	-	57.2	82.9	93.0
HICE-S [Zeng *et al.*, 2024]	SAM+Alpha-CLIP ViT-L	✓		56.4	37.2	-	-	57.9	86.1	93.1
RefHICE-S [Zeng *et al.*, 2024]	SAM+Alpha-CLIP ViT-L	✓	✓	57.7	**38.2**	-	-	58.7	**87.3**	**96.4**
HiFi-S [Yao *et al.*, 2024]	SAM+BLIP-2	✓		**58.4**	-	-	-	**65.8**	83.0	-
LLMs-based										
CLAIR [Chan *et al.*, 2023]	GPT-3.5		✓	48.3	-	-	**52.7**	61.0	78.7	81.4
FLEUR [Lee *et al.*, 2024]	LLaVA v1.5-13B	✓		53.0	38.6	-	-	63.5	83.2	96.8
RefFLEUR [Lee *et al.*, 2024]	LLaVA v1.5-13B	✓	✓	**51.9**	**38.8**	-	-	**64.2**	**85.5**	**97.3**

Table 1: A quantitative comparison between metrics, reporting correlation scores on Flickr8k-Expert, Flickr8k-CF, Polaris, Nebula, and Composite, along with accuracy on Pascal-50S and FOIL. The best scores within each category are in bold, overall best scores are underlined.

recent approaches leverage the advanced reasoning and extensive pre-training capabilities of LLMs to produce more robust evaluation scores. For example, CLAIR [Chan *et al.*, 2023] utilizes an LLM to rate the alignment of a candidate caption with a set of reference captions. Notably, CLAIR evaluates captions without considering image content. In response, FLEUR [Lee *et al.*, 2024] is a reference-free, explainable evaluation metric for image captioning, which directly leverages a score generated by an MLLM and adjusted with a smoothing function to better align with human judgments.

2.4 Hallucination-oriented Metrics

In image captioning, hallucination refers to the inclusion of information, objects, or details in a generated caption that are not present in the corresponding image, compromising the reliability and quality of the description. Accurately identifying captions with potential object hallucinations is therefore essential. Metrics such as CHAIR [Rohrbach *et al.*, 2018] and ALOHa [Petryk *et al.*, 2024] are specifically designed to address this challenge. Specifically, the CHAIR metric calculates what proportion of words generated is actually in the image according to the ground-truth sentences and detected object. However, this metric is restricted to a fixed set of COCO objects and their synonyms. To overcome this limitation, ALOHa introduces an open-vocabulary approach that utilizes LLMs to detect object hallucinations. Specifically, ALOHa extracts groundable objects from the candidate caption using an LLM, measures their semantic similarity to reference objects in captions or detections, and applies Hungarian matching to compute the final hallucination score.

3 Experimental Evaluation

This section presents a comprehensive analysis of captioning evaluation metrics through quantitative experiments. The evaluation spans multiple dimensions, including correlation with human judgments, ranking accuracy, and sensitivity to object hallucinations. Experiments are conducted on diverse datasets, with results summarized in Table 1. Additionally, we assess how well existing metrics evaluate captions generated by modern MLLMs, with results presented in Table 2.

3.1 Correlation with Human Judgment

We analyze the correlation of metrics with human judgment, highlighting their varying behaviours. Experimental results

are reported on standard captioning evaluation datasets, including Flickr8k-Expert, Flickr8k-CF, and Composite, along with the more recent Polaris and Nebula datasets.

Datasets. Flickr8k-Expert [Hodosh *et al.*, 2013] contains 17k annotations for 5,664 images, with each pair scored from 1 (no correlation) to 4 (accurate depiction). Flickr8k-CF [Hodosh *et al.*, 2013] provides 145k binary quality judgments for 48k pairs across 1,000 unique images, using the mean proportion of "yes" votes as the alignment score. The Composite dataset [Aditya *et al.*, 2015] includes 12k human ratings for image-caption pairs (with around 4k unique images), assessed on a 1–5 scale. However, these datasets lack model-generated captions, leading to a domain gap when applying them to train evaluation metrics. To address this limitation, the Polaris dataset [Wada *et al.*, 2024] introduces 131k human judgments from 550 evaluators (*i.e.*, around ten times larger than previous datasets) covering both human-written and machine-generated captions from ten image captioning models. Expanding on Polaris, the Nebula dataset [Matsuda *et al.*, 2024] includes 32,978 images with human judgments from 805 annotators. In line with prior studies [Hessel *et al.*, 2021; Wada *et al.*, 2024], we use Kendall's correlation coefficient τ_b for Flickr8k-CF and Kendall's τ_c for the other datasets.

