A Novel Evaluation Framework for Image2Text Generation

Jia-Hong Huang[1,†], Hongyi Zhu[1,†], Yixian Shen[1], Stevan Rudinac[1], Alessio M. Pacces[1] and Evangelos Kanoulas[1,*]

[1] *University of Amsterdam, Netherlands*

Abstract

Evaluating the quality of automatically generated image descriptions is challenging, requiring metrics that capture various aspects such as grammaticality, coverage, correctness, and truthfulness. While human evaluation offers valuable insights, its cost and time-consuming nature pose limitations. Existing automated metrics like BLEU, ROUGE, METEOR, and CIDEr aim to bridge this gap but often show weak correlations with human judgment. We address this challenge by introducing a novel evaluation framework rooted in a modern large language model (LLM), such as GPT-4 or Gemini, capable of image generation. In our proposed framework, we begin by feeding an input image into a designated image captioning model, chosen for evaluation, to generate a textual description. Using this description, an LLM then creates a new image. By extracting features from both the original and LLM-created images, we measure their similarity using a designated similarity metric. A high similarity score suggests that the image captioning model has accurately generated textual descriptions, while a low similarity score indicates discrepancies, revealing potential shortcomings in the model's performance. Human-annotated reference captions are not required in our proposed evaluation framework, which serves as a valuable tool for evaluating the effectiveness of image captioning models. Its efficacy is confirmed through human evaluation.

Keywords

Image Captioning, Metrics for Automated Evaluation, Large Language Models,

1. Introduction

The evaluation of sentences generated through automated methods remains a formidable challenge in the realm of image captioning. Current metrics for evaluating image descriptions aim to gauge multiple desirable attributes, such as grammaticality, covering crucial aspects, correctness, truthfulness, and more. Human evaluation plays a pivotal role in quantifying these properties, utilizing separate Likert scales or pairwise scales [1, 2, 3, 4, 5, 6]. However, due to the expensive, challenging-to-reproduce, and time-consuming nature of human studies, there is a growing need for automated evaluation measures. For practical utility, these automated metrics should align closely with human judgment. Therefore, the challenge in designing such an automatic metric lies in integrating the aforementioned diverse evaluations attributes into a unified measure of sentence quality.

Several automated metrics, including BLEU [7], ROUGE [8], METEOR [4], CIDEr [9], and more, have been introduced to assess image descriptions generated by automated approaches. BLEU, initially designed for machine translation, relies on precision, while ROUGE, originating from the summarization community, is a recall-based metric. METEOR is tailored for assessing the overall quality of image descriptions. Nonetheless, research has indicated a weak correlation between these metrics and human judgment [10, 4, 11, 12]. In contrast, the consensus-based metric CIDEr measures the similarity between a generated sentence and a set of ground truth sentences authored by humans, demonstrating high agreement with human consensus. However, preparing a set of ground truth sentences in advance is a prerequisite for CIDEr. If the quantity of human-authored ground truth sentences is insufficient, CIDEr may struggle to effectively evaluate image descriptions [9]. A similar limitation is observed in

LLM4Eval: The First Workshop on Large Language Models for Evaluation in Information Retrieval, 18 July 2024, Washington DC, United States.

[*] Corresponding author.

[†] These authors contributed equally.

j.huang@uva.nl (J. Huang); h.zhu@uva.nl (H. Zhu); y.shen@uva.nl (Y. Shen); s.rudinac@uva.nl (S. Rudinac); a.m.pacces@uva.nl (A. M. Pacces); E.Kanoulas@uva.nl (E. Kanoulas)

