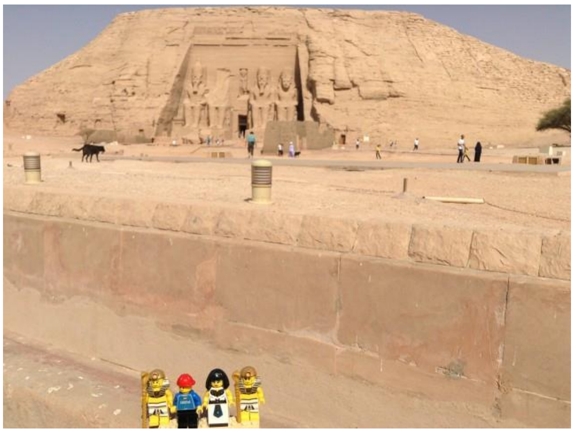
Surveying the Landscape of Image Captioning Evaluation: A Comprehensive Taxonomy and Novel Ensemble Method

Uri Berger[◇†] and **Gabriel Stanovsky**[◇] and **Omri Abend**[◇] and **Lea Frermann**[†]

[◇]School of Computer Science and Engineering, The Hebrew University of Jerusalem
[†]School of Computing and Information Systems, University of Melbourne
{uri.berger2, gabriel.stanovsky, omri.abend}@mail.huji.ac.il
lea.frermann@unimelb.edu.au

Abstract

The task of image captioning has recently been gaining popularity, and with it the complex task of evaluating the quality of image captioning models. In this work, we present the first survey and taxonomy of over 70 different image captioning metrics and their usage in hundreds of papers. We find that despite the diversity of proposed metrics, the vast majority of studies rely on only five popular metrics, which we show to be weakly correlated with human judgements. Instead, we propose ENSEMBEVAL — an ensemble of evaluation methods achieving the highest reported correlation with human judgements across 5 image captioning datasets, showing there is a lot of room for improvement by leveraging a diverse set of metrics.[1]



Lego figures in the middle of the desert
A pyramid with statues of ancient kings
Some people and a dog in front of an old monument

Figure 1: An image with various valid captions that have no overlap in content words, exemplifying some of the challenges in evaluating image captioning.

1 Introduction

Image captioning is a multi-modal task in which models generate textual captions describing an image. Evaluating the output of image captioning is challenging for several reasons. First, as in other generative tasks, multiple outputs may be valid for the same input, as illustrated in Figure 1, where different valid captions for the same image have no overlap in content words. Second, it involves bridging across modalities, requiring evaluators to compare text and images, unlike most generative tasks that only involve textual information. Attesting to the difficulty and importance of image captioning is the sheer volume of metrics proposed for this task, and their usage in hundreds of image captioning models (Figure 2).

In this work, we survey the different approaches proposed for evaluating image captioning, and provide a first taxonomy explaining over 70 metrics, which were largely developed independently of one another. Furthermore, based on our observations we devise an ensemble-based metric, ENSEMBEVAL, which achieves the highest reported correlation with human ratings.

In Section 2, we start our survey with a systematic approach to examine all relevant papers from 15 major venues in NLP, vision and machine learning between 2010 and 2024. This approach yields a body of work consisting of 71 different automatic evaluation metrics and 5 human evaluation paradigms, used in over 300 papers.

Next, we organize both automatic metrics and human evaluation in principled taxonomies (Section 3). Our taxonomy is the first to comprehensively cover all automatic metrics used in image captioning research, organized by a novel criterion: the specific property each metric aims to quantify. This enables future researchers to choose metrics based on the specific aspect their model is intended to improve. In addition, we propose the first taxonomy for human evaluation metrics, cat-

[1]Our code and data are available on github.com/uriberger/caption_evaluation.

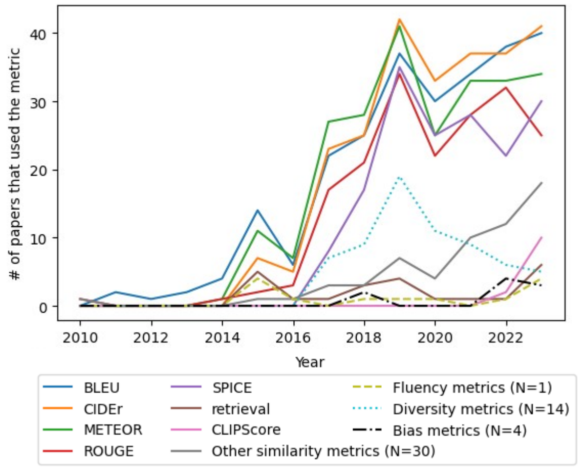


Figure 2: Metrics usage over the years. Metrics used in less than two papers are omitted.

egorizing them into groups such as comparative evaluation and scale rating evaluation.

We then survey the use of evaluation methods across a wide range of 314 papers (Section 4). Interestingly, we find that despite the wealth of different metrics proposed for the task, the vast majority of examined papers use only five simple metrics (BLEU, METEOR, ROUGE, CIDEr, SPICE), although there has been a the recent increase in the adoption of alternative metrics. Furthermore, we find that the use of human evaluation is declining in recent years and that significance and inter-annotator agreement are rarely reported.

Finally, in a series of evaluations, we show that the five most popular metrics show only a weak correlation with human ratings, whereas lesser-known metrics show a much higher correlation. This underscores the value of a systematic and comprehensive survey in identifying these lesser-known but more effective metrics. We then propose an ensemble of selected metrics, ENSEMBEVAL, optimized for diversity using a feature selection algorithm. We show that ENSEMBEVAL achieves the highest reported correlation with human ratings to date.

To recap, our contributions are twofold. First, we introduce the first comprehensive taxonomy of automatic captioning evaluation metrics, categorized by the property they assess. This will help future model developers select the most suitable metric based on the specific property they want their model to excel in. Second, our proposed ensemble approach based on this survey can be readily used as a state-of-the-art evaluation of image captioning models.

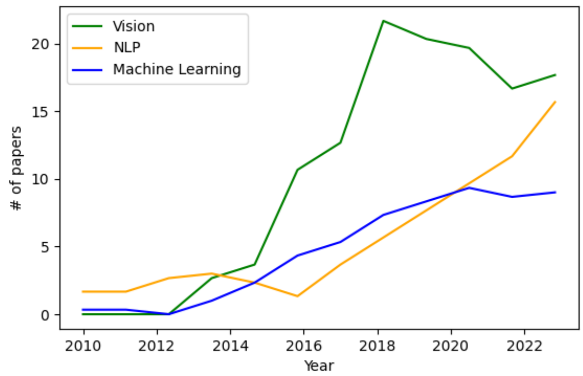


Figure 3: Annual number of image captioning papers per community, smoothed by convolving with a window size of 3 years.

Community	Conferences
NLP	AACL, ACL, CoNLL, EACL EMNLP, NAACL, TACL, *SEM
CV	CVPR, ICCV, ECCV
ML	Neurips, ICML, ICLR, AAAI

Table 1: Venues included in our review. NLP: Natural Language Processing. CV: Computer Vision. ML: Machine Learning.

2 Methodology

We perform a systematic review and gather usage statistics of metrics in conference papers on English image captioning. Overall, we examined 314 papers from 2010 to 2024, and identified 71 distinct metrics.

Venues. We focus on a large set of 15 venues spanning three distinct research communities (Table 1). Figure 3 shows the annual number of papers for each community.

Search strategy. To identify image captioning related papers within a venue, we search the venue proceedings for papers with the substrings *caption* or *description* in their titles, assuming relevant titles are likely to include variations of the phrases *image captioning* or *image descriptions*.

Filtering. Next, we manually filter out papers that we deem irrelevant to our review. Specifically, we exclude: 1) papers on other types of captioning such as video captioning; 2) papers unrelated to captioning that happen to have *caption* or *description* in their titles; 3) papers that focus on languages other than English; and 4) papers that did not conduct any experiments.

Years reviewed. We begin collecting data starting from 2024 backwards until reaching the first year with no relevant papers (2009).

Collected data. We collect two types of information. First, we intend to describe the types of metrics used in previous work. For each metric introduced or used in the papers we examined, we document its name, implementation details and the paper in which it was introduced. To the best of our knowledge, we are the first to systematically collect such a comprehensive set of metrics, as previous research has focused on a few well-known metrics (see Section 6).

Second, we seek to analyze the patterns of metrics usage. Therefore, we record which metrics were used in each examined paper.[2]

3 Taxonomy

In this section, we introduce a taxonomy of the metrics from the systematic review outlined in Section 2. This work is novel in two respects: we are the first to comprehensively include all metrics proposed and used in previous studies, rather than focusing on a limited set of well-known metrics. Second, we are the first to describe a taxonomy for human evaluation methods.

3.1 Automatic Evaluation

The release of large benchmarks like MSCOCO (Lin et al., 2014) has made manual evaluation of captioning models impractical. Hence, automatic tools for assessing generated captions have been developed, and some tools from other text generation tasks have been adapted for the image captioning task.

We turn to describing the various automatic evaluation methods. We categorize metrics based on the specific property they aim to evaluate.[3] The taxonomy is presented visually in Figure 4.

3.1.1 Relation with Ground-Truth

The largest and most widely used class of automatic methods use a ground-truth source (either a reference caption or the image) as a basis for evaluating the candidate. We categorize these methods based on the relation they examine (similarity/extrinsic) and the nature of the ground-truth (reference, image, both) they use.

Similarity to reference. Early automatic methods, still the most common today, compare the candidate caption to a reference caption.

Lexical similarity: Some metrics compare the candidate and reference captions at the word level using a straightforward textual comparison, most naïvely using exact match (Kang et al., 2023). More advanced methods include computing n-gram overlap (*BLEU*: Papineni et al. 2002, *NIST*: Doddington 2002, *ROUGE*: Lin 2004, *METEOR*: Banerjee and Lavie 2005), n-gram TF-IDF (*CIDEr*: Vedantam et al. 2015, *CIDEr-r*: Oliveira dos Santos et al. 2021), and calculating minimal edits to match the reference (*TER*, Snover et al., 2006). Other metrics include the number of reference words in the candidate (*Rword*, Cho et al., 2022) and semantic comparisons including synonyms (*CHAIR*, Rohrbach et al., 2018). Finally, some studies measure the precision and recall of specific word categories, including parts of speech (e.g., *Exact noun/verb overlap* Chan et al., 2023a), objects (Wang et al., 2021c) and named entities (common in news image captioning, e.g., Zhang and Wan, 2023).

Phrase-level semantic similarity: Others compare the semantics of sentence elements, most commonly using phrase embeddings, either with context (*BERTScore*: Zhang et al. 2019, *BERTScore++*: Yi et al. 2020) or without (*WMD*: Kusner et al. 2015, *NW*: Cornia et al. 2019a, *SSD*: Takmaz et al. 2020, *SemSim*: Nag Chowdhury et al. 2021, *Fuzzy noun/verb overlap*: Chan et al. 2023a). *VILBERTScore* (Lee et al., 2020) enriches phrase embeddings by injecting visual information. The SPICE family (*SPICE*: Anderson et al. 2016, *SPICE-U*: Wang et al. 2020d) compares the components of scene graphs of the candidate and reference captions. *ALOHa* (Petryk et al., 2024) prompts Large Language Models (LLMs) to identify object phrases in both candidate and references and computes the similarity of these phrases' embeddings.

Sentence-level similarity: Some metrics compare semantics at the sentence level. *MP-NetScore* (Black et al., 2024a) compares sentence level embeddings. Others (*CLAIR*: Chan

[2] We only document metrics that were used in the main body of the paper, omitting metrics used in the Appendix.

[3] One metric (Yngve score, Liu et al., 2019a) could not fit into any of the proposed categories, while another (Nubia) is assigned to two categories.

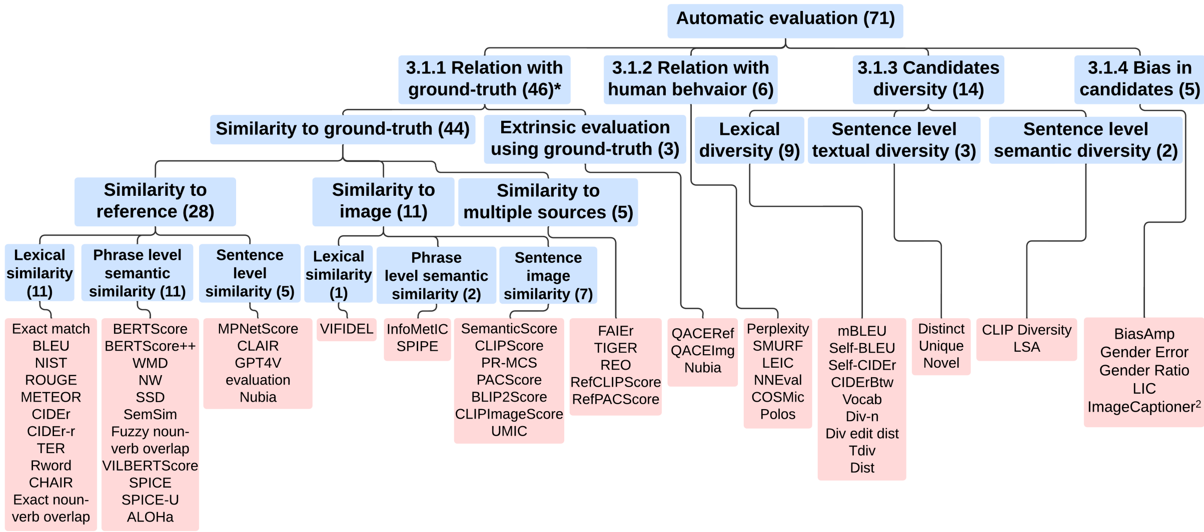


Figure 4: A taxonomy of automatic metrics for image evaluation. For each category, we indicate the number of metrics within that category in parentheses. Top categories are preceded by the relevant section number. Categories are in blue, metrics in red. * – Nubia is assigned to two distinct categories under *Relation with ground-truth*.

et al. 2023b, *GPT4V evaluation*: Ge et al. 2024a) prompt LLMs to compare the candidate and reference caption. *Nubia* (Kane et al., 2020) explicitly trains a model to evaluate sentence similarity.

Extracted information similarity: In specific domains (e.g., medical image captioning), it is common to extract relevant information from both the reference and candidate captions for comparison (e.g., a list of diseases from the captions in clinical image captioning, Nishino et al., 2022).

Similarity to image. Recent work suggests comparing the candidate caption to the image instead of a reference caption. These metrics are also known as reference-free metrics.

Lexical similarity: *VIFIDEL* (Madhyastha et al., 2019) compares object phrases in the candidate with names of objects identified in the image.

Phrase-level semantic similarity: *InfoMetIC* (Hu et al., 2023a) compares text and visual token embeddings. *SPIPE* (Xie et al., 2022) is a variation of SPICE where the scene graphs of the candidate and image are compared.