Experimental Analysis and Key Takeaways. Among rule-based methods, CIDEr demonstrates the highest performance across most datasets, with the exception of Flickr8k-Expert and Composite, where SPICE outperforms it by +1 and +2.6 points, respectively. Within the learnable metrics, RefPAC-S++ (ViT-L) achieves the best results on Flickr8k-Expert, Flickr8k-CF, and Polaris, while ViLBERT-S yields the highest scores among metrics evaluated on the Nebula dataset. Notably, reference-based metrics generally outperform their reference-free counterparts. However, among the reference-free methods, InfoMetIC and PAC-S variants demonstrate superior performance, emphasizing the importance of refining large pre-trained backbones in the absence of reference captions. When scaling the backbone size, PAC-S variants maintain superior performance, demonstrating their robustness and scalability. In a supervised setting, comparisons are more complex due to differences in backbone sizes. Overall, DENEB (ViT-L) achieves the highest results across multiple datasets. Interestingly, despite leveraging a supervised approach and incorporating an additional learnable module, DENEB is outperformed by RefPAC-S++ (ViT-L), which is trained in an unsupervised manner. This highlights the effectiveness of leveraging CLIP pre-training and enhancing it with regularization through additional generated visual-textual pairs. For fine-grained evaluation, the HiFi metric, employing a hierarchical parsing graph, achieves the best results in a reference-free setting, surpassing HICE-S by +7.9 and BRIDGE (ViT-L) by +8.6 on the Composite dataset. On the Flickr8k-CF dataset, the best performance is instead achieved by RefHICE-S. Among LLM-based metrics, FLEUR performs best, highlighting the importance of leveraging a multimodal approach that incorporates input images, unlike the CLAIR metric which relies solely on reference captions for evaluation. This underscores the critical role of visual information in the captioning task.

Overall, no single metric consistently outperforms all others across datasets. Interestingly, on the Flickr8k-CF dataset, comparable results are achieved using both RefPAC-S++ and the MLLM-based metric FLEUR. This comparison is particularly noteworthy, as the FLEUR metric, based on a LLaVA model with 13B parameters, achieves similar performance to metrics that utilize a smaller CLIP model with around 400M parameters. This highlights that for specific tasks like captioning evaluation, refining a pre-trained embedding space may be more effective than relying on a larger, multi-task embedding. Moreover, there is a significant performance gap between rule-based metrics and newer approaches, highlighting the importance of leveraging input images and the advantages of large-scale pre-training in achieving superior performance.

3.2 Pairwise Ranking

We focus on the pairwise ranking ability of current captioning metrics, measuring performance on the Pascal-50S dataset.

Dataset. Pascal-50S [Vedantam *et al.*, 2015] evaluates captioning metrics using pairwise preference judgments. It consists of 4,000 sentence pairs linked to 1,000 images, each with 50 reference captions. Human judges determine which caption better describes the image, categorizing pairs into human-correct, human with one incorrect caption, human vs. machine-generated, and machine-generated. In this setting, we compute accuracy instead of correlation scores, reporting the averaged results across the four categories.

Experimental Analysis and Key Takeaways. Among rule-based metrics, METEOR achieves the highest accuracy, outperforming CIDEr by +1 point. For learnable unsupervised metrics, the results deviate from trends seen in correlation-based evaluations, where larger backbones and reference-based metrics typically excel. In this case, InfoMetIC, a reference-free metric using the ViT-B backbone, emerges as the best among all metrics. Its ability to identify incorrect semantic words and unmentioned visual regions proves advantageous for this task. Conversely, for other learnable and LLM-based metrics, the results closely align with those observed in human correlation evaluations, with the highest overall accuracy achieved by DENEB (ViT-L).

3.3 Sensitivity to Object Hallucinations

We evaluate robustness to hallucinations on the FOIL dataset.

Dataset. FOIL [Shekhar *et al.*, 2017] consists of image-caption pairs derived from COCO [Lin *et al.*, 2014], where captions are intentionally altered by introducing a single erroneous word, that makes the modified caption closely resembles the original but includes a specific mistake. In our evaluation, we compute the percentage of times the original caption gets the highest score, according to each metric.