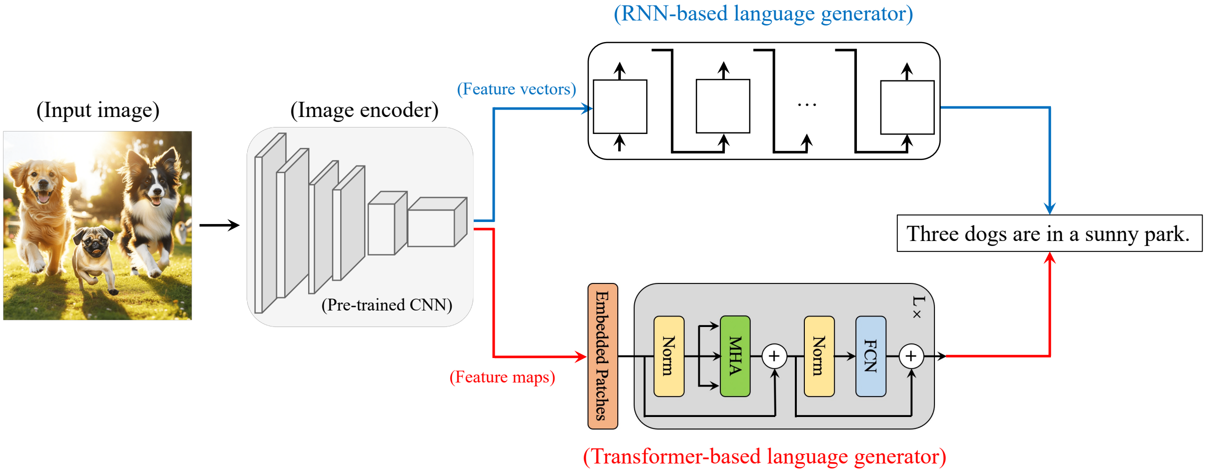


Figure 1: Flowchart for image captioning. Existing image captioning architectures can be broadly categorized into two groups: those based on the recurrent neural network (RNN) and those based on the transformer architecture. To aid comprehension, we represent RNN-based methods with blue paths and transformer-based approaches with red paths. The process involves feeding an input image through an image encoder for feature extraction, followed by a language generator to produce text-based descriptions using the extracted image features.

the CLAIR method [13] and other aforementioned approaches. Some metrics involve caption ranking [12] but are limited in evaluating novel image descriptions.

In addressing the above challenge, we present a novel framework for evaluating image descriptions. This framework is rooted in the utilization of a modern LLM approach, e.g., GPT-4 [14] or Gemini [15], capable of generating images. The advancement of LLMs [16, 17], exemplified by models like GPT-4, empowers us to provide textual descriptions, i.e., prompt, for generating images that closely correspond and align with the semantic meaning conveyed in the given text. The underlying design philosophy of the proposed framework hinges on the idea that if an image captioning model is validated as effective, the generated image description by the model should be sufficiently accurate to reconstruct the same or a highly similar image compared to the original input image, relying on LLMs. The ongoing evolution of LLM technology forms the bedrock of the proposed framework.

Starting with the definition of the image captioning task, as illustrated in Figure 1, our proposed framework begins by taking an image as input. Subsequently, this input undergoes processing through a given image captioning model, generating a textual description for the initial image. Following this, a given LLM, such as GPT-4, is employed to generate an image based on the textual description. Then, we extract the image features from both the original input image and the LLM-generated image, and assess their similarity using the cosine similarity metric. It is worth noting that human-annotated reference captions are not needed in our proposed evaluation framework. In the proposed evaluation framework, a high cosine similarity score is anticipated if the generated text-based description is of sufficient quality, signifying that the LLM can accurately reproduce an image highly similar to the original input. Conversely, if the generated text-based description lacks accuracy, the image produced by the LLM will deviate from the original input image and lead to a low cosine similarity score. This incongruity suggests the suboptimal performance of the image captioning model. Consequently, the proposed framework proves valuable for evaluating the efficacy of a given image captioning model.

The main contributions of this work are summarized as follows:

- **Innovative Framework for Image Captioning Model Evaluation:** We present a novel framework that relies on the utilization of an LLM, such as GPT-4 or Gemini, to evaluate the quality of image descriptions generated by an image captioning model. The proposed evaluation framework does not necessitate human-annotated reference captions.
- **Human Evaluation of the Framework:** To verify the effectiveness of our evaluation framework,

we introduce a human-annotated dataset and conduct human evaluations.

- **Comprehensive Experiments on Established Datasets:** We perform extensive experiments to demonstrate the efficacy of the proposed evaluation framework using widely-used image captioning datasets.