Sentence-image similarity: Several recent works embeded sentences and images in a joint visual-textual space and computed the similarity of candidate-image embeddings (*Semantic Score*: Dognin et al. 2019, *CLIPScore*: Hessel et al. 2021, *PR-MCS*: Kim et al. 2023, *PAC-Score*: Sarto et al. 2023, *BLIP2Score*: Zeng et al. 2024a). *CLIPImageScore* (Ge et al., 2024a) use text-to-image methods to generate an image based on the candidate and computes the embedding similarity between this generated image and the original image. *UMIC* (Lee et al., 2021b) explicitly trains models to predict similarity of candidate caption and image.

Retrieval-based methods: A popular evaluation protocol involves using text-to-image retrieval on the test set (e.g., Kornblith et al., 2023). In this setting, a candidate caption is matched with each image in the test set using an image-text matching model (most commonly CLIP), and the images are ranked by their matching score with the candidate. Then, recall@n is applied as the evaluation metric, with typical values for n being 1, 5, and 10.

Similarity to multiple sources. Some methods use both the image and reference captions as sources for comparison. *FAIEr* (Wang et al., 2021b) compares the candidate scene graph with a fusion of the image and reference scene graphs. *TIGER* (Jiang et al., 2019b) and *REO* (Jiang et al., 2019a) compare the candidate-image joint embedding vector with the reference-image joint embedding vector. *RefCLIPScore* (Hessel et al., 2021) and *RefPACScore* (Sarto et al., 2023) compute the harmonic mean of candidate-image similarity with candidate-reference similarity.

Extrinsic evaluation using ground-truth. Several NLP tasks focus on understanding the relation between two input sources. Some metrics use models trained for this task with the candidate and a ground-truth source as the inputs.

Question generation and question answering. Lee et al. (2021a) generate questions from the candidate and score it by how well a question-answering model answers using the reference caption (*QACERef*) or the image (*QACEImg*).

Natural language Inference (NLI). Nubia (Kane et al., 2020) use NLI models with the reference as the premise and the candidate as the hypothesis.

3.1.2 Relation With Human Behavior

The primary aim of captioning systems developed in recent years is to imitate human behavior, both in the text they generate and in the evaluation process. Metrics in this category strive to explicitly assess these properties.

Candidate fluency. Some works (e.g., Ou et al., 2023) evaluate the candidate's *perplexity* by employing an off-the-shelf large language model (commonly GPT-2, Radford et al., 2019) to compute the probability of the candidate's tokens. *SMURF* (Feinglass and Yang, 2021) computes the activation flow through a Transformer model self-attention layers, assumed to increase when the candidate differs greatly from typical captions.

Is the candidate human-like? *LEIC* (Cui et al., 2018) and *NNEval* (Sharif et al., 2018) measure how likely the candidate is to deceive a model that discriminates between human-generated and machine-generated sentences.

Human rating prediction. *COSMic* (Inan et al., 2021) and *Polos* (Wada et al., 2024) explicitly train a model to predict human scores.

3.1.3 Candidate Diversity

Several methods measure the diversity of generated captions. All assume a set of generated captions S for which diversity is measured, where S can either consist of candidates for the entire test set, a set of similar images or a single image.

Lexical diversity: While some works use lexical similarity methods such as BLEU (*mBLEU*: Shetty et al. 2017, *self-BLEU*: Zhu et al. 2018) and CIDEr (*Self-CIDEr*: Wang and Chan 2019, *CIDErBtw*: Wang et al. 2020a) on pairs of captions from S, others specifically designed methods to measure lexical diversity in S (*Vocab*: Shetty et al. 2017, *Div-n*: Shetty et al. 2017, *diversity edit distance*: Dai et al. 2018a, *Tdiv*: Liu et al. 2019a, *Dist*: Liu et al. 2019a).

Sentence-level textual diversity: Other methods compute the percentage of distinct candidates in S (*Distinct*, AKA *Unique*, Wang et al., 2017a), or the percentage of sentences not seen in the training set (*Novel*, Wang et al., 2017a).

Sentence-level semantic diversity: Recent methods propose to compare the semantics of candidates as an indicator for diversity. *LSA* (Wang and Chan, 2019) computes a matrix K where $K_{i,j}$ is the dot-product between the bag-of-words vectors of captions i, j in S, and calculates the singular vector decomposition (SVD) of K. *CLIP diversity* (Li et al., 2023b) computes the similarity of the CLIP embeddings of captions in S.

3.1.4 Bias in Candidates

Various metrics have been proposed to quantify the extent of bias evident in captioning models. *BiasAmp* (Zhao et al., 2017) measures the amplification of bias by the model compared to the training set by comparing the correlation between predefined attributes and activities (e.g., female-cooking) in model- and human-generated captions. Hendricks et al. (2018) introduce two metrics: *Gender error*, which assumes that images are labeled as male or female and calculates the number of gender misclassified words in the candidates, and *Gender ratio*, which defines male or female sentences based on the inclusion of predefined gender-related words and computes the ratio of male candidates to female candidates. Others measure bias amplification by training models to predict protected attributes given a caption (*LIC*, Hirota et al., 2022) or an image-caption pair (*ImageCaptioner*^{2}, Abdelrahman et al., 2024), and computing the difference between accuracies when training on candidates versus references.

3.2 Human Evaluation

While automatic evaluation offers clear benefits in terms of scale and consistency, it still serves as a surrogate for human evaluation. To ensure that the improvements shown by automatic methods are genuine, many image captioning studies also apply human evaluation methods to a subset of the data. Here, we categorize various frameworks for human evaluation.

Scale rating. Several studies direct human participants to evaluate candidates on a discrete scale based on various attributes, such as relevance (Maeda et al., 2023), fluency (Wu et al., 2023b) and descriptiveness (Yue et al., 2023).

Comparative. A different approach presents pairs of captions to human participants and asks them to decide which one is better without knowing their source. Early studies compared candidates to references (Kuznetsova et al., 2014; Yatskar et al., 2014), while recent research compares candidates with captions generated by baseline models (Tanaka et al., 2024; Ge et al., 2024a).

Yes/no questions. Certain studies involve human participants answering yes/no questions such as whether the candidate exhibits human-like qualities (Yao et al., 2019) or describes all the objects in the image (Chen et al., 2022).

Retrieval. Another popular approach is to ask participants to perform text-to-image retrieval given the candidate caption (Wang and Chan, 2019; Ou et al., 2023).

Answer questions using the candidate. Nie et al. (2020) prompts participants with questions and requires them to answer using the candidate provided, without showing the images.

4 Metrics Usage Analysis

In this section, we analyze the data from Section 2 to identify usage patterns and present the main findings, excluding 2024 as the data was collected before the year ended. Appendix B lists metric usage for all 314 reviewed papers.

4.1 Automatic Evaluation

A set of 5 metrics dominates. Figure 2 displays the number of papers using different metrics each year. Given the large number of metrics identified (71), we only plot usage for the seven most common metrics, clustering all other metrics by the categories from Section 3. Metrics used by no more than one paper are omitted.

The figure shows that since 2015, five metrics – BLEU, CIDEr, METEOR, ROUGE, and SPICE (henceforth, the five dominant metrics) – have been used substantially more frequently than all other metrics. This trend has begun to shift in recent years, with an increase in the use of other metrics starting from 2020. Notably, four of the dominant metrics (all except SPICE) are lexical similarity metrics, criticized for failing to capture semantic similarity (Giménez and Màrquez, 2007), and none consider the image during evaluation.

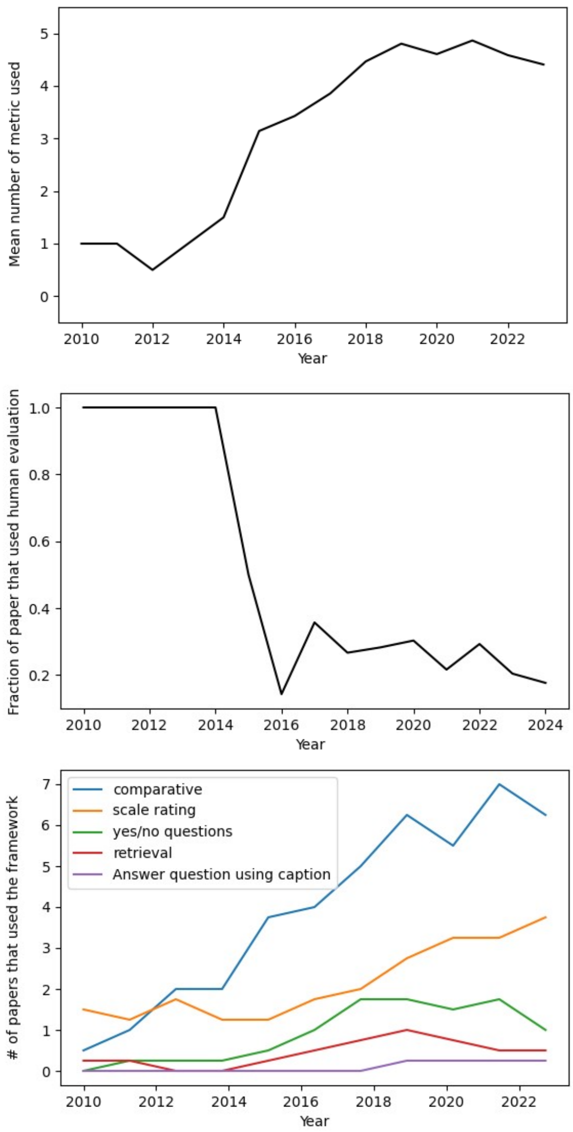


Figure 5: Metrics usage analysis plots. Top: Mean number of metrics used per paper each year. Center: Fraction of papers that performed human evaluation each year. Bottom: Human evaluation frameworks usage over the years, smoothed by convolving with a window size of four years.

The number of metrics per paper has settled at 4-5. Figure 5 (top) shows the mean number of metrics used per paper each year. There is a gradual increase until the mean number reached approximately 4.5 in 2019, remaining in the range of 4-5 since then.

4.2 Human Evaluation

Human evaluation usage is decreasing. Figure 5 (center) shows the fraction of papers with human evaluation each year. From 2010 to 2014, all papers conducted human evaluation. However, the publication of large image-caption datasets made this more challenging, and beginning at 2015 (a year after MSCOCO's release) there has been a general decline in the use of human evaluation, likely due to improvements in automatic evaluation metrics.

Comparative is the most common human framework. Figure 5 (bottom) presents the number of papers using each human evaluation framework per year. The comparative framework is most frequently used, closely followed by scale rating.

Significance and agreement are rarely reported. Out of the 94 papers that used human evaluation (29.9% of the papers we documented), only 19 reported significance. Reporting significance is crucial in human evaluation because, in most cases, only a small subset of the test set is assessed due to the high cost of human labor. Moreover, the results may be unreliable if the annotator agreement is low; however, only 6 of these 94 papers reported annotator agreement.

5 Experiments

Following previous work (Anderson et al., 2016; Hessel et al., 2021), we use correlation with human ratings as an indicator of metric quality. While inspired by previous work, our experiments include metrics not studied previously.

5.1 Experimental Setup

Datasets. We experiment with the following 7 human rating datasets.

Flickr8k-Expert and Flickr8k-CF (Hodosh et al., 2013) include human ratings for captions on the 1000 images of the Flickr8k test set. The captions are human-generated captions selected from the test set, where each image is associated with a set of captions (not necessarily initially generated for that image). Flickr8k-Expert includes 5,822 captions rated by a small and controlled group of native English speakers on a scale from 1 to 4. Flickr8k-CF comprises 47,830 captions, each rated by three or more crowd-sourced workers using a binary label (describes/doesn't describe the image). For each caption, we take the percentage of positive ratings as the final ground truth score.

Composite (Aditya et al., 2015) includes 11,985 human ratings for captions of 3,925 images from Flickr8k, Flickr30k and MSCOCO. Each image is paired with two model-generated and one original reference captions. Each caption is rated for correctness and thoroughness on a scale from 1 to 5. Following previous studies (Anderson et al., 2016), we use the correctness rating (also termed "relevance" in some prior research) as the score.

THumB (Kasai et al., 2022) features human ratings of captions for 500 images sourced from MSCOCO. Each image is associated with four model-generated and one original reference captions. Ratings cover various dimensions, such as precision, recall, and an overall score. We take the overall score as the score for this caption.

Polaris (Wada et al., 2024) contains 131,020 human ratings for captions generated for 13,691 images from MSCOCO and nocaps (Agrawal et al., 2019). The captions were generated by 10 different models. The dataset is partitioned into training, validation and test sets.

Pascal50S (Vedantam et al., 2015) contains 4000 pairs of candidate captions for images from the Pascal Sentence dataset (Rashtchian et al., 2010), evenly distributed across four categories: both captions are human-generated for the image (HC), both are human-generated but only one corresponds to the image (HI), one is human-generated while the other is machine-generated (HM), both are machine-generated (MM). Each pair has 48 human judgments indicating which of the two candidates is more similar to a reference caption. Following previous work (Hessel et al., 2021) we take the majority vote for each candidate pair (ties are broken randomly) and randomly select 5 references out of the available 48 for reference-based metrics.

The Reformulations dataset (introduced in a work currently under review) includes model-generated captions for 1,405 images from MSCOCO and Flickr30k, along with human-generated reformulations, i.e., corrected versions of the captions (if any errors exist). Evaluation metrics are expected to favor the caption after human correction. The dataset contains 5,208 caption-reformulation pairs; we omit 864 pairs where the reformulation is identical to the caption.

More details on the version we used for each dataset can be found in Appendix A.1.

Metrics selection. We now select the metrics to be evaluated out of the 71 identified metrics. First, we exclude bias metrics, as bias is not captured by the human ratings datasets. Next, we filter out all

Metric	Category	Coef
BLIP2Score	Image similarity	0.83
Polos	Human rating prediction	0.82
PACScore	Image similarity	0.29
Exact NO	Lexical similarity	0.08
BLEU1	Lexical similarity	0.07
Fuzzy VO	Phrase semantic similarity	0.02
BLEU4	Lexical similarity	0.02
CIDEr	Lexical similarity	-0.02
ROUGE	Lexical similarity	-0.07
RefCLIPScore	Multiple source similarity	-0.16

Table 2: Metrics selected by the feature selection algorithm to predict human ratings, along with their categories and linear regression coefficients.

metrics that receive multiple candidates as input (such as diversity metrics), as the human ratings are provided for a single candidate. Finally, we exclude all metrics that rely on a closed model API access (CLAIR, GPT4V evaluation, ALOHa), focusing on publicly accessible evaluation methods.

After filtering, 56 metrics remain. To ensure fair comparison, we conduct all the experiments for each metric ourselves, even if previous studies have reported results for this metric. Due to the effort required and the scope of this study, we focus on metrics frequently employed in image captioning research. Specifically, we select metrics used in at least two papers per year on average since their publication.[4] This narrows our selection to 20 metrics.[5] For details on the implementation of each metric, refer to Appendix A.2.

5.2 ENSEMBEVAL: An Ensemble of Evaluation Methods

We now examine whether a more diverse evaluation method, incorporating multiple existing metrics, outperforms individual metrics. We propose an ensemble approach we name ENSEMBEVAL, using a linear combination of these metrics.