Experimental Analysis and Key Takeaways. In this task, CIDEr outperforms all other rule-based metrics by a substantial margin, with a gain of around +10 points. Among learnable metrics, RefHICE achieves the highest accuracy, followed by DENEB and RefPAC-S, both using the ViT-L backbone. Reference-free versions of HICE and PAC-S experience a significant drop, indicating the need for reference captions to detect hallucinated objects. The results also highlight

		LLM	Length	BLEU-4	METEOR	CIDEr	CLIP-S	PAC-S++	RefCLIP-S	RefPAC-S++	Polos	BRIDGE
Captioning Models	Show and Tell [Vinyals *et al.*, 2015]	-	9.1	31.4	25.0	97.2	0.715	0.654	0.779	0.752	0.585	0.788
	Show, Attend and Tell [Xu *et al.*, 2015]	-	9.1	33.4	26.2	104.6	0.727	0.670	0.790	0.766	0.609	0.804
	Up-Down [Anderson *et al.*, 2018]	-	9.5	36.7	27.9	122.7	0.740	0.680	0.804	0.778	0.640	0.821
	SGAE [Yang *et al.*, 2019]	-	9.4	38.6	28.8	129.8	0.750	0.691	0.812	0.786	0.655	0.833
	AoANet [Huang *et al.*, 2019]	-	9.5	39.1	29.0	128.9	0.753	0.693	0.813	0.787	0.660	0.836
	$\mathcal{M}^2$ Transformer [Cornia *et al.*, 2020]	-	9.7	39.1	29.2	131.2	0.757	0.699	0.813	0.791	0.629	0.841
	X-Transformer [Pan *et al.*, 2020]	-	9.6	39.7	29.5	132.8	0.762	0.701	0.819	0.792	0.668	0.845
	VinVL [Zhang *et al.*, 2021]	-	10.0	41.0	31.1	140.9	0.784	0.715	0.836	0.805	0.708	0.865
	COS-Net [Li *et al.*, 2022]	-	9.6	42.0	30.6	141.1	0.773	0.712	0.829	0.803	0.692	0.859
	BLIP-2 [Li *et al.*, 2023]	Flan T5-XL	9.6	**43.8**	**31.7**	**146.0**	**0.782**	**0.719**	**0.838**	**0.810**	**0.716**	**0.868**
General Purpose MLLMs	InstructBLIP [Dai *et al.*, 2023]	Vicuna-7B	10.2	**41.2**	**31.8**	**142.2**	0.786	0.721	0.838	**0.810**	**0.714**	0.871
	LLaVA-1.5 [Liu *et al.*, 2024]	Vicuna-7B	14.5	8.1	28.0	69.6	0.785	0.707	0.828	0.794	0.666	0.867
	IDEFICS [Laurençon *et al.*, 2023]	Llama-7B	8.8	36.8	28.3	125.1	0.788	0.711	**0.840**	0.803	0.699	0.863
	IDEFICS-2 [Laurençon *et al.*, 2024b]	Mistral-7B	12.2	19.1	24.3	74.1	0.800	0.711	0.819	0.780	0.626	0.865
	IDEFICS-3 [Laurençon *et al.*, 2024a]	Llama-3-8B	14.2	17.4	24.5	62.2	0.801	0.710	0.811	0.771	0.615	0.865
	Llama-3.2 [Dubey *et al.*, 2024]	Llama-3.2-11B	22.2	14.3	25.8	46.0	**0.827**	**0.734**	0.828	0.796	0.692	**0.900**
	InstructBLIP [Dai *et al.*, 2023]	Vicuna-7B	43.3	9.8	**25.0**	2.8	0.828	**0.738**	0.817	**0.792**	**0.709**	**0.912**
	LLaVA-1.5 [Liu *et al.*, 2024]	Vicuna-7B	49.5	8.3	23.0	0.3	0.813	0.722	0.808	0.783	0.689	0.898
	IDEFICS [Laurençon *et al.*, 2023]	Llama-7B	29.0	**11.4**	24.8	**43.9**	0.792	0.719	0.815	0.788	0.679	0.867
	IDEFICS-2 [Laurençon *et al.*, 2024b]	Mistral-7B	22.1	7.3	20.0	31.7	0.728	0.683	0.773	0.760	0.575	0.794
	IDEFICS-3 [Laurençon *et al.*, 2024a]	Llama-3-8B	123.6	2.0	12.2	8.3	0.777	0.697	0.770	0.748	0.605	0.856
	Llama-3.2 [Dubey *et al.*, 2024]	Llama-3.2-11B	31.4	10.2	23.9	30.5	**0.832**	0.736	**0.818**	0.790	0.689	0.906
	Humans		*10.4*	-	*24.1*	*87.6*	*0.782*	*0.710*	*0.822*	*0.792*	*0.654*	*0.856*