2. Related Work

In this section, we begin by reviewing existing related literature, covering topics such as the existing image captioning methods, the evolution of automated metrics, and the latest advancements in LLM technology.

2.1. Image Captioning Methods

The encoder-decoder network architecture has become a cornerstone in the field of image captioning, as evidenced by various studies [18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34]. Typically, these networks employ a CNN as the encoder for extracting global image features, and an RNN as the decoder for generating word sequences. [35] introduces a method for generating referring expressions, which are descriptions for specific objects or regions within an image. In [36], the bidirectional LSTM-based method for image captioning takes advantage of both past and future information to learn long-term visual-language interactions. Attention mechanisms have significantly enhanced the performance of image captioning models. [37] introduces an area-based attention model that predicts the next word and the corresponding image regions at each RNN timestep. While these advancements represent significant strides, they predominantly focus on single-image based description generation. However, certain abstract concepts or descriptions might not be fully captured using only image data [38, 39]. [28, 27] have explored the use of expert-defined keyword sequences to augment model capabilities in generating more accurate and contextually relevant descriptions. Recent advancements have also explored transformer-based architectures, such as Vision Transformers (ViT), which have shown promise in capturing finer details and global context in images for caption generation [40]. Furthermore, the integration of multimodal learning approaches, where models are trained on both visual and textual data, has led to significant improvements in generating contextually richer and more nuanced image descriptions [41].

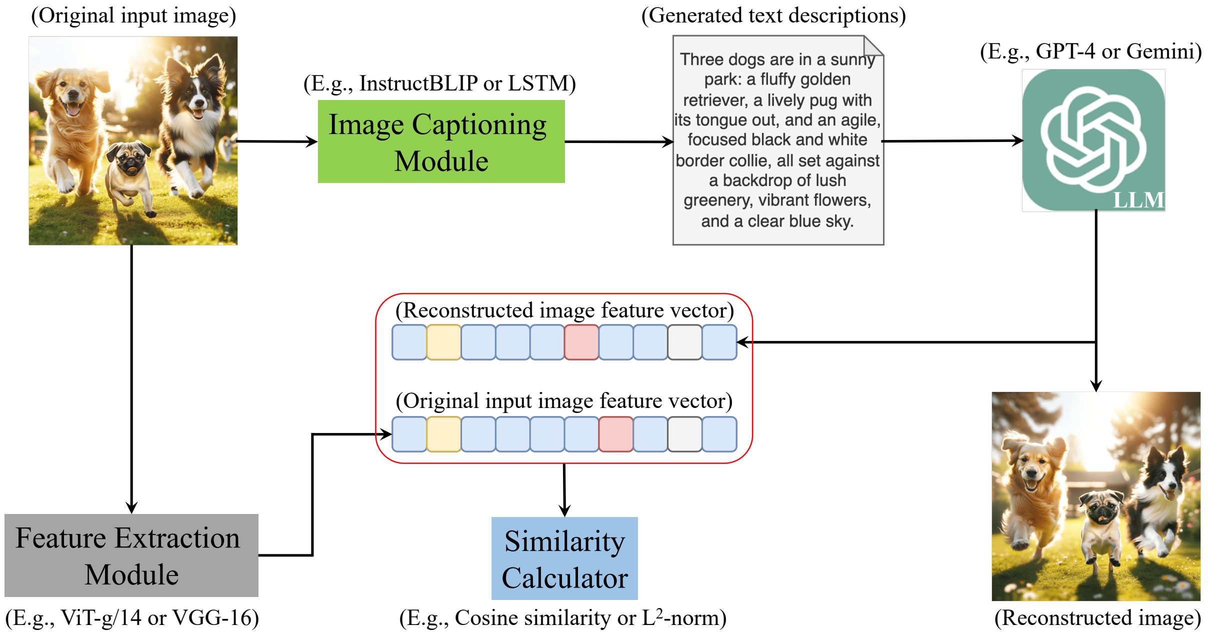
The domain of medical image captioning has witnessed significant advancements, particularly through methods that meld human expertise with algorithmic prowess. [42] has developed a Hybrid Retrieval-Generation Reinforced Agent, which integrates human prior knowledge with AI-based caption generation for medical images. This agent alternates between a generative module and a retrieval mechanism that utilizes a template database reflecting human expertise, thereby producing multi-faceted, sequential sentences. [39] has contributed to this field with a multi-task learning framework that simultaneously predicts tags and generates captions. Their method, which focuses on abnormal areas in chest radiology images using an attention mechanism and a hierarchical LSTM, offers detailed descriptions. These methods primarily focus on generating reports for chest radiology images, which are structurally different in terms of object size and detail compared to retinal images [38, 43, 27]. Additionally, the color features in chest radiology and retinal images differ significantly, with the former being predominantly grey-scale and the latter being colorful [38, 27]. Most existing methods rely primarily on the image input for caption generation. Recent advancements also include the enhancement of the CNN-RNN framework with the TransFuser model [28]. This model adeptly combines features from different modalities and addresses the challenge of incorporating unordered keyword sequences with visual inputs, minimizing information loss [28]. This development represents a significant stride in medical image captioning, reflecting the growing complexity and capability of these methods. Further progress in deep learning, particularly the application of ViTs, has offered promising results in medical imaging [44]. ViTs excel in capturing intricate details and providing a broader context for more accurate medical image analysis and caption generation.