Due to strong correlations among certain metrics, using all metrics in the ensemble might create redundancy. Therefore, we employ a sequential feature selection algorithm to identify a subset of metrics that accurately predict human ratings. At each step, the algorithm adds the best metric to the ensemble based on the cross-validation score of a linear regression estimator of human ratings. The process stops when the estimator's score does not increase by at least ε. We use the Polaris train set[6] for selecting the metrics and the validation set to determine the optimal ε value (0.0001).

Next, we find the metrics coefficients by training a linear regression model to predict human ratings, again using the Polaris train dataset. More details on our feature selection and linear regression procedures can be found in Appendix A.3.

Table 2 displays the selected metrics, their taxonomy categories, and their linear regression coefficients. The metrics span different categories, indicating that predicting human ratings could benefit from using different types of metrics. Interestingly, some metrics received a negative coefficient, most notably RefCLIPScore which is widely regarded as highly correlated with human ratings (Kasai et al., 2022).

Considering the frequent use of the five dominant metrics in prior research, we include an ensemble of these metrics (the dominant ensemble) as a baseline. As before, we determine the coefficients for this ensemble using linear regression.

5.3 Correlation with Human Ratings

We now compute the correlation between metric scores and human ratings across all datasets containing human ratings (all but Pascal50S and the Reformulations datasets).[7] We follow previous work (Anderson et al., 2016; Hessel et al., 2021; Kasai et al., 2022; Wada et al., 2024) and use Kendall-C correlation for Flickr8k-Expert, Composite and Polaris, Kendall-B correlation for Flickr8k-CF and Pearson correlation for THumB.

Table 3 presents the results. We omit results for five metrics[8] that were proposed ad hoc in an experiment (rather than in a dedicated paper) and showed weak performance. We will now discuss the findings derived from this data.

An ensemble improves over individual metrics. Our proposed ENSEMBEVAL achieves state-of-the-art results across all datasets, indicating that

[4]Not including the publication year and 2024.

[5]BLEU1, BLEU2, BLEU3, BLEU4, BLIP2Score, CIDEr, CLIPImageScore, CLIPScore, Exact noun overlap (NO), Exact verb overlap (VO), Fuzzy noun overlap (NO), Fuzzy verb overlap (VO), METEOR, MPNetScore, PACScore, ROUGE, RefCLIPScore, RefPACScore, SPICE, Polos. We omit two metrics for which we could not obtain a full implementation: PR-MCS and InfoMetIC.

[6]This is the only human rating dataset that is split into train/val/test.

[7]For Polaris we use the test set.

[8]CLIPImageScore, Exact NO, Exact VO, Fuzzy NO, Fuzzy VO.

Metric	Flickr8k (Expert)	Flickr8k (CF)	Composite	THumB	Polaris
BLEU4	30.8	16.9	30.6	10.4	46.2
BLEU3	31.5	17.0	30.9	11.8	46.6
BLEU2	32.5	17.9	31.1	15.8	47.1
BLEU1	32.3	17.9	31.4	19.5	45.4
ROUGE	32.3	19.9	32.5	18.7	46.2
METEOR	41.8	22.3	39.0	18.5	51.2
CIDEr	43.9	24.6	37.5	22.4	52.1
Dominant ensemble	43.7	23.3	38.5	23.1	52.3
SPICE	44.9	24.4	40.4	21.0	50.9
CLIPScore	51.4	34.4	53.8	31.9	51.5
PACScore	54.3	36.0	55.7	31.4	52.4
MPNetScore	54.7	35.3	54.7	40.4	53.6
RefCLIPScore	53.0	36.4	55.4	40.7	54.1
RefPACScore	55.9	37.6	57.3	42.4	55.2
BLIP2Score	52.5	36.7	61.5	44.9	53.7
Polos	56.4	37.8	58.0	43.4	57.8
EnsembEval (ours)	**58.5**	**38.7**	**61.7**	**46.6**	**58.8**
Δ from 2nd	+2.1	+0.9	+0.2	+1.7	+1.0

Table 3: Correlation with human ratings across different datasets. The best performing metric is in bold. Metrics are sorted by their average score.

integrating various aspects of caption quality yields a more "human-like" score.

Dominant metrics are weaker than recent alternatives. The five dominant metrics, as well as their ensemble, exhibit weaker correlations with human ratings compared to other metrics. Moreover, BLEU4, the most common BLEU variation, shows the weakest correlation with human ratings.

Lesser-known metrics prove to be valuable. While some metrics were introduced in dedicated papers, others were proposed ad hoc in the experiments section. Previous studies comparing metrics performance have only discussed the former, missing strong correlations with human ratings observed with ad hoc metrics like MPNetScore and BLIP2Score. This underscores the importance of a systematic review of all relevant papers.

5.4 Accuracy on Pairwise Comparison

We also perform a pairwise comparison task on the Pascal50S and Reformulations datasets to assess metrics' accuracy in assigning a higher score to the human-preferred candidate in each pair. For Pascal50S, we use majority vote to indicate human preference and report the mean and standard deviation across five random instances of tie-breaking and reference selection. For Reformulations, we consider the reformulated caption as preferred.

The results are presented in Table 4. EnsembEval attained the highest score on the Reformulations dataset, as well as when comparing human-generated to machine-generated captions in Pascal 50 (HM). In all other cases MPNetScore and BLIP2Score performed best. As noted earlier, these metrics were absent from previous studies since they weren't introduced in dedicated papers.

6 Related Work

6.1 Image Captioning Surveys

To the best of our knowledge, no previous work has exclusively surveyed image captioning evaluation methods. However, most image captioning surveys include a dedicated section for automatic evaluation methods, typically mentioning the five dominant metrics along with one or two additional ones (e.g., Hossain et al., 2019; Liu et al., 2019b; Ghandi et al., 2023; Sharma and Padha, 2023).

A few studies (Stefanini et al., 2022; Xu et al., 2023) provide a more detailed taxonomy of automatic metrics. Our work differs from them in two ways: First, we conduct a systematic review to gather all metric usage, identifying strong metrics not covered previously (as discussed in Section 5.3). Second, while these studies group metrics into loosely defined categories (e.g., "standard" metrics), we clearly define our categorization criteria based on the properties each metric

Metric	Pascal50S HC	HI	HM	MM	Mean	Reformulations
BLEU4	59.8 ± 0.7	92.3 ± 0.7	85.6 ± 0.5	57.3 ± 1.2	73.7 ± 0.5	49.8
BLEU3	60.6 ± 0.9	93.0 ± 0.7	88.1 ± 0.5	57.6 ± 1.0	74.8 ± 0.6	51.8
ROUGE	63.1 ± 0.8	95.2 ± 0.6	91.9 ± 0.3	60.3 ± 1.0	77.6 ± 0.4	53.1
BLEU2	62.9 ± 1.7	94.0 ± 0.4	90.1 ± 0.2	58.6 ± 1.0	76.4 ± 0.7	54.9
BLEU1	62.8 ± 1.3	95.1 ± 0.5	91.5 ± 0.4	59.8 ± 0.7	77.3 ± 0.2	56.0
CIDEr	65.6 ± 1.7	98.0 ± 0.4	90.9 ± 0.3	64.9 ± 1.1	79.8 ± 0.4	54.3
SPICE	61.2 ± 1.9	94.3 ± 0.7	85.5 ± 0.7	49.5 ± 0.7	72.6 ± 0.7	67.6
METEOR	64.4 ± 1.7	97.6 ± 0.4	94.1 ± 0.8	65.3 ± 0.9	80.4 ± 0.6	66.3
Dominant ensemble	65.5 ± 1.7	97.7 ± 0.3	93.0 ± 0.5	68.0 ± 0.7	81.1 ± 0.6	66.0
RefPACScore	67.6 ± 0.7	99.6 ± 0.1	96.0 ± 0.2	75.7 ± 0.6	84.7 ± 0.3	73.3
RefCLIPScore	64.1 ± 1.2	99.6 ± 0.1	95.8 ± 0.4	72.9 ± 0.6	83.1 ± 0.4	74.9
BLIP2Score	60.5 ± 0.2	$\mathbf{99.8 \pm 0.0}$	96.3 ± 0.0	71.2 ± 0.5	82.0 ± 0.1	76.1
Polos	69.5 ± 1.2	99.6 ± 0.0	97.0 ± 0.3	78.5 ± 0.6	86.1 ± 0.1	73.0
MPNetScore	$\mathbf{71.9 \pm 0.6}$	99.8 ± 0.1	96.3 ± 0.5	$\mathbf{79.0 \pm 0.6}$	$\mathbf{86.7 \pm 0.2}$	72.7
PACScore	60.4 ± 0.1	99.3 ± 0.0	96.8 ± 0.0	72.9 ± 0.4	82.4 ± 0.1	77.2
CLIPScore	56.1 ± 0.2	99.3 ± 0.0	96.3 ± 0.0	71.2 ± 0.4	80.7 ± 0.1	80.8
EnsembEval (ours)	68.7 ± 0.6	99.8 ± 0.1	$\mathbf{98.3 \pm 0.3}$	77.3 ± 0.3	86.0 ± 0.2	**81.4**

Table 4: Accuracy in pairwise comparison. In Pascal50S we report mean and standard deviation across five random instances of tie-breaking and reference selection. HC: both captions are human-generated for the image, HI: both are human-generated but only one corresponds to the image, HM: one caption is human-generated while the other is machine-generated, MM: both captions are machine-generated. In each dataset, best scoring metric is in bold. Metrics are sorted by their average score.

aims to measure.

We are also the first to develop a taxonomy for human evaluation frameworks. Before our study, Bernardi et al. (2016) discussed human evaluation but only mentioned the scale rating framework.

Prior to our study, Staniūtė and Šešok (2019) reported metric usage from selected papers; however, they manually selected popular papers rather than identifying them systematically. Similar to our work, two studies (Sharma, 2021; Al-Shamayleh et al., 2024) systematically identify image captioning papers and report metric usage. However, they cover much fewer papers (79 and 41, respectively) and metrics (3 and 8, respectively), and do not report trends or provide in-depth analysis.

6.2 Image Captioning Metrics Analysis

Several previous studies have analyzed the effectiveness of image captioning metrics. Hodosh et al. (2013) compare BLEU and ROUGE scores to human ratings and find that these metrics fall short in measuring content quality. Elliott and Keller (2014) compute the correlation of lexical similarity metrics with human ratings and find that METEOR had the strongest correlation. Kilickaya et al. (2017) compare the five dominant metrics and WMD, finding that n-gram based metrics exhibit lower performance than SPICE and WMD.

Recently, as reference-free metrics like CLIPScore gain prominence, there has been growing criticism directed at this line of research. Deutsch et al. (2022) argue that reference-free metrics can be exploited at test time to find outputs that maximize their scores. Ahmadi and Agrawal (2024) find these metrics sensitive to visual grounding errors but not to caption implausibility. Ma et al. (2024b) show that optimizing for reference-free metrics leads to unreadable captions.

7 Conclusion

We conducted a comprehensive survey of image captioning evaluation methods, finding that the vast majority of papers use only 5 metrics which are weakly correlated with human judgment. Instead, we propose EnsembEval, which leverages a diverse set of metrics in a data-driven manner to achieve state-of-the-art results.

Acknowledgements

This work was supported in part by the Israel Science Foundation (grant no. 2424/21), by a grant from the Israeli Planning and Budgeting Committee (PBD) and by the HUJI-UoM joint PhD pro-

gram.

References

Eslam Abdelrahman, Pengzhan Sun, Li Erran Li, and Mohamed Elhoseiny. 2024. Imagecaptioner2: Image captioner for image captioning bias amplification assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 20902–20911.

Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos. 2015. From images to sentences through scene description graphs using commonsense reasoning and knowledge. *arXiv preprint arXiv:1511.03292*.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Saba Ahmadi and Aishwarya Agrawal. 2024. An examination of the robustness of reference-free image captioning evaluation metrics. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 196–208, St. Julian's, Malta. Association for Computational Linguistics.

Hiba Ahsan, Daivat Bhatt, Kaivan Shah, and Nikita Bhalla. 2021. Multi-modal image captioning for the visually impaired. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 53–60, Online. Association for Computational Linguistics.

Ahmet Aker and Robert Gaizauskas. 2010. Generating image descriptions using dependency relational patterns. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1250–1258, Uppsala, Sweden. Association for Computational Linguistics.

Ahmad Sami Al-Shamayleh, Omar Adwan, Mohammad A. Alsharaiah, Abdelrahman H. Hussein, Qasem M. Kharma, and Christopher Ifeanyi Eke. 2024. A comprehensive literature review on image captioning methods and metrics based on deep learning technique. *Multimedia Tools and Applications*, 83(12):34219–34268.

Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone. 2020. Cross-modal coherence modeling for caption generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6535, Online. Association for Computational Linguistics.

Aliki Anagnostopoulou, Mareike Hartmann, and Daniel Sonntag. 2023. Towards adaptable and interactive image captioning with data augmentation and episodic memory. In *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 245–256, Toronto, Canada (Hybrid). Association for Computational Linguistics.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2017. Guided open vocabulary image captioning with constrained beam search. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 936–945, Copenhagen, Denmark. Association for Computational Linguistics.

Peter Anderson, Stephen Gould, and Mark Johnson. 2018a. Partially-supervised image captioning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018b. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Jyoti Aneja, Harsh Agrawal, Dhruv Batra, and Alexander Schwing. 2019. Sequential latent spaces for modeling the intention during diverse image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Jyoti Aneja, Aditya Deshpande, and Alexander G. Schwing. 2018. Convolutional image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Manuele Barraco, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2023. With a little help from your own past: Prototypical memory networks for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3021–3031.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442.

Romain Bielawski and Rufin VanRullen. 2023. CLIP-based image captioning via unsupervised cycle-consistency in the latent space. In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 266–275, Toronto, Canada. Association for Computational Linguistics.

Ali Furkan Biten, Lluis Gomez, Marcal Rusinol, and Dimosthenis Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alexander Black, Jing Shi, Yifei Fan, Tu Bui, and John Collomosse. 2024a. Vixen: Visual text comparison network for image difference captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 846–854.

Alexander Black, Jing Shi, Yifei Fan, Tu Bui, and John Collomosse. 2024b. Vixen: Visual text comparison network for image difference captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(2):846–854.

Emanuele Bugliarello and Desmond Elliott. 2021. The role of syntactic planning in compositional image captioning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 593–607, Online. Association for Computational Linguistics.

Michele Cafagna, Kees van Deemter, and Albert Gatt. 2022. Understanding cross-modal interactions in V&L models that generate scene descriptions. In *Proceedings of the Workshop on Unimodal and Multimodal Induction of Linguistic Structures (UM-IoS)*, pages 56–72, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 2022. 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16464–16473.

Tingjia Cao, Ke Han, Xiaomei Wang, Lin Ma, Yanwei Fu, Yu-Gang Jiang, and Xiangyang Xue. 2020. Feature deformation meta-networks in image captioning of novel objects. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):10494–10501.