Table 2: Evaluation scores on the COCO test set comparing traditional captioning models with general-purpose MLLMs. For MLLMs, we assess both short captions and longer, unconstrained descriptions generated using the default prompt of the model.

the advantages of stronger backbones. For instance, PAC-S (ViT-L) outperforms PAC-S (ViT-B) by +2 points, demonstrating the impact of model architecture. Notably, rule-based metrics like CIDEr are significantly surpassed by modern multimodal approaches, with a performance gap of approximately 10 points compared to RefPAC-S (ViT-L) and nearly 12 points against the LLM-based RefFLEUR. This disparity highlights the limitations of rule-based approaches in effectively capturing the nuanced semantics of image captions, particularly in detecting subtle errors such as hallucinations. Metrics like RefPAC-S and RefFLEUR benefit not only from powerful backbones but also from their ability to align text with visual content, allowing them to detect discrepancies much more accurately than traditional metrics. These results empathize the growing need for evaluation metrics that assess not just linguistic fluency but also semantic accuracy and visual relevance through advanced multimodal understanding.

3.4 System-level Correlation

We assess whether well-established captioning metrics effectively evaluate captions generated by modern MLLMs, especially when their style deviates from traditional COCO captions. The potential limitations of these metrics originate from their design: rule-based metrics rely on reference captions that adhere to COCO-style annotations, while learnable metrics are primarily fine-tuned on COCO or similar datasets. As a result, their ability to evaluate captions with diverse linguistic structures remains uncertain. To investigate this, we assess MLLM-generated captions in two formats: short captions, comparable in length to COCO ones, and longer, unconstrained descriptions[1]. This enables us to analyze metric performance across varying caption styles and lengths.

In Table 2, we evaluate popular captioning models on the COCO test set using a range of metrics to assess their effectiveness and highlight differences in their behavior[2]. Our analysis includes both traditional captioning models and modern MLLMs designed for broader tasks. Additionally, we establish a human-based baseline by randomly selecting one human-annotated caption from the five provided in COCO and comparing it against the remaining references[3]. For evaluation, we employ standard rule-based scores such as BLEU, METEOR and CIDEr, as well as two widely used learnable metrics, CLIP-S and PAC-S++, along with their reference-based counterparts. Additionally, we include two recent evaluation methods: Polos, a supervised metric, and BRIDGE, which focuses on fine-grained visual details. This selection allows for assessing how different evaluation methods capture caption quality across various models.

As it can be seen, for models explicitly trained for the image captioning task, such as $\mathcal{M}^2$ Transformer, the generated captions tend to align with COCO in both length and style. Under these conditions, traditional evaluation metrics like CIDEr, effectively recognize the superior quality of BLIP-2 captions. Similarly, recently developed metrics leveraging large-scale pre-trained models maintain strong performance.

When evaluating captions generated by MLLMs interesting patterns emerge. If the captions remain similar in length to the ones contained in COCO, rule-based metrics still perform effectively: CIDEr, for instance, assigns a score of 142.2 to captions generated by InstructBLIP. However, as caption length increases, these metrics exhibit a sharp decline in scores. Notably, when evaluating LLaVA-1.5, which produces significantly longer captions than COCO ones, CIDEr

[1]For short captions, we use the prompt "`Briefly describe the image`". For longer ones, we rely on the MLLM default prompt (*e.g.* "`What is the content of the image?`").

[2]Note that for BLIP-2 we employ the captioning-specific model fine-tuned on COCO image-caption pairs.

[3]BLEU is not reported for the human-based baseline as its value is sensitive to the number of reference captions.