The evaluation framework proposed in this paper is versatile and capable of assessing any existing image captioning approaches.

2.2. Automatic Metrics for Image Captioning

The evolution of image captioning has been significantly influenced by the development and application of automatic metrics for evaluating caption quality [7, 8, 45, 9, 46, 47]. These metrics guide the training of captioning models and provide a scalable means for performance assessment. The BLEU score, a pioneering metric by [7], gauges n-gram precision in generated text against a reference. ROUGE, developed by [8], emphasizes recall through the overlap of N-grams and longest common subsequences. Subsequent innovations introduced refined approaches. METEOR, by [45], aligns more closely with human judgment by incorporating synonym matching and stemming.In [9], the CIDEr metric, specifically designed for image captioning, assesses the similarity of generated captions to a set of reference captions. The SPICE metric by [46] evaluates semantic content and the depiction of objects, attributes, and relationships. Additionally, the NLG-Eval toolkit by [47] provides a comprehensive suite of metrics for a more holistic evaluation of natural language generation. However, these metrics have limitations. Metrics like BLEU and ROUGE often fail to capture the contextual nuances of captions [7, 8]. The challenge of evaluating creativity and novelty in caption generation is also evident, as automated metrics may penalize deviations from standard references [9, 46]. Recently, advancements like BERTScore [48] and CLIPScore [49], which utilize contextual embeddings and visual-textual alignment, respectively, have been proposed to address these challenges.

In this study, human evaluation is employed to validate the effectiveness of the proposed evaluation framework.


(Original input image)
(E.g., InstructBLIP or LSTM)
Image Captioning Module
(Generated text descriptions)
(E.g., GPT-4 or Gemini)
LLM
(Reconstructed image feature vector)
(Original input image feature vector)
Feature Extraction Module
(E.g., ViT-g/14 or VGG-16)
Similarity Calculator
(E.g., Cosine similarity or L2-norm)
(Reconstructed image)

Three dogs are in a sunny park: a fluffy golden retriever, a lively pug with its tongue out, and an agile, focused black and white border collie, all set against a backdrop of lush greenery, vibrant flowers, and a clear blue sky.

Figure 2: Flowchart of the proposed evaluation framework. The proposed framework consists of four main components: an image captioning module, an image feature extractor, a large language model (LLM), and a similarity calculator. The image captioning module employs a chosen model to process an input image and generate textual descriptions. The image feature extractor is tasked with extracting features from the input image. The LLM utilizes the text descriptions produced by the image captioning model to generate the corresponding image. Finally, the similarity calculator computes the similarity between the features of the input image and the image generated by the LLM.