David Chan, Austin Myers, Sudheendra Vijayanarasimhan, David Ross, and John Canny. 2023a. IC3: Image captioning by committee consensus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8975–9003, Singapore. Association for Computational Linguistics.

David Chan, Suzanne Petryk, Joseph Gonzalez, Trevor Darrell, and John Canny. 2023b. CLAIR: Evaluating image captions with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13638–13646, Singapore. Association for Computational Linguistics.

Arjun Chandrasekaran, Devi Parikh, and Mohit Bansal. 2018. Punny captions: Witty wordplay in image descriptions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 770–775, New Orleans, Louisiana. Association for Computational Linguistics.

Soravit Changpinyo, Bo Pang, Piyush Sharma, and Radu Soricut. 2019. Decoupled box proposal and featurization with ultrafine-grained semantic labels improve image captioning and visual question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1468–1474, Hong Kong, China. Association for Computational Linguistics.

Chen Chen, Shuai Mu, Wanpeng Xiao, Zexiong Ye, Liesi Wu, and Qi Ju. 2019a. Improving image captioning with conditional generative adversarial nets. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8142–8150.

Cheng-Kuan Chen, Zhufeng Pan, Ming-Yu Liu, and Min Sun. 2019b. Unsupervised stylish image description generation via domain layer norm. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8151–8158.

Fuhai Chen, Rongrong Ji, Jiayi Ji, Xiaoshuai Sun, Baochang Zhang, Xuri Ge, Yongjian Wu, Feiyue Huang, and Yan Wang. 2019c. Variational structured semantic inference for diverse image captioning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Fuhai Chen, Rongrong Ji, Xiaoshuai Sun, Yongjian Wu, and Jinsong Su. 2018a. Groupcap: Group-based image captioning with structured relevance and diversity constraints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh. 2018b. Attacking visual language grounding with adversarial examples: A case study on neural image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2587–2597, Melbourne, Australia. Association for Computational Linguistics.

Hui Chen, Guiguang Ding, Sicheng Zhao, and Jungong Han. 2018c. Temporal-difference learning with sampling baseline for image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Jia Chen and Qin Jin. 2020. Better captioning with sequence-level exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jia Chen, Yike Wu, Shiwan Zhao, and Qin Jin. 2021a. Language resource efficient learning for captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1887–1895, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jianfu Chen, Polina Kuznetsova, David Warren, and Yejin Choi. 2015. Déjà image-captions: A corpus of expressive descriptions in repetition. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 504–514, Denver, Colorado. Association for Computational Linguistics.

Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. 2022. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18030–18040.

Long Chen, Zhihong Jiang, Jun Xiao, and Wei Liu. 2021b. Human-like controllable image captioning with verb-specific semantic roles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16846–16856.

Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. 2017a. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Minghai Chen, Guiguang Ding, Sicheng Zhao, Hui Chen, Qiang Liu, and Jungong Han. 2017b. Reference based lstm for image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Shi Chen and Qi Zhao. 2018. Boosted attention: Leveraging human attention for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. 2020. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, Gang Yu, and Tao Chen. 2023a. End-to-end 3d dense captioning with vote2cap-detr. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11124–11133.

Tianlang Chen, Zhongping Zhang, Quanzeng You, Chen Fang, Zhaowen Wang, Hailin Jin, and Jiebo Luo. 2018d. "factual" or "emotional": Stylized image captioning with adaptive learning and attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Xinlei Chen and C. Lawrence Zitnick. 2015. Mind's eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xinpeng Chen, Lin Ma, Wenhao Jiang, Jian Yao, and Wei Liu. 2018e. Regularizing rnns for caption generation by reconstructing the past with the present. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhenyu Chen, Ali Gholami, Matthias Niessner, and Angel X. Chang. 2021c. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3193–3203.

Zhenyu Chen, Ronghang Hu, Xinlei Chen, Matthias Nießner, and Angel X. Chang. 2023b. Unit3d: A unified transformer for 3d dense captioning and visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18109–18119.

Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. 2022. Fine-grained image captioning with CLIP reward. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 517–527, Seattle, United States. Association for Computational Linguistics.

Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. 2017. Attend to you: Personalized image captioning with context sequence memory networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 2018. Pragmatically informative image captioning with character-level inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 439–443, New Orleans, Louisiana. Association for Computational Linguistics.

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2019a. Show, control and tell: A framework for generating controllable and grounded captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8307–8316.

Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2019b. Show, control and tell: A framework for generating controllable and grounded captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. 2018. Learning to evaluate image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bo Dai, Sanja Fidler, and Dahua Lin. 2018a. A neural compositional paradigm for image captioning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. 2017. Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Bo Dai and Dahua Lin. 2017. Contrastive learning for image captioning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Bo Dai, Deming Ye, and Dahua Lin. 2018b. Rethinking the form of latent states in image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Riccardo Del Chiaro, Bartł omiej Twardowski, Andrew Bagdanov, and Joost van de Weijer. 2020. Ratt: Recurrent attention to transient tasks for continual image captioning. In *Advances in Neural Information Processing Systems*, volume 33, pages 16736–16748. Curran Associates, Inc.

Chaorui Deng, Ning Ding, Mingkui Tan, and Qi Wu. 2020. Length-controllable image captioning. In *Computer Vision – ECCV 2020*, pages 712–729, Cham. Springer International Publishing.

Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander G. Schwing, and David Forsyth. 2019. Fast, diverse and accurate image captioning guided by part-of-speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Roberto Dessì, Michele Bevilacqua, Eleonora Gualdoni, Nathanaël Carraz Rakotonirina, Francesca Franzon, and Marco Baroni. 2023. Cross-domain image captioning with discriminative finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6935–6944.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. On the limitations of reference-free evaluations of generated text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10960–10977, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 100–105, Beijing, China. Association for Computational Linguistics.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145.

Pierre Dognin, Igor Melnyk, Youssef Mroueh, Jerret Ross, and Tom Sercu. 2019. Adversarial semantic alignment for improved image captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Desmond Elliott and Frank Keller. 2013. Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1292–1302, Seattle, Washington, USA. Association for Computational Linguistics.

Desmond Elliott and Frank Keller. 2014. Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 452–457, Baltimore, Maryland. Association for Computational Linguistics.

Jiashuo Fan, Yaoyuan Liang, Leyao Liu, Shaolun Huang, and Lei Zhang. 2023. Rca-noc: Relative contrastive alignment for novel object captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15510–15520.

Zhihao Fan, Zhongyu Wei, Siyuan Wang, and Xuanjing Huang. 2019. Bridging by word: Image grounded vocabulary construction for visual captioning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6514–6524, Florence, Italy. Association for Computational Linguistics.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollar, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lin Liang, Zhe Gan, Lijuan Wang, Yezhou Yang, and Zicheng Liu. 2022. Injecting semantic concepts into end-to-end image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18009–18019.

Junjie Fei, Teng Wang, Jinrui Zhang, Zhenyu He, Chengjie Wang, and Feng Zheng. 2023a. Transferable decoding with visual entities for zero-shot image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3136–3146.

Zhengcong Fei. 2021a. Memory-augmented image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):1317–1324.

Zhengcong Fei. 2021b. Partially non-autoregressive image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):1309–1316.

Zhengcong Fei. 2022. Attention-aligned transformer for image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1):607–615.

Zhengcong Fei, Mingyuan Fan, Li Zhu, Junshi Huang, Xiaoming Wei, and Xiaolin Wei. 2023b. Uncertainty-aware image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1):614–622.

Zhengcong Fei, Xu Yan, Shuhui Wang, and Qi Tian. 2022. Deecap: Dynamic early exiting for efficient image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12216–12226.

Joshua Feinglass and Yezhou Yang. 2021. SMURF: SeMantic and linguistic UndeRstanding fusion for caption evaluation via typicality analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2250–2260, Online. Association for Computational Linguistics.

Steven Y. Feng, Kevin Lu, Zhuofu Tao, Malihe Alikhani, Teruko Mitamura, Eduard Hovy, and Varun Gangal. 2022. Retrieve, caption, generate: Visual grounding for enhancing commonsense in text generation models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10618–10626.

Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. 2019. Unsupervised image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yansong Feng and Mirella Lapata. 2010. How many words is a picture worth? automatic caption generation for news images. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1239–1249, Uppsala, Sweden. Association for Computational Linguistics.

Zhongtian Fu, Kefei Song, Luping Zhou, and Yang Yang. 2024. Noise-aware image captioning with progressively exploring mismatched words. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(11):12091–12099.

Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017a. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. 2017b. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Junlong Gao, Shiqi Wang, Shanshe Wang, Siwei Ma, and Wen Gao. 2019a. Self-critical n-step training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lianli Gao, Kaixuan Fan, Jingkuan Song, Xianglong Liu, Xing Xu, and Heng Tao Shen. 2019b. Deliberate attention networks for image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8320–8327.

Yiqi Gao, Xinglin Hou, Yuanmeng Zhang, Tiezheng Ge, Yuning Jiang, and Peng Wang. 2022. CapOnImage: Context-driven dense-captioning on image. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3449–3465, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hongwei Ge, Zehang Yan, Kai Zhang, Mingde Zhao, and Liang Sun. 2019. Exploring overall contextual information for image captioning in human-like cognitive style. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Yunhao Ge, Xiaohui Zeng, Jacob Samuel Huffman, Tsung-Yi Lin, Ming-Yu Liu, and Yin Cui. 2024a. Visual fact checker: Enabling high-fidelity detailed caption generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14033–14042.

Yunhao Ge, Xiaohui Zeng, Jacob Samuel Huffman, Tsung-Yi Lin, Ming-Yu Liu, and Yin Cui. 2024b. Visual fact checker: Enabling high-fidelity detailed caption generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14033–14042.

Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar. 2023. Deep learning approaches on image captioning: A review. *ACM Computing Surveys*, 56(3):1–39.

Jesús Giménez and Lluís Màrquez. 2007. Linguistic features for automatic evaluation of heterogenous MT systems. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 256–264, Prague, Czech Republic. Association for Computational Linguistics.

Jiuxiang Gu, Jianfei Cai, Gang Wang, and Tsuhan Chen. 2018a. Stack-captioning: Coarse-to-fine learning for image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Jiuxiang Gu, Shafiq Joty, Jianfei Cai, and Gang Wang. 2018b. Unpaired image captioning by language pivoting. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. 2019. Unpaired image captioning via scene graph alignments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Jiuxiang Gu, Gang Wang, Jianfei Cai, and Tsuhan Chen. 2017. An empirical study of language cnn for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Longteng Guo, Jing Liu, Peng Yao, Jiangwei Li, and Hanqing Lu. 2019. Mscap: Multi-style image captioning with unpaired stylized text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Longteng Guo, Jing Liu, Xinxin Zhu, Peng Yao, Shichen Lu, and Hanqing Lu. 2020. Normalized and geometry-aware self-attention network for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tszhang Guo, Shiyu Chang, Mo Yu, and Kun Bai. 2018. Improving reinforcement learning based image captioning with natural language prior. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 751–756, Brussels, Belgium. Association for Computational Linguistics.

Zixin Guo, Tzu-Jui Wang, and Jorma Laaksonen. 2022. CLIP4IDC: CLIP for image difference captioning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 33–42, Online only. Association for Computational Linguistics.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning images taken by people who are blind. In *Computer Vision – ECCV 2020*, pages 417–434, Cham. Springer International Publishing.

Sen He, Hamed R. Tavakoli, Ali Borji, and Nicolas Pugeault. 2019. Human attention in image captioning: Dataset and analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. 2016. Deep compositional captioning: Describing novel object categories without paired training data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image captioning: Transforming objects into words. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yusuke Hirota, Yuta Nakashima, and Noa Garcia. 2022. Quantifying societal bias amplification in image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13450–13459.

Yusuke Hirota, Yuta Nakashima, and Noa Garcia. 2023. Model-agnostic gender debiased image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15191–15200.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Ukyo Honda, Yoshitaka Ushiku, Atsushi Hashimoto, Taro Watanabe, and Yuji Matsumoto. 2021. Removing word-level spurious alignment between images and pseudo-captions in unsupervised image captioning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3692–3702, Online. Association for Computational Linguistics.

MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36.

Mehrdad Hosseinzadeh and Yang Wang. 2021. Image change captioning by learning from an auxiliary task. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR), pages 2725–2734.

Jingyi Hou, Xinxiao Wu, Xiaoxun Zhang, Yayun Qi, Yunde Jia, and Jiebo Luo. 2020. Joint commonsense and relation reasoning for image and video captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):10973–10980.

Anwen Hu, Shizhe Chen, Liang Zhang, and Qin Jin. 2023a. InfoMetIC: An informative metric for reference-free image caption evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3171–3185, Toronto, Canada. Association for Computational Linguistics.

Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. Scaling up vision-language pretraining for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17980–17989.

Xiaowei Hu, Xi Yin, Kevin Lin, Lei Zhang, Jianfeng Gao, Lijuan Wang, and Zicheng Liu. 2021. Vivo: Visual vocabulary pre-training for novel object captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):1575–1583.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A. Smith, and Jiebo Luo. 2023b. Promptcap: Prompt-guided image captioning for vqa with gpt-3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2963–2975.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019a. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Lun Huang, Wenmin Wang, Yaxian Xia, and Jie Chen. 2019b. Adaptively aligned image captioning via adaptive attention time. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Qiuyuan Huang, Pengchuan Zhang, Dapeng Wu, and Lei Zhang. 2018. Turbo learning for captionbot and drawingbot. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Xiaoke Huang, Jianfeng Wang, Yansong Tang, Zheng Zhang, Han Hu, Jiwen Lu, Lijuan Wang, and Zicheng Liu. 2024. Segment and caption anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13405–13417.

EunJeong Hwang and Vered Shwartz. 2023. MemeCap: A dataset for captioning and interpreting memes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1445, Singapore. Association for Computational Linguistics.

Mert Inan, Piyush Sharma, Baber Khalid, Radu Soricut, Matthew Stone, and Malihe Alikhani. 2021. COSMic: A coherence-aware generation metric for image descriptions. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3419–3430, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiayi Ji, Yunpeng Luo, Xiaoshuai Sun, Fuhai Chen, Gen Luo, Yongjian Wu, Yue Gao, and Rongrong Ji. 2021. Improving image captioning by leveraging intra- and inter-layer global representation in transformer network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):1655–1663.

Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. 2015. Guiding the long-short term memory model for image caption generation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Ming Jiang, Junjie Hu, Qiuyuan Huang, Lei Zhang, Jana Diesner, and Jianfeng Gao. 2019a. REO-relevance, extraness, omission: A fine-grained evaluation for image captioning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1475–1480, Hong Kong, China. Association for Computational Linguistics.

Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. 2019b. TIGEr: Text-to-image grounding for image caption evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2141–2152, Hong Kong, China. Association for Computational Linguistics.