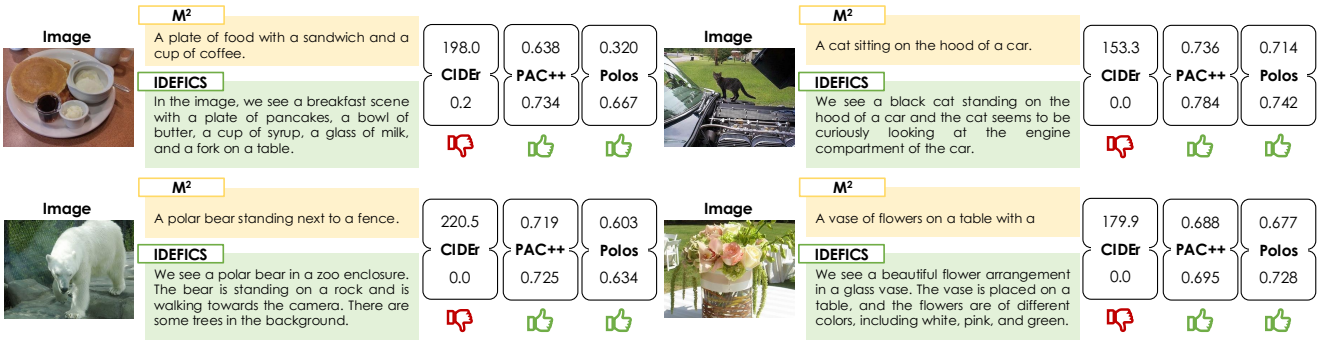

Figure 2: Qualitative examples showing the differences between captions generated by a traditional captioning model (*i.e.*, $\mathcal{M}^2$ Transformer) and MLLM-style captions (*i.e.*, IDEFICS). Length and style variations affect rule-based metrics, while learnable metrics remain robust.

score drops dramatically to just 0.3. Qualitative results showing this trend are reported in Fig. 2. This decline stems from rule-based metrics relying on COCO-style captions, which are shorter and simpler than detailed MLLM outputs, resulting in misleading evaluations.

In contrast, more recent metrics that incorporate large-scale components exhibit greater robustness to variations in caption length. A slight decline in performance is still observed: for example, PAC-S++ decreases from 0.710 to 0.697 when caption length increases nearly ninefold (*i.e.*, for IDEFICS-3). This primarily reflects a lower confidence level rather than a failure of the metric. In fact, although these metrics were not explicitly designed for evaluating long, highly detailed captions, they still maintain a high degree of reliability. For learnable reference-based metrics such as RefCLIP-S, RefPAC-S++, and Polos, performance degrades more significantly compared to their reference-free counterparts, reflecting a trend similar to that observed in rule-based metrics. However, unlike rule-based solutions, these metrics retain their ability to distinguish high-quality captions, correctly rewarding models such as Llama-3.2 and InstructBLIP.

Overall, these results indicate that recent learnable metrics designed for image captioning remain valuable for evaluating longer and more detailed captions generated by MLLMs, demonstrating reliability despite variations in length and style. However, reference-free metrics are generally preferable, as they more effectively distinguish high-quality captions and exhibit greater robustness to variations in length.

4 Future Directions and Open Problems

Despite progress in image captioning evaluation, several challenges remain, offering opportunities for future research.

Benchmark Evolution. Traditional evaluation datasets mainly feature short captions similar to those in COCO, creating a gap with the longer, more detailed outputs of modern MLLMs. While recent efforts, like Polaris and Nebula, incorporate captions from standard models, they often maintain the concise style of COCO descriptions. A valuable direction would involve creating new benchmarks specifically designed to assess the quality of metrics on longer, richer captions that better reflect the MLLM outputs, addressing challenges like synonyms, paraphrases, and domain-specific terms.

Explainability in Metrics. Most evaluation metrics generate scores without offering insights into the rationale behind their assessments, limiting their usefulness for improving captioning models. A few metrics, such as CLAIR, have advanced in this area by leveraging the interpretative capabilities of LLMs to offer explanations alongside scores. Building upon these advancements, future research should focus on further enhancing explainability, facilitating a deeper understanding of the strengths and weaknesses of evaluation models.

Detecting Hallucinations. Hallucination remains a major challenge in captioning, with models often generating inaccurate details. While benchmarks like FOIL help identify simpler errors, modern MLLMs produce more complex hallucinations, such as distorted relationships or invented elements. To address this, more diverse datasets are needed to assess these complex errors, including test cases designed to challenge evaluation metrics on altered attributes, incorrect relationships, or non-existent entities and actions. Additionally, although some metrics targeting hallucinations have emerged, more advanced solutions are required to capture the diversity and semantic richness of MLLM-generated captions.