2.3. Large Language Models

The advent of LLMs has significantly reshaped the landscape of natural language processing (NLP) and Artificial Intelligence (AI). Pioneering models such as GPT, developed by [14], and BERT by [50], have marked critical milestones in this evolution. These models, characterized by their vast number of parameters and advanced deep learning architectures, have enhanced the capacity to understand and generate human language, excelling in diverse tasks like translation, summarization, and question-answering [50, 51]. The efficacy of LLMs such as GPT, which utilizes a transformer-based architecture, stems from their comprehensive training across a broad spectrum of internet text, enabling the generation of coherent and contextually pertinent language [51]. BERT's introduction of bidirectional transformers has revolutionized pre-training in language understanding, showing remarkable efficiency in tasks requiring intricate contextual comprehension [50]. The incorporation of attention mechanisms, as conceptualized by [17], has further refined these models' ability for nuanced understanding and text generation. In the realm of image captioning, the deployment of LLMs like GPT-3 has brought transformative changes. GPT-3's adeptness in image captioning tasks is a testament to its sophisticated transformer-based architecture and comprehensive training on a wide array of internet text. This extensive training enables GPT-3 to intricately understand and generate content that accurately aligns with both textual and visual contexts, producing coherent, contextually relevant, and detailed image descriptions [51]. The fusion of LLMs with advanced computer vision techniques has been a significant leap forward, leading to the development of more sophisticated systems. These systems are now better equipped to interpret and describe complex visual data with greater accuracy and nuance [52]. This integration highlights the evolving capability of AI to understand and convey the subtleties of visual information, mirroring a more human-like perception and articulation of images. This advancement in image captioning technology is pivotal in enhancing how machines process and narrate visual data, bridging the gap between visual perception and linguistic expression. Furthermore, the use of LLMs goes beyond generating captions to evaluating their quality. A notable method in this regard is CLAIR [13], which leverages zero-shot language modeling to assess caption quality. CLAIR shows a stronger correlation with human judgment compared to traditional metrics like BLEU, ROUGE, METEOR, and CIDEr. By soliciting an LLM to rate how likely a candidate caption accurately describes an image relative to a set of reference captions, CLAIR outperforms language-only measures, approaching human-level correlation. However, CLAIR requires a set of human-annotated reference captions to function, without which it cannot be applied.

In this work, the proposed approach leverages modern LLMs like GPT-4 for an innovative and comprehensive evaluation. We use LLMs to reverse-engineer the image captioning process, generating images from textual descriptions to assess caption accuracy. This method offers a unique advantage in evaluating the semantic richness and contextual relevance of captions. By comparing the generated images with the original, our approach provides a direct, visual assessment of caption quality, moving beyond mere textual analysis. This novel methodology not only aligns with human perception but also embraces the creativity and diversity inherent in image captioning, offering a more rounded and practical evaluation framework.

3. Methodology

The proposed evaluation framework comprises several key components: an image captioning module, an LLM-based text-to-image generator, an image feature extraction module, and a similarity calculator, as depicted in Figure 2. Each of these components will be introduced in detail in the following subsections. Furthermore, to ensure the validity of the evaluation results based on our framework—specifically, their alignment with human judgment—we introduce a human-annotated image captioning dataset to validate the effectiveness of the proposed framework.

3.1. Image Captioning Module

The module incorporates an image captioning model, which will undergo evaluation using the proposed framework. This module takes an image as input and generates a text-based description as output. To facilitate user comprehension of the proposed evaluation framework, we utilize the InstructBLIP model [53] as an illustrative example in Section 4. This demonstration showcases the entire process of leveraging the proposed framework to evaluate a given image captioning model, making it easily understandable for users.

3.2. LLM-based Text-to-Image Generator

Numerous studies [14, 15, 16] have demonstrated the proficiency of LLM-based image generators, exemplified by models like GPT-4, in producing high-quality images that closely align with the semantic meaning of provided text-based prompts. Specifically, DALL-E, functioning as an image generation model within GPT-4, a variant of GPT-3 boasting 12 billion parameters, is engineered by OpenAI to generate images based on textual descriptions, drawing from a dataset comprising text-image pairs. Its versatile capabilities include crafting anthropomorphized versions of animals and objects, seamlessly combining unrelated concepts, rendering text, and applying transformations to existing images. In the context of the proposed framework, the LLM-based image generator utilizes the text-based image description generated by a preceding image captioning model. If the image captioning model performs well, generating a high-quality and accurate image description, the LLM-based image generator subsequently creates an image that is similar to the original input image. This connection highlights the interplay between effective image captioning and the generation of corresponding images by the LLM-based approach.