Wenhao Jiang, Lin Ma, Xinpeng Chen, Hanwang Zhang, and Wei Liu. 2018a. Learning to guide decoding for image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. 2018b. Recurrent fusion network for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Yang Jiao, Shaoxiang Chen, Zequn Jie, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. 2022. More: Multi-order relation mining for dense captioning in 3d scenes. In *Computer Vision – ECCV 2022*, pages 528–545, Cham. Springer Nature Switzerland.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajanoh, and Mohamed Coulibali. 2020. NUBIA: NeUral based interchangeability assessor for text generation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 28–37, Online (Dublin, Ireland). Association for Computational Linguistics.

Wooyoung Kang, Jonghwan Mun, Sungjun Lee, and Byungseok Roh. 2023. Noise-aware learning from web-crawled image-text data for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2942–2952.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jungo Kasai, Keisuke Sakaguchi, Lavinia Dunagan, Jacob Morrison, Ronan Le Bras, Yejin Choi, and Noah A. Smith. 2022. Transparent human evaluation for image captioning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3464–3478, Seattle, United States. Association for Computational Linguistics.

Lei Ke, Wenjie Pei, Ruiyu Li, Xiaoyong Shen, and Yu-Wing Tai. 2019. Reflective decoding network for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2017. Re-evaluating automatic metrics for image captioning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 199–209, Valencia, Spain. Association for Computational Linguistics.

Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. 2019a. Dense relational captioning: Triple-stream networks for relationship-based captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. 2019b. Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2012–2023, Hong Kong, China. Association for Computational Linguistics.

Hoeseong Kim, Jongseok Kim, Hyungseok Lee, Hyunsung Park, and Gunhee Kim. 2021a. Viewpoint-agnostic change captioning with cycle consistency. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2095–2104.

Hyounghun Kim, Abhay Zala, Graham Burri, and Mohit Bansal. 2021b. Fixmypose: Pose correctional captioning and retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):13161–13170.

Yongil Kim, Yerin Hwang, Hyeongu Yun, Seunghyun Yoon, Trung Bui, and Kyomin Jung. 2023. PR-MCS: Perturbation robust metric for MultiLingual image captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12237–12258, Singapore. Association for Computational Linguistics.

Simon Kornblith, Lala Li, Zirui Wang, and Thao Nguyen. 2023. Guiding image captioning models toward more specific captions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15259–15269.

Chia-Wen Kuo and Zsolt Kira. 2022. Beyond a pre-trained object detector: Cross-modal textual and visual context for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17969–17979.

Chia-Wen Kuo and Zsolt Kira. 2023. Haav: Hierarchical aggregation of augmented views for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11039–11049.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France. PMLR.

Polina Kuznetsova, Vicente Ordonez, Alexander Berg, Tamara Berg, and Yejin Choi. 2012. Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 359–368, Jeju Island, Korea. Association for Computational Linguistics.

Polina Kuznetsova, Vicente Ordonez, Alexander Berg, Tamara Berg, and Yejin Choi. 2013. Generalizing image captions for image-text parallel corpus. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 790–796, Sofia, Bulgaria. Association for Computational Linguistics.

Polina Kuznetsova, Vicente Ordonez, Tamara L. Berg, and Yejin Choi. 2014. TreeTalk: Composition and compression of trees for image descriptions. *Transactions of the Association for Computational Linguistics*, 2:351–362.

Iro Laina, Christian Rupprecht, and Nassir Navab. 2019. Towards unsupervised image captioning with shared multimodal embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Remi Lebret, Pedro Pinheiro, and Ronan Collobert. 2015. Phrase-based image captioning. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2085–2094, Lille, France. PMLR.

Hwanhee Lee, Thomas Scialom, Seunghyun Yoon, Franck Dernoncourt, and Kyomin Jung. 2021a. QACE: Asking questions to evaluate an image caption. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4631–4638, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Trung Bui, and Kyomin Jung. 2021b. UMIC: An unreferenced metric for image captioning via contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 220–226, Online. Association for Computational Linguistics.

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. 2020. ViLBERTScore: Evaluating image caption using vision-and-language BERT. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages

34–39, Online. Association for Computational Linguistics.

Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C.H. Hoi. 2023a. LAVIS: A one-stop library for language-vision intelligence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 31–41, Toronto, Canada. Association for Computational Linguistics.

Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. 2019a. Entangled transformer for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Jiaxuan Li, Duc Minh Vo, Akihiro Sugimoto, and Hideki Nakayama. 2024a. Evcap: Retrieval-augmented image captioning with external visual-name memory for open-world comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13733–13742.

Linghui Li, Sheng Tang, Lixi Deng, Yongdong Zhang, and Qi Tian. 2017. Image caption with global-local attention. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Nannan Li, Zhenzhong Chen, and Shan Liu. 2019b. Meta learning for image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8626–8633.

Runjia Li, Shuyang Sun, Mohamed Elhoseiny, and Philip Torr. 2023b. Oxfordtvg-hic: Can machine make humorous captions from images? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20293–20303.

Siming Li, Girish Kulkarni, Tamara L Berg, Alexander C Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 220–228, Portland, Oregon, USA. Association for Computational Linguistics.

Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. 2023c. Decap: Decoding clip latents for zero-shot captioning via text-only training.

Wenyan Li, Jonas Lotz, Chen Qiu, and Desmond Elliott. 2024b. The role of data curation in image captioning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1074–1088, St. Julian's, Malta. Association for Computational Linguistics.

Xiangyang Li, Shuqiang Jiang, and Jungong Han. 2019c. Learning object context for dense captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8650–8657.

Yehao Li, Yingwei Pan, Ting Yao, and Tao Mei. 2022. Comprehending and ordering semantics for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17990–17999.

Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. 2019d. Pointing novel objects in image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Bing Liu, Dong Wang, Xu Yang, Yong Zhou, Rui Yao, Zhiwen Shao, and Jiaqi Zhao. 2022. Show, deconfound and tell: Image captioning with causal inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18041–18050.

Chenxi Liu, Junhua Mao, Fei Sha, and Alan Yuille. 2017a. Attention correctness in neural image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Fenglin Liu, Xuancheng Ren, Yuanxin Liu, Houfeng Wang, and Xu Sun. 2018a. simNet: Stepwise image-topic merging network for generating detailed and comprehensive image captions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 137–149, Brussels, Belgium. Association for Computational Linguistics.

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2021. Visual news: Benchmark and challenges in news image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6761–6771, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Junhao Liu, Kai Wang, Chunpu Xu, Zhou Zhao, Ruifeng Xu, Ying Shen, and Min Yang. 2020. Interactive dual generative adversarial networks for image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11588–11595.

Lixin Liu, Jiajun Tang, Xiaojun Wan, and Zongming Guo. 2019a. Generating diverse and descriptive image captions using visual paraphrases. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. 2017b. Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Xiaoxiao Liu, Qingyang Xu, and Ning Wang. 2019b. A survey on deep neural network-based image captioning. *The Visual Computer*, 35(3):445–470.

Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. 2018b. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Zhiyue Liu, Jinyuan Liu, and Fanrong Ma. 2024. Improving cross-modal alignment with synthetic pairs for text-only image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(4):3864–3872.

Di Lu, Spencer Whitehead, Lifu Huang, Heng Ji, and Shih-Fu Chang. 2018. Entity-aware image caption generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4013–4023, Brussels, Belgium. Association for Computational Linguistics.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jianjie Luo, Yehao Li, Yingwei Pan, Ting Yao, Jianlin Feng, Hongyang Chao, and Tao Mei. 2023a. Semantic-conditional diffusion networks for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23359–23368.

Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. 2018. Discriminability objective for training descriptive captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. 2023b. Scalable 3d captioning with pretrained models. In *Advances in Neural Information Processing Systems*, volume 36, pages 75307–75337. Curran Associates, Inc.

Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong Ji. 2021. Dual-level collaborative transformer for image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(3):2286–2293.

Chih-Yao Ma, Yannis Kalantidis, Ghassan AlRegib, Peter Vajda, Marcus Rohrbach, and Zsolt Kira. 2020. Learning to generate grounded visual captions without localization supervision. In *Computer Vision – ECCV 2020*, pages 353–370, Cham. Springer International Publishing.

Feipeng Ma, Yizhou Zhou, Fengyun Rao, Yueyi Zhang, and Xiaoyan Sun. 2024a. Image captioning with multi-context synthetic data. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5):4089–4097.

Zheng Ma, Changxin Wang, Yawen Ouyang, Fei Zhao, Jianbing Zhang, Shujian Huang, and Jiajun Chen. 2024b. Cobra effect in reference-free image captioning metrics. *arXiv preprint arXiv:2402.11572*.

Ziping Ma, Furong Xu, Jian Liu, Ming Yang, and Qingpei Guo. 2024c. Sycoca: Symmetrizing contrastive captioners with attentive masking for multimodal alignment. *arXiv preprint arXiv:2401.02137*.

Pranava Madhyastha, Josiah Wang, and Lucia Specia. 2019. VIFIDEL: Evaluating the visual fidelity of image descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6539–6550, Florence, Italy. Association for Computational Linguistics.

Pranava Swaroop Madhyastha, Josiah Wang, and Lucia Specia. 2018. End-to-end image captioning exploits distributional similarity in multimodal space. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 381–383, Brussels, Belgium. Association for Computational Linguistics.

Koki Maeda, Shuhei Kurita, Taiki Miyanishi, and Naoaki Okazaki. 2023. Query-based image captioning from multi-context 360$cdegree$ images. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6940–6954, Singapore. Association for Computational Linguistics.

Shweta Mahajan and Stefan Roth. 2020. Diverse image captioning with context-object split latent spaces. In *Advances in Neural Information Processing Systems*, volume 33, pages 3613–3624. Curran Associates, Inc.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2014. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*.

Rebecca Mason and Eugene Charniak. 2014a. Domain-specific image captioning. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 11–20, Ann Arbor, Michigan. Association for Computational Linguistics.

Rebecca Mason and Eugene Charniak. 2014b. Nonparametric method for data-driven image captioning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 592–598, Baltimore, Maryland. Association for Computational Linguistics.

Alexander Mathews, Lexing Xie, and Xuming He. 2016. Senticap: Generating image descriptions with sentiments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Alexander Mathews, Lexing Xie, and Xuming He. 2018. Semstyle: Learning to generate stylised image captions using unaligned text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Effrosyni Mavroudi and René Vidal. 2022. Weakly-supervised generation and grounding of visual descriptions with conditional generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15544–15554.

Luke Melas-Kyriazi, Alexander Rush, and George Han. 2018. Training for diversity in image paragraph captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 757–761, Brussels, Belgium. Association for Computational Linguistics.

Zihang Meng, David Yang, Xuefei Cao, Ashish Shah, and Ser-Nam Lim. 2022. Object-centric unsupervised image captioning. In *Computer Vision – ECCV 2022*, pages 219–235, Cham. Springer Nature Switzerland.

Victor Milewski, Marie-Francine Moens, and Iacer Calixto. 2020. Are scene graphs good enough to improve image captioning? In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 504–515, Suzhou, China. Association for Computational Linguistics.

Suvir Mirchandani, Licheng Yu, Mengjiao Wang, Animesh Sinha, Wenwen Jiang, Tao Xiang, and Ning Zhang. 2022. FaD-VLP: Fashion vision-and-language pre-training towards unified retrieval and captioning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10484–10497, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Xufeng Han, Alyssa Mensch, Alex Berg, Tamara Berg, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756, Avignon, France. Association for Computational Linguistics.

Abdelrahman Mohamed, Fakhraddin Alwajih, El Moatez Billah Nagoudi, Alcides Inciarte, and Muhammad Abdul-Mageed. 2023. Violet: A vision-language model for Arabic image captioning with gemini decoder. In *Proceedings of ArabicNLP 2023*, pages 1–11, Singapore (Hybrid). Association for Computational Linguistics.

Youssef Mohamed, Faizan Farooq Khan, Kilichbek Haydarov, and Mohamed Elhoseiny. 2022. It is okay to not be okay: Overcoming emotional bias in affective image captioning by contrastive data collection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21263–21272.

Jonghwan Mun, Minsu Cho, and Bohyung Han. 2017. Text-guided attention model for image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Sreyasi Nag Chowdhury, Rajarshi Bhowmik, Hareesh Ravi, Gerard de Melo, Simon Razniewski, and Gerhard Weikum. 2021. Exploiting image–text synergy for contextual image captioning. In *Proceedings of the Third Workshop on Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, pages 30–37, Kyiv, Ukraine. Association for Computational Linguistics.

Shunta Nagasawa, Yotaro Watanabe, and Hitoshi Iyatomi. 2021. Validity-based sampling and smoothing methods for multiple reference image captioning. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 36–41, Mexico City, Mexico. Association for Computational Linguistics.

Edwin G. Ng, Bo Pang, Piyush Sharma, and Radu Soricut. 2021. Understanding guided image captioning performance across domains. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 183–193, Online. Association for Computational Linguistics.

Khanh Nguyen, Ali Furkan Biten, Andres Mafla, Lluis Gomez, and Dimosthenis Karatzas. 2023. Show, interpret and tell: Entity-aware contextualised image captioning in wikipedia. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):1940–1948.

Kien Nguyen, Subarna Tripathi, Bang Du, Tanaya Guha, and Truong Q. Nguyen. 2021. In defense of scene graphs for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1407–1416.

Minh Thu Nguyen, Duy Phung, Minh Hoai, and Thien Huu Nguyen. 2020. Structural and functional decomposition for personality image captioning in a communication game. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4587–4593, Online. Association for Computational Linguistics.

Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. 2022. Grit: Faster and better image captioning transformer using dual visual features. In *Computer Vision – ECCV 2022*, pages 167–184, Cham. Springer Nature Switzerland.

Allen Nie, Reuben Cohn-Gordon, and Christopher Potts. 2020. Pragmatic issue-sensitive image captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1924–1938, Online. Association for Computational Linguistics.

Mitja Nikolaus, Mostafa Abdou, Matthew Lamm, Rahul Aralikatte, and Desmond Elliott. 2019. Compositional generalization in image captioning. In *Proceedings of the 23rd Conference* *on Computational Natural Language Learning (CoNLL)*, pages 87–98, Hong Kong, China. Association for Computational Linguistics.

Toru Nishino, Yasuhide Miura, Tomoki Taniguchi, Tomoko Ohkuma, Yuki Suzuki, Shoji Kido, and Noriyuki Tomiyama. 2022. Factual accuracy is not enough: Planning consistent description order for radiology report generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7123–7138, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Nukrai, Ron Mokady, and Amir Globerson. 2022. Text-only training for image captioning using noise-injected CLIP. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4055–4063, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Jiefu Ou, Benno Krojer, and Daniel Fried. 2023. Pragmatic inference with a CLIP listener for contrastive captioning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1904–1917, Toronto, Canada. Association for Computational Linguistics.

Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. 2020. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

George Pantazopoulos, Alessandro Suglia, and Arash Eshghi. 2022. Combine to describe: Evaluating compositional generalization in image captioning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 115–131, Dublin, Ireland. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Dong Huk Park, Trevor Darrell, and Anna Rohrbach. 2019. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Hyeryun Park, Kyungmo Kim, Jooyoung Yoon, Seongkeun Park, and Jinwook Choi. 2020. Feature Difference Makes Sense: A medical image captioning model exploiting feature difference and tag information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 95–102, Online. Association for Computational Linguistics.

Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek. 2017. Areas of attention for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Suzanne Petryk, David Chan, Anish Kachinthaya, Haodi Zou, John Canny, Joseph Gonzalez, and Trevor Darrell. 2024. ALOHa: A new measure for hallucination in captioning models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 342–357, Mexico City, Mexico. Association for Computational Linguistics.

Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. 2016. Variational autoencoder for deep learning of images, labels and captions. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Yayun Qi, Wentian Zhao, and Xinxiao Wu. 2024. Relational distant supervision for image captioning without image-text pairs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5):4524–4532.

Yu Qin, Jiajun Du, Yonghua Zhang, and Hongtao Lu. 2019. Look back and predict forward in image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Haoyi Qiu, Zi-Yi Dou, Tianlu Wang, Asli Celikyilmaz, and Nanyun Peng. 2023. Gender biases in automatic evaluation metrics for image captioning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8358–8375, Singapore. Association for Computational Linguistics.

Longtian Qiu, Shan Ning, and Xuming He. 2024. Mining fine-grained image-text alignment for zero-shot captioning via text-only training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5):4605–4613.

Tingyu Qu, Tinne Tuytelaars, and Marie-Francine Moens. 2024. Visually-aware context modeling for news image captioning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2927–2943, Mexico City, Mexico. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Abisek Rajakumar Kalarani, Pushpak Bhattacharyya, Niyati Chhaya, and Sumit Shekhar. 2023. "let's not quote out of context": Unified vision-language pretraining for context assisted image captioning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 695–706, Toronto, Canada. Association for Computational Linguistics.

Rita Ramos, Emanuele Bugliarello, Bruno Martins, and Desmond Elliott. 2024. PAELLA: Parameter-efficient lightweight language-agnostic captioning model. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3549–3564, Mexico City, Mexico. Association for Computational Linguistics.

Rita Ramos, Desmond Elliott, and Bruno Martins. 2023a. Retrieval-augmented image captioning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3666–3681, Dubrovnik, Croatia. Association for Computational Linguistics.

Rita Ramos, Bruno Martins, and Desmond Elliott. 2023b. LMCap: Few-shot multilingual image captioning by retrieval augmented language model prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1635–1651, Toronto, Canada. Association for Computational Linguistics.

Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. 2023c. Smallcap: Lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2840–2849.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147, Los Angeles. Association for Computational Linguistics.

Yuchen Ren, Zhendong Mao, Shancheng Fang, Yan Lu, Tong He, Hao Du, Yongdong Zhang, and Wanli Ouyang. 2023. Crossing the gap: Domain generalization for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2871–2880.

Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. 2017. Deep reinforcement learning-based image captioning with embedding reward. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

Anna Rohrbach, Marcus Rohrbach, Siyu Tang, Seong Joon Oh, and Bernt Schiele. 2017. Generating descriptions with grounded and co-referenced people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dan Ruta, Andrew Gilbert, Pranav Aggarwal, Naveen Marri, Ajinkya Kale, Jo Briggs, Chris Speed, Hailin Jin, Baldo Faieta, Alex Filipkowski, Zhe Lin, and John Collomosse. 2022. Stylebabel: Artistic style tagging and captioning. In *Computer Vision – ECCV 2022*, pages 219–236, Cham. Springer Nature Switzerland.

Fawaz Sammani and Luke Melas-Kyriazi. 2020. Show, edit and tell: A framework for editing image captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gabriel Oliveira dos Santos, Esther Luna Colombini, and Sandra Avila. 2021. CIDEr-R: Robust consensus-based image description evaluation. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 351–360, Online. Association for Computational Linguistics.

Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2023. Positive-augmented contrastive learning for image and video captioning evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6914–6924.

Naeha Sharif, Lyndon White, Mohammed Bennamoun, and Syed Afaq Ali Shah. 2018. Nneval: Neural network based evaluation metric for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Himanshu Sharma. 2021. A survey on image captioning datasets and evaluation metrics. In *IOP Conference Series: Materials Science and Engineering*, volume 1116, page 012184. IOP Publishing.

Himanshu Sharma and Devanand Padha. 2023. A comprehensive survey on image captioning: from handcrafted to deep learning-based techniques, a taxonomy and open research issues. *Artificial Intelligence Review*, 56(11):13619–13661.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Tingke Shen, Amlan Kar, and Sanja Fidler. 2019. Learning to caption images through a lifetime by asking questions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. 2017. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Jiahe Shi, Yali Li, and Shengjin Wang. 2021a. Partial off-policy learning: Balance accuracy and diversity for human-oriented image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2187–2196.

Xiangxi Shi, Xu Yang, Jiuxiang Gu, Shafiq Joty, and Jianfei Cai. 2020a. Finding it at another side: A viewpoint-adapted matching encoder for change captioning. In *Computer Vision – ECCV 2020*, pages 574–590, Cham. Springer International Publishing.

Zhan Shi, Hui Liu, Martin Renqiang Min, Christopher Malon, Li Erran Li, and Xiaodan Zhu. 2021b. Retrieval, analogy, and composition: A framework for compositional generalization in image captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1990–2000, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhan Shi, Hui Liu, and Xiaodan Zhu. 2021c. Enhancing descriptive image captioning with natural language inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 269–277, Online. Association for Computational Linguistics.

Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. 2020b. Improving image captioning with better use of caption. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7454–7464, Online. Association for Computational Linguistics.

Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. 2019. Engaging image captioning via personality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: A dataset for image captioning with reading comprehension. In *Computer Vision – ECCV 2020*, pages 742–758, Cham. Springer International Publishing.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Lingyun Song, Jun Liu, Buyue Qian, and Yihe Chen. 2019. Connecting language to images: A progressive attention-guided network for simultaneous image captioning and language grounding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8885–8892.

Zeliang Song, Xiaofei Zhou, Zhendong Mao, and Jianlong Tan. 2021. Image captioning with context-aware auxiliary guidance. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(3):2584–2592.

Raimonda Staniūtė and Dmitrij Šešok. 2019. A systematic literature review on image captioning. *Applied Sciences*, 9(10):2024.

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2022. From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):539–559.

Qing Sun, Stefan Lee, and Dhruv Batra. 2017. Bidirectional beam search: Forward-backward inference in neural sequence models for fill-in-the-blank image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ece Takmaz, Sandro Pezzelle, Lisa Beinborn, and Raquel Fernández. 2020. Generating Image Descriptions via Sequential Cross-Modal Alignment Guided by Human Gaze. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4664–4677, Online. Association for Computational Linguistics.

Kohtaro Tanaka, Kohei Uehara, Lin Gu, Yusuke Mukuta, and Tatsuya Harada. 2024. Content-specific humorous image captioning using incongruity resolution chain-of-thought. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2348–2367, Mexico City, Mexico. Association for Computational Linguistics.

Hamed R. Tavakoli, Rakshith Shetty, Ali Borji, and Jorma Laaksonen. 2017. Paying attention to descriptions generated by image captioning models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Alasdair Tran, Alexander Mathews, and Lexing Xie. 2020. Transform and tell: Entity-aware news image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yunbin Tu, Liang Li, Li Su, Zheng-Jun Zha, Chenggang Yan, and Qingming Huang. 2023.

Self-supervised cross-view representation reconstruction for change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2805–2815.

Yunbin Tu, Liang Li, Chenggang Yan, Shengxiang Gao, and Zhengtao Yu. 2021a. R^3Net:relation-embedded representation reconstruction network for change captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9319–9329, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yunbin Tu, Tingting Yao, Liang Li, Jiedong Lou, Shengxiang Gao, Zhengtao Yu, and Chenggang Yan. 2021b. Semantic relation-aware difference representation learning for change captioning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 63–73, Online. Association for Computational Linguistics.

Yoshitaka Ushiku, Masataka Yamaguchi, Yusuke Mukuta, and Tatsuya Harada. 2015. Common subspace for model and similarity: Phrase learning for caption generation from images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2017. Captioning images with diverse objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gilad Vered, Gal Oren, Yuval Atzmon, and Gal Chechik. 2019. Joint optimization for cooperative image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Duc Minh Vo, Hong Chen, Akihiro Sugimoto, and Hideki Nakayama. 2022. Noc-rek: Novel object captioning with retrieved vocabulary from external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18000–18008.

Duc Minh Vo, Quoc-An Luong, Akihiro Sugimoto, and Hideki Nakayama. 2023. A-cap: Anticipation captioning with commonsense knowledge. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10824–10833.

Yuiga Wada, Kanta Kaneda, Daichi Saito, and Komei Sugiura. 2024. Polos: Multimodal metric learning from human feedback for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13559–13568.

Dalin Wang, Daniel Beck, and Trevor Cohn. 2019a. On the role of scene graphs in image captioning. In *Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, pages 29–34, Hong Kong, China. Association for Computational Linguistics.

Jing Wang, Jinhui Tang, Mingkun Yang, Xiang Bai, and Jiebo Luo. 2021a. Improving ocr-based image captioning by incorporating geometrical relationship. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1306–1315.

Jiuniu Wang, Wenjia Xu, Qingzhong Wang, and Antoni B Chan. 2020a. Compare and reweight: Distinctive image captioning using similar images sets. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 370–386. Springer.

Jiuniu Wang, Wenjia Xu, Qingzhong Wang, and Antoni B. Chan. 2020b. Compare and reweight: Distinctive image captioning using similar images sets. In *Computer Vision – ECCV 2020*, pages 370–386, Cham. Springer International Publishing.

Josiah Wang, Pranava Swaroop Madhyastha, and Lucia Specia. 2018. Object counts! bringing explicit detections back into image captioning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2180–2193, New Orleans, Louisiana. Association for Computational Linguistics.

Li Wang, Zechen Bai, Yonghua Zhang, and Hongtao Lu. 2020c. Show, recall, and tell: Image captioning with recall mechanism. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12176–12183.

Liwei Wang, Alexander Schwing, and Svetlana Lazebnik. 2017a. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ning Wang, Jiajun Deng, and Mingbo Jia. 2024. Cycle-consistency learning for captioning and grounding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):5535–5543.

Ning Wang, Jiahao Xie, Jihao Wu, Mingbo Jia, and Linlin Li. 2023a. Controllable image captioning via prompting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):2617–2625.

Ning Wang, Jiangrong Xie, Hang Luo, Qinglin Cheng, Jihao Wu, Mingbo Jia, and Linlin Li. 2023b. Efficient image captioning for edge devices. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):2608–2616.

Qingzhong Wang and Antoni B. Chan. 2019. Describing like humans: On diversity in image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sijin Wang, Ziwei Yao, Ruiping Wang, Zhongqin Wu, and Xilin Chen. 2021b. Faier: Fidelity and adequacy ensured image caption evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14050–14059.

Ting Wang, Weidong Chen, Yuanhe Tian, Yan Song, and Zhendong Mao. 2023c. Improving image captioning via predicting structured concepts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 360–370, Singapore. Association for Computational Linguistics.

Weixuan Wang, Zhihong Chen, and Haifeng Hu. 2019b. Hierarchical attention network for image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8957–8964.

Yiyu Wang, Jungang Xu, and Yingfei Sun. 2022a. End-to-end transformer based model for image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):2585–2594.

Yufei Wang, Zhe Lin, Xiaohui Shen, Scott Cohen, and Garrison W. Cottrell. 2017b. Skeleton key: Image captioning by skeleton-attribute decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yufei Wang, Ian Wood, Stephen Wan, and Mark Johnson. 2021c. ECOL-R: Encouraging copying in novel object captioning with reinforcement learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1222–1234, Online. Association for Computational Linguistics.

Yuxuan Wang, Difei Gao, Licheng Yu, Weixian Lei, Matt Feiszli, and Mike Zheng Shou. 2022b. Geb+: A benchmark for generic event boundary captioning, grounding and retrieval. In *Computer Vision – ECCV 2022*, pages 709–725, Cham. Springer Nature Switzerland.

Zeyu Wang, Berthy Feng, Karthik Narasimhan, and Olga Russakovsky. 2020d. Towards unique and informative captioning of images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020,* *Proceedings, Part VII 16*, pages 629–644. Springer.

Zeyu Wang, Berthy Feng, Karthik Narasimhan, and Olga Russakovsky. 2020e. Towards unique and informative captioning of images. In *Computer Vision – ECCV 2020*, pages 629–644, Cham. Springer International Publishing.

Zhen Wang, Long Chen, Wenbo Ma, Guangxing Han, Yulei Niu, Jian Shao, and Jun Xiao. 2022c. Explicit image caption editing. In *Computer Vision – ECCV 2022*, pages 113–129, Cham. Springer Nature Switzerland.

Jianzong Wu, Xiangtai Li, Henghui Ding, Xia Li, Guangliang Cheng, Yunhai Tong, and Chen Change Loy. 2023a. Betrayed by captions: Joint caption grounding and generation for open vocabulary instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21938–21948.

Mingrui Wu, Xuying Zhang, Xiaoshuai Sun, Yiyi Zhou, Chao Chen, Jiaxin Gu, Xing Sun, and Rongrong Ji. 2022. Difnet: Boosting visual information flow for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18020–18029.

Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. 2023b. Cross2StrA: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2593–2608, Toronto, Canada. Association for Computational Linguistics.

Yujia Xie, Luowei Zhou, Xiyang Dai, Lu Yuan, Nguyen Bach, Ce Liu, and Michael Zeng. 2022. Visual clues: Bridging vision and language foundations for image paragraph captioning. In *Advances in Neural Information Processing Systems*, volume 35, pages 17287–17300. Curran Associates, Inc.

Guanghui Xu, Shuaicheng Niu, Mingkui Tan, Yucheng Luo, Qing Du, and Qi Wu. 2021. Towards accurate text-based image captioning with content diversity exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12637–12646.

Jitao Xu, François Buet, Josep Crego, Elise Bertin-Lemée, and François Yvon. 2022. Joint generation of captions and subtitles with dual decoding. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 74–82, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France. PMLR.

Liming Xu, Quan Tang, Jiancheng Lv, Bochuan Zheng, Xianhua Zeng, and Weisheng Li. 2023. Deep image captioning: A review of methods, trends and future challenges. *Neurocomputing*, 546:126287.

Semih Yagcioglu, Erkut Erdem, Aykut Erdem, and Ruket Cakici. 2015. A distributed representation based query expansion approach for image captioning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 106–111, Beijing, China. Association for Computational Linguistics.

Kun Yan, Lei Ji, Huaishao Luo, Ming Zhou, Nan Duan, and Shuai Ma. 2021. Control image captioning spatially and temporally. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2014–2025, Online. Association for Computational Linguistics.

Bang Yang, Fenglin Liu, Xian Wu, Yaowei Wang, Xu Sun, and Yuexian Zou. 2023a. MultiCapCLIP: Auto-encoding prompts for zero-shot multilingual visual captioning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11908–11922, Toronto, Canada. Association for Computational Linguistics.