Metrics Personalization. Current evaluation metrics rely on reference captions or image-caption alignment, but they fail to accommodate the diversity of user preferences or task-specific requirements. Future research should focus on the development of personalized evaluation metrics, that can let users prioritize factors such as detail, brevity, stylistic preferences, or domain relevance. By incorporating these custom priorities, evaluation can become more adaptable and meaningful across a variety of applications and user needs.

5 Conclusion

This survey examines image captioning evaluation metrics with the advent of MLLMs, comparing their performance on standard datasets and analyzing their adaptability to evolving caption styles. Our analysis reveals that traditional rule-based metrics struggle to evaluate longer, more detailed captions, while learnable metrics remain effective. Moreover, we highlight key open challenges and promising future directions, emphasizing the need for more robust, context-aware evaluation frameworks that can assess the growing diversity and complexity of captioning models.

Acknowledgments

We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources. This work has been conducted under a research grant co-funded by Leonardo S.p.A. and supported by the PNRR-M4C2 project FAIR and the PRIN 2022-PNRR project MUCES (CUP E53D23016290001), funded by EU - Next-Generation EU. The authors also thank Davide Caffagni and Lorenzo Baraldi for their valuable support.

References

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

[Agrawal *et al.*, 2019] Harsh Agrawal, Karan Desai, Xinlei Chen, Rishabh Jain, Dhruv Batra, Devi Parikh, et al. nocaps: novel object captioning at scale. In *ICCV*, 2019.

[Anderson *et al.*, 2016] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic Propositional Image Caption Evaluation. In *ECCV*, 2016.

[Anderson *et al.*, 2018] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, et al. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *CVPR*, 2018.

[Banerjee and Lavie, 2005] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL Workshops*, 2005.

[Chan *et al.*, 2023] David Chan, Suzanne Petryk, Joseph E Gonzalez, Trevor Darrell, et al. CLAIR: Evaluating Image Captions with Large Language Models. In *EMNLP*, 2023.

[Chen *et al.*, 2020] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: UNiversal Image-TExt Representation Learning. In *ECCV*, 2020.

[Cornia *et al.*, 2020] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-Memory Transformer for Image Captioning. In *CVPR*, 2020.

[Dai *et al.*, 2023] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, et al. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv:2305.06500*, 2023.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 2018.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021.

[Dubey *et al.*, 2024] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, et al. The Llama 3 Herd of Models. *arXiv:2407.21783*, 2024.

[Hessel *et al.*, 2021] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *EMNLP*, 2021.

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

[Hu *et al.*, 2023] Anwen Hu, Shizhe Chen, Liang Zhang, and Qin Jin. InfoMetIC: An Informative Metric for Reference-free Image Caption Evaluation. In *ACL*, 2023.

[Huang *et al.*, 2019] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on Attention for Image Captioning. In *ICCV*, 2019.

[Jiang *et al.*, 2019] Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, et al. TIGEr: Text-to-Image Grounding for Image Caption Evaluation. In *EMNLP*, 2019.

[Kim *et al.*, 2022] Jin-Hwa Kim, Yunji Kim, Jiyoung Lee, Kang Min Yoo, and Sang-Woo Lee. Mutual Information Divergence: A Unified Metric for Multimodal Generative Models. In *NeurIPS*, 2022.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, et al. Segment Anything. In *ICCV*, 2023.

[Laurençon *et al.*, 2023] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, et al. OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents. In *NeurIPS*, 2023.

[Laurençon *et al.*, 2024a] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. In *NeurIPS Workshops*, 2024.

[Laurençon *et al.*, 2024b] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? In *NeurIPS*, 2024.

[Lee *et al.*, 2018] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked Cross Attention for Image-Text Matching. In *ECCV*, 2018.

[Lee *et al.*, 2020] Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, et al. ViLBERTScore: Evaluating Image Caption Using Vision-and-Language BERT. In *EMNLP Workshops*, 2020.

[Lee *et al.*, 2021] Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Trung Bui, and Kyomin Jung. UMIC: An Unreferenced Metric for Image Captioning via Contrastive Learning. In *ACL*, 2021.

[Lee *et al.*, 2024] Yebin Lee, Imseong Park, and Myungjoo Kang. FLEUR: An Explainable Reference-Free Evaluation Metric for Image Captioning Using a Large Multimodal Model. In *ACL*, 2024.