3.3. Image Feature Extraction Module

The image feature extraction module primarily consists of a pre-trained image encoder. This module takes an image as input and produces a feature vector representing the input image as output. To enhance user understanding of the proposed evaluation framework, we employ ViT-g/14 [54] as a demonstrative example for image feature extraction in Section 4. ViT-g/14 is a vanilla ViT pre-trained for reconstructing masked-out image-text aligned vision features conditioned on visible image patches. Through this pretext task, the model efficiently scales up to one billion parameters, achieving notable performance across various vision downstream tasks, including image recognition, video action recognition, object detection, instance segmentation, and semantic segmentation, all without extensive supervised training. This demonstration in Section 4 highlights the complete process, encompassing image feature extraction for calculating similarity scores between the input and generated images. It illustrates how the proposed framework can be leveraged to assess a given image captioning model, providing users with a clear understanding. It is worth noting that the image feature extractor can be substituted with other pre-trained CNNs, such as VGG-16 [55] or ResNet-52 [56].

3.4. Similarity Calculator

Cosine similarity, as defined in Equation (1), serves as a metric for quantifying the similarity between two vectors in a multi-dimensional space. It evaluates the cosine of the angle between these vectors, offering insight into their degree of similarity or dissimilarity. The advantage of cosine similarity lies in its ability to assess directional similarity rather than magnitude, rendering it robust against variations in scale and orientation. This characteristic makes it a widely adopted metric in diverse domains, including image processing and NLP. In these fields, cosine similarity is frequently employed to assess the similarity between images, documents, or sentences represented as vectors in high-dimensional spaces. The cosine similarity value $\text{CosSim}(\cdot\,,\cdot) \in [-1, 1]$, where a value of 1 signifies that the vectors are identical, 0 indicates orthogonality (i.e., no similarity), and -1 indicates complete dissimilarity or

“A dog wearing a leash laying next to an orange frisbee”,

“A dog with a collar on laying down next to a frisbee”,

“A dog lies in the grass next to a Frisbee.”,

“A frisbee on the ground next to a dog sitting in the grass”,

“A dog that is laying on the ground next to a Frisbee.”

“All of these sheep have coats that are ready for shearing.”,

“some sheep standing around by a wooden wall”,

“Five sheep are standing and sitting in their enclosure.”,

“One sheep lies down as four others stand near.”,

“A group of five sheep wait outside a barn.”

“A skateboarder is jumping down a flight of stairs.”,

“A skaterboarder getting major air over some stairs during a night time shoot”,

“A skate boarder jumping down some stairs at night”,

“A skateboarder riding down a flight of stone stairs”,

“A young man skateboarding over a flight of steps”

“A person is on a living room couch watching TV and there is a stuffed panda bear and a purse on the table.”,

“A child watches television while a panda bear sits by a purse.”,

“A simple living room with a panda on the coffee table”,

“A black and white stuffed koala bear is in the room.”,

“A stuffed panda is on the living room table.”

Figure 3: Dataset examples. To provide a clearer insight into the introduced human-annotated dataset, we have randomly selected four examples for illustrative purposes. Each image in the dataset is accompanied by five human-annotated descriptions that vividly depict the content of the image.

opposition.

$$\text{CosSim}(\mathbf{i_o}, \mathbf{i_g}) = \frac{\mathbf{i_o} \cdot \mathbf{i_g}}{||\mathbf{i_o}||\,||\mathbf{i_g}||}, \tag{1}$$

where $\mathbf{i_o} \cdot \mathbf{i_g}$ denotes the dot product (also known as the inner product) of the original input image feature vector $\mathbf{i_o}$ and the LLM-generated image feature vector $\mathbf{i_g}$. $||\mathbf{i_o}||$ and $||\mathbf{i_g}||$ represent the Euclidean norms (also known as the magnitudes or lengths) of vectors $\mathbf{i_o}$ and $\mathbf{i_g}$, respectively. In words, cosine similarity measures the cosine of the angle between two vectors, which represents their similarity in direction and magnitude.