Cheng-Fu Yang, Yao-Hung Hubert Tsai, Wan-Cyuan Fan, Russ R Salakhutdinov, Louis-Philippe Morency, and Frank Wang. 2022. Paraphrasing is all you need for novel object captioning. In *Advances in Neural Information Processing Systems*, volume 35, pages 6492–6504. Curran Associates, Inc.

Fan Yang, Shalini Ghosh, Emre Barut, Kechen Qin, Prashan Wanigasekara, Chengwei Su, Weitong Ruan, and Rahul Gupta. 2024. Masking latent gender knowledge for debiasing image captioning. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 227–238, Mexico City, Mexico. Association for Computational Linguistics.

Li-Chuan Yang, Chih-Yuan Yang, and Jane Yung-jen Hsu. 2021a. Object relation attention for image paragraph captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3136–3144.

Linjie Yang, Kevin Tang, Jianchao Yang, and Li-Jia Li. 2017. Dense captioning with joint inference and visual context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xu Yang, Chongyang Gao, Hanwang Zhang, and Jianfei Cai. 2021b. Auto-parsing network for image captioning and visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2197–2207.

Xu Yang, Jiawei Peng, Zihua Wang, Haiyang Xu, Qinghao Ye, Chenliang Li, Songfang Huang, Fei Huang, Zhangzikang Li, and Yu Zhang. 2023b. Transforming visual scene graphs to image captions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12427–12440, Toronto, Canada. Association for Computational Linguistics.

Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019a. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Xin Geng. 2023c. Exploring diverse in-context configurations for image captioning. In *Advances in Neural Information Processing Systems*, volume 36, pages 40924–40943. Curran Associates, Inc.

Xu Yang, Hanwang Zhang, and Jianfei Cai. 2019b. Learning to collocate neural modules for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Xuewen Yang, Svebor Karaman, Joel Tetreault, and Alejandro Jaimes. 2021c. Journalistic guidelines aware news image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5162–5175, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xuewen Yang, Heming Zhang, Di Jin, Yingru Liu, Chi-Hao Wu, Jianchao Tan, Dongliang Xie, Jue Wang, and Xin Wang. 2020. Fashion captioning: Towards generating accurate descriptions with semantic rewards. In *Computer Vision – ECCV 2020*, pages 1–17, Cham. Springer International Publishing.

Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei Florencio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. 2021d. Tap: Text-aware pre-training for text-vqa and text-caption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8751–8761.

Zhilin Yang, Ye Yuan, Yuexin Wu, William W Cohen, and Russ R Salakhutdinov. 2016. Review networks for caption generation. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Zhuolin Yang, Wei Ping, Zihan Liu, Vijay Korthikanti, Weili Nie, De-An Huang, Linxi Fan, Zhiding Yu, Shiyi Lan, Bo Li, Mohammad Shoeybi, Ming-Yu Liu, Yuke Zhu, Bryan Catanzaro, Chaowei Xiao, and Anima Anandkumar. 2023d. Re-ViLM: Retrieval-augmented visual language model for zero and few-shot image captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11844–11857, Singapore. Association for Computational Linguistics.

Li Yao, Nicolas Ballas, Kyunghyun Cho, John R Smith, and Yoshua Bengio. 2016. Empirical performance upper bounds for image and video captioning.

Linli Yao, Weiying Wang, and Qin Jin. 2022. Image difference captioning with pre-training and contrastive learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):3108–3116.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2017a. Incorporating copying mechanism in image captioning for learning novel objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2019. Hierarchy parsing for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017b. Boosting image captioning with attributes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Mark Yatskar, Michel Galley, Lucy Vanderwende, and Luke Zettlemoyer. 2014. See no evil, say no evil: Description generation from densely labeled images. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 110–120, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Yanzhi Yi, Hangyu Deng, and Jinglu Hu. 2020. Improving image captioning evaluation by considering inter references variance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 985–994, Online. Association for Computational Linguistics.

Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, and Jing Shao. 2019. Context and attribute grounded dense captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Licheng Yu, Eunbyung Park, Alexander C. Berg, and Tamara L. Berg. 2015. Visual madlibs: Fill in the blank description generation and question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Zhihao Yuan, Xu Yan, Yinghong Liao, Yao Guo, Guanbin Li, Shuguang Cui, and Zhen Li. 2022. X-trans2cap: Cross-modal knowledge transfer using transformer for 3d dense captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8563–8573.

Zihao Yue, Anwen Hu, Liang Zhang, and Qin Jin. 2023. Learning descriptive image captioning via semipermeable maximum likelihood estimation. In *Advances in Neural Information Processing Systems*, volume 36, pages 79124–79141. Curran Associates, Inc.

Zequn Zeng, Yan Xie, Hao Zhang, Chiyu Chen, Bo Chen, and Zhengjue Wang. 2024a. Meacap: Memory-augmented zero-shot image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14100–14110.

Zequn Zeng, Yan Xie, Hao Zhang, Chiyu Chen, Bo Chen, and Zhengjue Wang. 2024b. Meacap: Memory-augmented zero-shot image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14100–14110.

Zequn Zeng, Hao Zhang, Ruiying Lu, Dongsheng Wang, Bo Chen, and Zhengjue Wang. 2023. Conzic: Controllable zero-shot image captioning by sampling-based polishing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23465–23476.

Hongkuan Zhang, Saku Sugawara, Akiko Aizawa, Lei Zhou, Ryohei Sasano, and Koichi Takeda. 2022a. Cross-modal similarity-based curriculum learning for image captioning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7599–7606, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Junzhe Zhang and Xiaojun Wan. 2023. Exploring the impact of vision features in news image captioning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12923–12936, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wei Zhang, Yue Ying, Pan Lu, and Hongyuan Zha. 2020. Learning long- and short-term user literal-preference with multimodal hierarchical transformer network for personalized image caption. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9571–9578.

Wenqiao Zhang, Haochen Shi, Jiannan Guo, Shengyu Zhang, Qingpeng Cai, Juncheng Li, Sihui Luo, and Yueting Zhuang. 2022b. Magic: Multimodal relational graph adversarial inference for diverse and unpaired text-based image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):3335–3343.

Wenqiao Zhang, Haochen Shi, Siliang Tang, Jun Xiao, Qiang Yu, and Yueting Zhuang. 2021a. Consensus graph representation learning for better grounded image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3394–3402.

Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. 2021b. Rstnet: Captioning with adaptive attention on visual and non-visual words. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15465–15474.

Dora Zhao, Angelina Wang, and Olga Russakovsky. 2021. Understanding and evaluating racial biases in image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14830–14840.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Sanqiang Zhao, Piyush Sharma, Tomer Levinboim, and Radu Soricut. 2019. Informative image captioning with external sources of information. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6485–6494, Florence, Italy. Association for Computational Linguistics.

Wentian Zhao, Xinxiao Wu, and Xiaoxun Zhang. 2020. Memcap: Memorizing style knowledge for image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12984–12992.

Yu Zhao, Hao Fei, Wei Ji, Jianguo Wei, Meishan Zhang, Min Zhang, and Tat-Seng Chua. 2023. Generating visual spatial description via holistic 3D scene understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7960–7977, Toronto, Canada. Association for Computational Linguistics.

Yu Zhao, Jianguo Wei, ZhiChao Lin, Yueheng Sun, Meishan Zhang, and Min Zhang. 2022. Visual spatial description: Controlled spatial-oriented image-to-text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1437–1449, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ervine Zheng and Qi Yu. 2023. Evidential interactive learning for medical image captioning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 42478–42491. PMLR.

Yue Zheng, Yali Li, and Shengjin Wang. 2019. Intention oriented image captions with guiding objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wenjie Zhong and Yusuke Miyao. 2021. Leveraging partial dependency trees to control image captions. In *Proceedings of the Second Workshop on Advances in Language and Vision Research*, pages 16–21, Online. Association for Computational Linguistics.

Yiwu Zhong, Liwei Wang, Jianshu Chen, Dong Yu, and Yin Li. 2020. Comprehensive image captioning via scene graph decomposition. In *Computer Vision – ECCV 2020*, pages 211–229, Cham. Springer International Publishing.

Zhipeng Zhong, Fei Zhou, and Guoping Qiu. 2023. Aesthetically relevant image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(3):3733–3741.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020a. Unified vision-language pre-training for image captioning and vqa. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):13041–13049.

Mingyang Zhou, Grace Luo, Anna Rohrbach, and Zhou Yu. 2022. Focus! relevant and sufficient context selection for news image captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6078–6088, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuanen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang. 2020b. More grounded image captioning by distilling image-text matching model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yucheng Zhou and Guodong Long. 2023. Style-aware contrastive learning for multi-style image captioning. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2257–2267, Dubrovnik, Croatia. Association for Computational Linguistics.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 1097–1100, New York, NY, USA. Association for Computing Machinery.

A Experiments: Additional Information

A.1 Datasets

Flickr8k. We use the official version of the Flickr8k dataset.[9] We follow previous work (Anderson et al., 2016; Hessel et al., 2021) when handling references that are also included in the candidate set: in Flickr8k-Expert we remove these sentences from the candidate set (158 examples), and in Flickr8k-CF we remove these sentences from the reference set.

Composite. We use the official version of the Composite dataset.[10] While previous work mentioned that there are three candidates per image and a single score per candidate, we find that the dataset contains four candidates per image for MSCOCO and Flickr30k, and two scores per candidate (correctness and throughness). We find that results from prior work was most closely replicated when we discarded the fourth candidate when one exists and used the correctness score as the total score. Also, following Hessel et al. (2021) we remove sentences that appear both as references and as candidates from the reference set.

THumB. We use the official version of the THumB dataset.[11] Following the practice in other datasets, we remove sentences that appear both as references and as candidates from the reference set.

Polaris. We use the HuggingFace version of the Polaris dataset.[12]

Pascal50S. The link to download the official version of the dataset in the CIDEr blog post[13] seems to be broken; we thank (REMOVED FOR ANONIMYZATION) for providing us with the original files.

A.2 Metrics implementations.

The five dominant metrics. For the five most dominant metrics we use the pycocoevalcap implementation.[14]

Exact/fuzzy noun/verb overlap. We use an implementation[15] provided to us by the author.

CLIPImageScore. Since no implementation is provided by the authors, we implement the metric ourselves. We use the SD-XL 1.0-base model[16] to generate the image based on the candidate.

(Ref) CLIP and PAC Score, Polos. We use the official implementations.[17] [18] [19]

BLIP2Score. Since no implementation is provided by the authors, we implement the metric ourselves. We use the BLIP2 image-text-matching model from the LAVIS package (Li et al., 2023a).

MPNetScore. Since no implementation is provided by the authors, we implement the metric ourselves. We use the all-mpnet-base-v2 model.[20]

A.3 Feature Selection and Linear Regression

We use the scikit-learn[21] implementations of feature selection and linear regression. We use forward-selection in feature selection.

[9] www.kaggle.com/datasets/dibyansudiptiman/flickr-8k
[10] imagesdg.wordpress.com/image-to-scene-description-graph
[11] github.com/jungokasai/THumB
[12] huggingface.co/datasets/yuwd/Polaris
[13] vrama91.github.io/cider/
[14] pypi.org/project/pycocoevalcap
[15] github.com/DavidMChan/caption-by-committee/blob/main/cbc/metrics/content_score.py
[16] huggingface.co/stabilityai/stable-diffusion-xl-base-1.0
[17] github.com/jmhessel/clipscore
[18] github.com/aimagelab/pacscore
[19] github.com/keio-smilab24/Polos
[20] huggingface.co/sentence-transformers/all-mpnet-base-v2
[21] scikit-learn.org/stable

A.4 Reproducibility Issues

In our experiments in Section 5.3 we tried our best to replicate the results reported by previous studies. We now describe the cases where we were enable to. We compare our results to Hessel et al. (2021), Kasai et al. (2022), Sarto et al. (2023), Hu et al. (2023a) and Wada et al. (2024).

The five dominant metrics. Across some of the datasets, we found small differences between our values and previous reported values, no larger than 0.2. One exception is the THumB dataset, where prior work (Kasai et al., 2022; Hu et al., 2023a) report no digits after the decimal point, so the differences are no larger than 1.0, except for BLEU4, where the difference was larger because the authors used the sacrebleu implemenation while we used the pycocoevalcap implementation, consistent with the common practice in image captioning.

(Ref) CLIP and PAC Score. Across all datasets our results are identical to those previous reported. One exception is the Polaris dataset, where our results for these metrics differed from those of Wada et al. (2024). Since we used the HuggingFace version of the dataset and the official implementations for these metrics, we assume this discrepancy results from the authors using a more preliminary version of the dataset.

B Metric Usage Records

We now list all the papers we examined and the metric categories they used. Abbreviations: LexSim: Lexical similarity, RefPhSim: Candidate-reference phrase semantic similarity, RefSenSim: Candidate-reference sentence-level similarity, InfoSim: Extracted information similarity, ImPhSim: Candidate-image phrase-level semantic similarity, SenImSim: Sentence-image similarity, Ret: Retrieval-based methods, MulSim: Candidate-multiple source similarity, Flu: Fluency, Div: Diversity, Yngve: Yngve score, Bias: Bias.