[Li *et al.*, 2022] Yehao Li, Yingwei Pan, Ting Yao, and Tao Mei. Comprehending and Ordering Semantics for Image Captioning. In *CVPR*, 2022.

[Li *et al.*, 2023] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *ICML*, 2023.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, et al. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014.

[Lin, 2004] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL Workshops*, 2004.

[Liu *et al.*, 2024] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. In *CVPR*, 2024.

[Lu *et al.*, 2019] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*, 2019.

[Matsuda *et al.*, 2024] Kazuki Matsuda, Yuiga Wada, and Komei Sugiura. DENEB: A Hallucination-Robust Automatic Evaluation Metric for Image Captioning. In *ACCV*, 2024.

[Pan *et al.*, 2020] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-Linear Attention Networks for Image Captioning. In *CVPR*, 2020.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[Petryk *et al.*, 2024] Suzanne Petryk, David M Chan, Anish Kachinthaya, Haodi Zou, et al. ALOHa: A New Measure for Hallucination in Captioning Models. In *NAACL*, 2024.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, et al. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021.

[Rennie *et al.*, 2017] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, et al. Self-Critical Sequence Training for Image Captioning. In *CVPR*, 2017.

[Rohrbach *et al.*, 2018] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, et al. Object Hallucination in Image Captioning. In *EMNLP*, 2018.

[Sarto *et al.*, 2023] Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation. In *CVPR*, 2023.

[Sarto *et al.*, 2024a] Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. BRIDGE: Bridging Gaps in Image Captioning Evaluation with Stronger Visual Cues. In *ECCV*, 2024.

[Sarto *et al.*, 2024b] Sara Sarto, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, et al. Positive-Augmented Contrastive Learning for Vision-and-Language Evaluation and Training. *arXiv:2410.07336*, 2024.

[Sharma *et al.*, 2018] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *ACL*, 2018.

[Shekhar *et al.*, 2017] Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, et al. FOIL it! Find One mismatch between Image and Language caption. In *ACL*, 2017.

[Stefanini *et al.*, 2022] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From Show to Tell: A Survey on Deep Learning-based Image Captioning. *IEEE Trans. PAMI*, 45(1):539–559, 2022.

[Sun *et al.*, 2024] Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, et al. Alpha-CLIP: A CLIP Model Focusing on Wherever You Want. In *CVPR*, 2024.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, et al. Attention is all you need. In *NeurIPS*, 2017.

[Vedantam *et al.*, 2015] Ramakrishna Vedantam, Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based Image Description Evaluation. In *CVPR*, 2015.

[Vinyals *et al.*, 2015] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[Wada *et al.*, 2024] Yuiga Wada, Kanta Kaneda, Daichi Saito, and Komei Sugiura. Polos: Multimodal Metric Learning from Human Feedback for Image Captioning. In *CVPR*, 2024.

[Wang *et al.*, 2021] Sijin Wang, Ziwei Yao, Ruiping Wang, Zhongqin Wu, and Xilin Chen. FAIEr: Fidelity and Adequacy Ensured Image Caption Evaluation. In *CVPR*, 2021.

[Xu *et al.*, 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *ICML*, 2015.

[Yang *et al.*, 2019] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-Encoding Scene Graphs for Image Captioning. In *CVPR*, 2019.

[Yao *et al.*, 2024] Ziwei Yao, Ruiping Wang, and Xilin Chen. HiFi-Score: Fine-Grained Image Description Evaluation with Hierarchical Parsing Graphs. In *ECCV*, 2024.

[Yi *et al.*, 2020] Yanzhi Yi, Hangyu Deng, and Jinglu Hu. Improving Image Captioning Evaluation by Considering Inter References Variance. In *ACL*, 2020.

[Zeng *et al.*, 2024] Zequn Zeng, Jianqiao Sun, Hao Zhang, Tiansheng Wen, Yudi Su, Yan Xie, Zhengjue Wang, and Bo Chen. HICEScore: A Hierarchical Metric for Image Captioning Evaluation. In *ACM Multimedia*, 2024.

[Zhang *et al.*, 2020] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In *ICLR*, 2020.

[Zhang *et al.*, 2021] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, et al. VinVL: Revisiting Visual Representations in Vision-Language Models. In *CVPR*, 2021.