3.5. Human-annotated Image Captioning Dataset

The Microsoft Common Objects in Context (MSCOCO) dataset is a comprehensive resource widely used across various image recognition tasks including object detection, segmentation, and captioning. Originally, the MSCOCO Captions dataset comprised over $330,000$ images, each meticulously annotated with 80 object categories. Notably, both the training and validation sets feature each image accompanied by five distinct human-generated captions. This dataset holds significant importance within the realm of computer vision research, serving as a cornerstone for the development and evaluation of numerous state-of-the-art object detection and segmentation models. In our study, we enhance the existing MSCOCO Caption dataset by incorporating an additional $30,000$ human-annotated image-description pairs. This augmented dataset serves as the basis for evaluating the alignment of our proposed evaluation method with human-annotated image descriptions. To aid in understanding the dataset, several examples from the dataset are provided in Figure 3.

4. Experiments and Analysis

In this section, our goal is to evaluate the effectiveness of the proposed evaluation framework designed for image captioning models. To achieve this, we will validate our framework using both the widely adopted human-annotated image captioning datasets and our newly introduced dataset, the details of which are outlined in the Section 3.5. Since all datasets have undergone human annotation, our primary objective in this assessment is to ascertain whether the evaluation results obtained through our proposed framework align with human consensus or judgment. To elaborate, a correct caption—matching the human-annotated counterpart—should yield a substantial cosine similarity score between the generated and original images, as measured by our evaluation framework. Conversely, an incorrect caption—deviating from the human-annotated version—should result in a comparatively smaller cosine similarity score. This approach allows us to empirically validate the effectiveness of our proposed evaluation framework in aligning with human judgment.

4.1. Experimental Settings

To illustrate the application of the proposed framework for evaluating an image captioning model, we employ the InstructBLIP [57] model in our image captioning module. This model is equipped with the pre-trained language model Vicuna-7B [58] to generate image descriptions. Image captions are generated using the prompt "<Image> A short image caption:", guiding the model to produce sentences of fewer than 100 tokens, excluding special symbols. For text-to-image generation, GPT-4 with the built-in diffusion model DALL-E-3 is employed. Notably, the diffusion model can be replaced by Stable Diffusion models [59], utilizing a fixed, pre-trained encoder (ViT-g/14) [60], and the entire diffusion model is pre-trained on the LAION-2B dataset [61]. Human evaluation serves as the validation method for the proposed framework. Each image in the dataset comes with five human-annotated image captions, and performance is quantified using the average cosine similarity score, as detailed in Section 4.3. The experiments are conducted using two NVIDIA-A6000 GPUs.

4.2. Datasets

MSCOCO Dataset [62]. The MSCOCO dataset comprises two primary components: the images and their corresponding annotations. The images are organized into a directory hierarchy, with top-level directories for the train, validation, and test sets. Annotations are provided in JSON format, with each file corresponding to a single image. Each annotation includes details such as the image file name, dimensions (width and height), a list of objects with their respective class labels (e.g., "person," "car"), bounding box coordinates (x, y, width, height), segmentation mask (in polygon or RLE format), keypoints and their positions (if available), and five captions describing the scene. Additional information provided by the MSCOCO dataset includes image super categories, license details, and coco-stuff annotations (pixel-wise annotations for stuff classes in addition to the 80 object classes). The MSCOCO dataset provides various types of annotations, including object detection with bounding box coordinates and full segmentation masks for 80 different objects, stuff image segmentation with pixel maps displaying 91 amorphous background areas, panoptic segmentation identifying items in images based on 80 "things" and 91 "stuff" categories, dense pose annotations featuring over $39,000$ photos and mapping between pixels and a template for over $56,000$ tagged persons, 3D model annotations and natural language descriptions for each image, and keypoint annotations for over $250,000$ persons annotated with key points such as the right eye, nose, and left hip.