2024. *NLP* Yang et al. (2024): LexSim, RefPhSim, Bias, Ramos et al. (2024): LexSim, Tanaka et al. (2024): Div, Qu et al. (2024): LexSim, Li et al. (2024b): LexSim, SenImSim

Vision Ge et al. (2024b): RefSenSim, SenImSim, Huang et al. (2024): LexSim, RefPhSim, Zeng et al. (2024b): LexSim, RefPhSim, SenImSim, Li et al. (2024a): LexSim, RefPhSim

Machine Learning Ma et al. (2024c): LexSim, RefPhSim, Black et al. (2024b): LexSim, RefSenSim, Liu et al. (2024): LexSim, RefPhSim, Ma et al. (2024a): LexSim, Qi et al. (2024): LexSim, Qiu et al. (2024): LexSim, RefPhSim, Wang et al. (2024): LexSim, RefPhSim, Fu et al. (2024): LexSim, RefPhSim

2023. *NLP* Wang et al. (2023c): LexSim, RefPhSim, Qiu et al. (2023): LexSim, RefPhSim, SenImSim, Hwang and Shwartz (2023): LexSim, RefPhSim, Chan et al. (2023a): LexSim, RefPhSim, Ret, Maeda et al. (2023): LexSim, RefPhSim, SenImSim, MulSim, Yang et al. (2023d): LexSim, RefPhSim, Mohamed et al. (2023): LexSim, Wu et al. (2023b): LexSim, Yang et al. (2023a): LexSim, RefPhSim, Yang et al. (2023b): LexSim, RefPhSim, Rajakumar Kalarani et al. (2023): LexSim, Ramos et al. (2023b): LexSim, Ou et al. (2023): Flu, Zhang and Wan (2023): LexSim, Bielawski and VanRullen (2023): LexSim, Anagnostopoulou et al. (2023): LexSim, Zhao et al. (2023): LexSim, RefPhSim, Ramos et al. (2023a): LexSim, RefPhSim, Zhou and Long (2023): LexSim, RefPhSim, Flu

Vision Wu et al. (2023a): LexSim, Li et al. (2023b): Flu, Div, Kornblith et al. (2023): LexSim, SenImSim, Ret, MulSim, Hu et al. (2023b): LexSim, RefPhSim, Tu et al. (2023): LexSim, RefPhSim, Fei et al. (2023a): LexSim, RefPhSim, Fan et al. (2023): LexSim, RefPhSim, Kang et al. (2023): LexSim, RefPhSim, SenImSim, Ret, Barraco et al. (2023): LexSim, RefPhSim, SenImSim, MulSim, Hu et al. (2023b): LexSim, RefPhSim, Chen et al. (2023b): LexSim, Dessì et al. (2023): LexSim, RefPhSim, Vo et al. (2023): LexSim, RefPhSim, SenImSim, Ret, MulSim, Luo et al. (2023a): LexSim, RefPhSim, Zeng et al. (2023): LexSim, RefPhSim, SenImSim, MulSim, Div, Kuo and Kira (2023): LexSim, RefPhSim, Chen et al. (2023a): LexSim, Ramos et al. (2023c): LexSim, RefPhSim, Hirota et al. (2023): LexSim, RefPhSim, SenImSim, Bias, Ren et al. (2023): LexSim, RefPhSim

Machine Learning Yang et al. (2023c): LexSim, Yue et al. (2023): SenImSim, Ret, Flu, Div, Luo et al. (2023b): SenImSim, Ret, Zheng and Yu (2023): LexSim, Li et al. (2023c): LexSim, RefPhSim, Fei et al. (2023b): LexSim, RefPhSim, Zhong et al. (2023): LexSim, RefPhSim, Nguyen et al. (2023): LexSim, RefPhSim, Wang et al. (2023b): LexSim, RefPhSim, Wang et al. (2023a): LexSim, RefPhSim

2022. *NLP* Gao et al. (2022): LexSim, Div, Zhang et al. (2022a): LexSim, RefPhSim, Mirchandani et al. (2022): LexSim, Nukrai et al. (2022): LexSim, Zhou et al. (2022): LexSim, Zhao et al. (2022): LexSim, RefPhSim, Nishino et al. (2022): LexSim, RefPhSim, InfoSim, Cafagna et al. (2022): LexSim, RefPhSim, Cho et al. (2022): LexSim, RefPhSim, SenImSim, Ret, MulSim, Pantazopoulos et al. (2022): LexSim, RefPhSim, Xu et al. (2022): LexSim, Guo et al. (2022): LexSim

Vision Jiao et al. (2022): LexSim, Wang et al. (2022b): LexSim, RefPhSim, Ruta et al. (2022): LexSim, Wang et al. (2022c): LexSim, RefPhSim, Nguyen et al. (2022): LexSim, RefPhSim, Meng et al. (2022): LexSim, RefPhSim, Chen et al. (2022): LexSim, Hirota et al. (2022): LexSim, Bias, Cai et al. (2022): LexSim, Wu et al. (2022): LexSim, RefPhSim, Chen et al. (2022): LexSim, Yuan et al. (2022): LexSim, Li et al. (2022): LexSim, RefPhSim, Fei et al. (2022): LexSim, RefPhSim, Liu et al. (2022): LexSim, RefPhSim, Vo et al. (2022): LexSim, RefPhSim, Mohamed et al. (2022): LexSim, Fang et al. (2022): LexSim, RefPhSim, Hu et al. (2022): LexSim, RefPhSim, Kuo and Kira (2022): LexSim, RefPhSim, Mavroudi and Vidal (2022): LexSim, RefPhSim

Machine Learning Yang et al. (2022): LexSim, RefPhSim, Yang et al. (2022): LexSim, RefPhSim, SenImSim, Div, Xie et al. (2022): ImPhSim, Fei (2022): LexSim, RefPhSim, Zhang et al. (2022b): LexSim, RefPhSim, Div, Yao et al. (2022): LexSim, Wang et al. (2022a): LexSim, RefPhSim, Feng et al. (2022): LexSim, RefPhSim, Flu

2021. *NLP* Yang et al. (2021c): LexSim, Liu et al. (2021): LexSim, Tu et al. (2021a): LexSim, RefPhSim, Chen et al. (2021a): LexSim, Shi et al. (2021b): LexSim, RefPhSim, Div, Ng et al. (2021): LexSim, RefPhSim, Div, Yan et al. (2021): LexSim, Shi et al. (2021c): LexSim, RefPhSim, Ret, Tu et al. (2021b): LexSim, RefPhSim, Bugliarello and Elliott (2021): LexSim, RefPhSim, Wang et al. (2021c): LexSim, RefPhSim, Honda et al. (2021): LexSim, RefPhSim, Nag Chowdhury et al. (2021): LexSim, RefPhSim, Ahsan et al. (2021): LexSim, RefPhSim, Zhong and Miyao (2021): LexSim, RefPhSim, Nagasawa et al. (2021): LexSim, RefPhSim

Vision Shi et al. (2021a): LexSim, Div, Kim et al. (2021a): LexSim, RefPhSim, Zhao et al. (2021): LexSim, RefPhSim, Yang et al. (2021b): LexSim, RefPhSim, Nguyen et al. (2021): LexSim, RefPhSim, Yang et al. (2021d): LexSim, Chen et al. (2021b): LexSim, RefPhSim, Div, Hosseinzadeh and Wang (2021): LexSim, RefPhSim, Xu et al. (2021): LexSim, RefPhSim, Div, Wang et al. (2021a): LexSim, RefPhSim, Chen et al. (2021c): LexSim, Zhang et al. (2021b): LexSim, RefPhSim

Machine Learning Hu et al. (2021): LexSim, RefPhSim, Ji et al. (2021): LexSim, RefPhSim, Fei (2021b): LexSim, RefPhSim, Fei (2021a): LexSim, RefPhSim, Song et al. (2021): LexSim, RefPhSim, Luo et al. (2021): LexSim, RefPhSim, Zhang et al. (2021a): LexSim, RefPhSim, Yang et al. (2021a): LexSim, Kim et al. (2021b): LexSim

2020. *NLP* Milewski et al. (2020): LexSim, RefPhSim, Alikhani et al. (2020): LexSim, Shi et al. (2020b): LexSim, RefPhSim, Park et al. (2020): LexSim, Nie et al. (2020): LexSim, Nguyen et al. (2020): LexSim, RefPhSim, Takmaz et al. (2020): LexSim, RefPhSim

Vision Wang et al. (2020b): LexSim, RefPhSim, Ret, Div, Sidorov et al. (2020): LexSim, RefPhSim, Wang et al. (2020e): LexSim, RefPhSim, Yang et al. (2020): LexSim, RefPhSim, Deng et al. (2020): LexSim, RefPhSim, Zhong et al. (2020): LexSim, RefPhSim, Div, Shi et al. (2020a): LexSim, RefPhSim, Gurari et al. (2020): LexSim, RefPhSim, Ma et al. (2020): LexSim, RefPhSim, Chen et al. (2020): LexSim, RefPhSim, Sammani and Melas-Kyriazi (2020): LexSim, RefPhSim, Guo et al. (2020): LexSim, RefPhSim, Chen and Jin (2020): LexSim, RefPhSim, Div, Cornia et al. (2020): LexSim, RefPhSim, Pan et al. (2020): LexSim, RefPhSim, Zhou et al. (2020b): LexSim, RefPhSim, Tran et al. (2020): LexSim

Machine Learning Del Chiaro et al. (2020): LexSim, Mahajan and Roth (2020): LexSim, RefPhSim, Div, Zhang et al. (2020): LexSim, RefPhSim, Zhou et al. (2020a): LexSim, RefPhSim, Zhao et al. (2020): LexSim, Flu, Wang et al. (2020c): LexSim, RefPhSim, Liu et al. (2020): LexSim, Hou et al. (2020): LexSim, RefPhSim, Cao et al. (2020): LexSim, RefPhSim

2019. *NLP* Zhao et al. (2019): LexSim, Fan et al. (2019): LexSim, Nikolaus et al. (2019): LexSim, RefPhSim, Changpinyo et al. (2019): LexSim, RefPhSim, Kim et al. (2019b): LexSim, RefPhSim, Wang et al. (2019a): LexSim, RefPhSim

Vision Shen et al. (2019): LexSim, Div, Liu et al. (2019a): LexSim, RefPhSim, Ret, Div, Yngve, Ge et al. (2019): LexSim, Yao et al. (2019): LexSim, RefPhSim, Yang et al. (2019b): LexSim, RefPhSim, Aneja et al. (2019): LexSim, RefPhSim, Div, Park et al. (2019): LexSim, RefPhSim, Huang et al. (2019a): LexSim, RefPhSim, Laina et al. (2019): LexSim, RefPhSim, Ke et al. (2019): LexSim, RefPhSim, He et al. (2019): LexSim, Vered et al. (2019): LexSim, RefPhSim, Ret, Li et al. (2019a): LexSim, RefPhSim, Agrawal et al. (2019): LexSim, RefPhSim, Gu et al. (2019): LexSim, RefPhSim, Cornia et al. (2019b): LexSim, RefPhSim, Zheng et al. (2019): LexSim, RefPhSim, Dognin et al. (2019): LexSim, SenImSim, Feng et al. (2019): LexSim, RefPhSim, Wang and Chan (2019): LexSim, RefPhSim, Div, Guo et al. (2019): LexSim, Flu, Yin et al. (2019): LexSim, Kim et al. (2019a): LexSim, Ret, Gao et al. (2019a): LexSim, RefPhSim, Qin et al. (2019): LexSim, RefPhSim, Yang et al. (2019a): LexSim, RefPhSim, Deshpande et al. (2019): LexSim, RefPhSim, Div, Biten et al. (2019): LexSim, RefPhSim, Li et al. (2019d): LexSim, RefPhSim, Shuster et al. (2019): LexSim, RefPhSim

Machine Learning Herdade et al. (2019): LexSim, RefPhSim, Chen et al. (2019c): LexSim, RefPhSim, Div, Huang et al. (2019b): LexSim, RefPhSim, Wang et al. (2019b): LexSim, RefPhSim, Song et al. (2019): LexSim, Li et al. (2019b): LexSim, RefPhSim, Ret, Li et al. (2019c): LexSim, Gao et al. (2019b): LexSim, RefPhSim, Chen et al. (2019a): LexSim, RefPhSim, Chen et al. (2019b): LexSim, RefPhSim

2018. *NLP* Sharma et al. (2018): LexSim, RefPhSim, Chen et al. (2018b): LexSim, Madhyastha et al. (2018): LexSim, RefPhSim, Liu et al. (2018a): LexSim, RefPhSim, Guo et al. (2018): LexSim, Melas-Kyriazi et al. (2018): LexSim, Lu et al. (2018): LexSim, Div, Wang et al. (2018): LexSim, RefPhSim, Cohn-Gordon et al. (2018): Ret, Chandrasekaran et al. (2018): NONE

Vision Chen and Zhao (2018): LexSim, Liu et al. (2018b): LexSim, RefPhSim, Ret, Div, Gu et al. (2018b): LexSim, Div, Dai et al. (2018b): LexSim, RefPhSim, Hendricks et al. (2018): LexSim, Bias, Jiang et al. (2018b): LexSim, RefPhSim, Chen et al. (2018d): LexSim, Yao et al. (2018): LexSim, RefPhSim, Chen et al. (2018e): LexSim, RefPhSim, Luo et al. (2018): LexSim, RefPhSim, Ret, Mathews et al. (2018): LexSim, RefPhSim, Flu, Chen et al. (2018a): LexSim, Aneja et al. (2018): LexSim, RefPhSim, Anderson et al. (2018b): LexSim, RefPhSim

Machine Learning Dai et al. (2018a): LexSim, RefPhSim, Div, Huang et al. (2018): LexSim, RefPhSim, Anderson et al. (2018a): LexSim, RefPhSim, Jiang et al. (2018a): LexSim, Div, Chen et al. (2018c): LexSim, Gu et al. (2018a): LexSim, RefPhSim

2017. *NLP* Anderson et al. (2017): LexSim, RefPhSim

Vision Shetty et al. (2017): LexSim, RefPhSim, Div, Liu et al. (2017b): LexSim, Gu et al. (2017): LexSim, RefPhSim, Pedersoli et al. (2017): LexSim, Tavakoli et al. (2017): LexSim, Yao et al. (2017b): LexSim, RefPhSim, Dai et al. (2017): LexSim, RefPhSim, Vedantam et al. (2017): LexSim, Gan et al. (2017a): LexSim, Ren et al. (2017): LexSim, Lu et al. (2017): LexSim, RefPhSim, Chunseong Park et al. (2017): LexSim, Yang et al. (2017): NONE, Gan et al. (2017b): LexSim, Chen et al. (2017a): LexSim, Venugopalan et al. (2017): LexSim, Yao et al. (2017a): LexSim, Sun et al. (2017): LexSim, Rennie et al. (2017): LexSim, Wang et al. (2017b): LexSim, RefPhSim, Div, Rohrbach et al. (2017): LexSim

Machine Learning Dai and Lin (2017): LexSim, Ret, Wang et al. (2017a): LexSim, RefPhSim, Div, Liu et al. (2017a): LexSim, Chen et al. (2017b): LexSim, Li et al. (2017): LexSim, Mun et al. (2017): LexSim

2016. *NLP*

Vision Hendricks et al. (2016): LexSim, Johnson et al. (2016): LexSim, Ret, You et al. (2016): LexSim

Machine Learning Yang et al. (2016): LexSim, Pu et al. (2016): LexSim, Flu, Yao et al. (2016): LexSim, Mathews et al. (2016): LexSim

2015. *NLP* Devlin et al. (2015): LexSim, Flu, Yagcioglu et al. (2015): LexSim, Chen et al. (2015): LexSim

Vision Jia et al. (2015): LexSim, Ushiku et al. (2015): LexSim, Yu et al. (2015): LexSim, Chen and Lawrence Zitnick (2015): LexSim, Ret, Flu, Vinyals et al. (2015): LexSim, Ret, Fang et al. (2015): LexSim, Flu, Donahue et al. (2015): LexSim, Ret, Karpathy and Fei-Fei (2015): LexSim, Ret

Machine Learning Xu et al. (2015): LexSim, Lebret et al. (2015): LexSim, Mao et al. (2014): LexSim, Ret, Flu

2014. *NLP* Mason and Charniak (2014b): LexSim, Mason and Charniak (2014a): LexSim, Yatskar et al. (2014): LexSim, Kuznetsova et al. (2014): LexSim

2013. *NLP* Kuznetsova et al. (2013): LexSim, Elliott and Keller (2013): LexSim

2012. *NLP* Kuznetsova et al. (2012): LexSim, Mitchell et al. (2012): NONE

2011. *NLP* Li et al. (2011): LexSim

Machine Learning Ordonez et al. (2011): LexSim

2010. *NLP* Feng and Lapata (2010): LexSim, Aker and Gaizauskas (2010): LexSim