Flickr30k Dataset [63]. The authors in [63] advocate for utilizing the visual denotations of linguistic expressions, represented by the set of images they describe, to define new denotational similarity metrics. These metrics, as demonstrated in [63], prove to be at least as advantageous as distributional similarities for tasks requiring semantic inference. The computation of these denotational similarities involves the construction of a denotation graph—a subsumption hierarchy over constituents and their denotations. This graph is established using a substantial corpus comprising $30,000$ images and $150,000$ descriptive captions. The creation of this denotation graph involves the development of an image caption corpus by the authors in [63], consisting of $158,915$ crowd-sourced captions elucidating $31,783$ images. This corpus serves as an extension of their previous work on the Flickr8k Dataset. The new images and captions specifically focus on individuals engaged in everyday activities and events.

4.3. Effectiveness Analysis of the Proposed Evaluation Framework

Human Evaluation Using the Proposed Dataset. The dataset introduced in this work, consisting of pairs of images and captions, has undergone human annotation. Each image is accompanied by five distinct human-generated captions. The details of our human evaluation process are outlined below. In Step 1, we directly utilize the human-annotated ground truth caption to generate an image through a text-to-image LLM, such as GPT-4 or Gemini. In Step 2, we extract the image features of both the ground truth caption's corresponding image and the image generated by the text-to-image LLM. In Step 3, we apply the cosine similarity formula from Section 3.4 to compute the cosine similarity scores between these two sets of image features. Given that the caption is a human-annotated ground truth description, accurately portraying the corresponding image, we expect the similarity score from Step 3 to be high. Conversely, if a caption inaccurately describes a given image, the cosine similarity score from Step 3 should be low. Consistency between the experimental result and these expectations indicates the effectiveness of the proposed evaluation framework in aligning with human consensus.

The evaluation results depicted in Figure 4 reveal notable insights. The blue lines in Figure 4 illustrate the impact of the provided captions on the cosine similarity scores. Specifically, when the provided caption matches the correct human-annotated description (upper blue line), the average cosine similarity score reaches approximately 0.67. Conversely, when the caption is incorrect (lower blue line), the average cosine similarity score drops to around 0.47. This discrepancy results in a similarity gap of approximately 0.2. These findings underscore the effectiveness of the proposed evaluation framework, as it closely aligns with human judgment. It is noteworthy that the robustness of this human evaluation method is attributed to the remarkable text-to-image generation capabilities of modern LLM models. Widely recognized models such as GPT-4 and Gemini have been extensively acclaimed in various studies and by the broader community [14, 15, 16].

Assessment Using MSCOCO and Flickr30k Datasets. Figure 4 reveals consistent trends in the evaluation results across MSCOCO, Flickr30k, and our dataset. Similar patterns are observed in MSCOCO and Flickr30k, where there is a notable decrease in the average cosine similarity when the model-generated image caption differs from the human-annotated ground truth caption. These findings affirm the effectiveness and reliability of the proposed evaluation framework for assessing image captioning models.

Qualitative Analysis. To gain deeper insights into the performance of the proposed evaluation framework, we present qualitative results in Figure 5 and Figure 6. In Figure 5, we observe that the human-annotated ground truth captions and the model-predicted captions exhibit poor alignment in these four examples. Given the accurate image generation capabilities of existing LLMs based on text-based prompts, the accuracy of model-generated image descriptions is crucial. However, in these instances, all predicted captions are incorrect, resulting in LLM-generated images that significantly differ from the ground truth images. Consequently, this discrepancy contributes to the low cosine-based similarity scores.

In Figure 6, these two examples illustrate a strong alignment between the model-generated descriptions and the human-generated ground truth captions. Hence, this alignment results in LLM-generated images that closely resemble the ground truth images. As a result, when calculating cosine similarity scores based on the image features extracted from the LLM-generated and ground truth images, the scores are notably high. We also calculate scores based on these metrics to highlight the advantage of our proposed method over the aforementioned text-based evaluation metrics. In Figure 6, we observe that despite the model-generated image captions closely matching the ground truth captions, the scores based on text-based evaluation metrics are comparatively low. This observation underscores the superiority of our proposed evaluation framework over existing text-based evaluation metrics for image captioning models.

$$\mathrm{BP} = \begin{cases} 1 & \text{if} \quad c > r \\ \exp(1 - \frac{r}{c}) & \text{if} \quad c \leq r \end{cases}; \mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \mathrm{log} p_n\right), \tag{2}$$

where r represents the effective length of the ground truth text, c signifies the length of the predicted

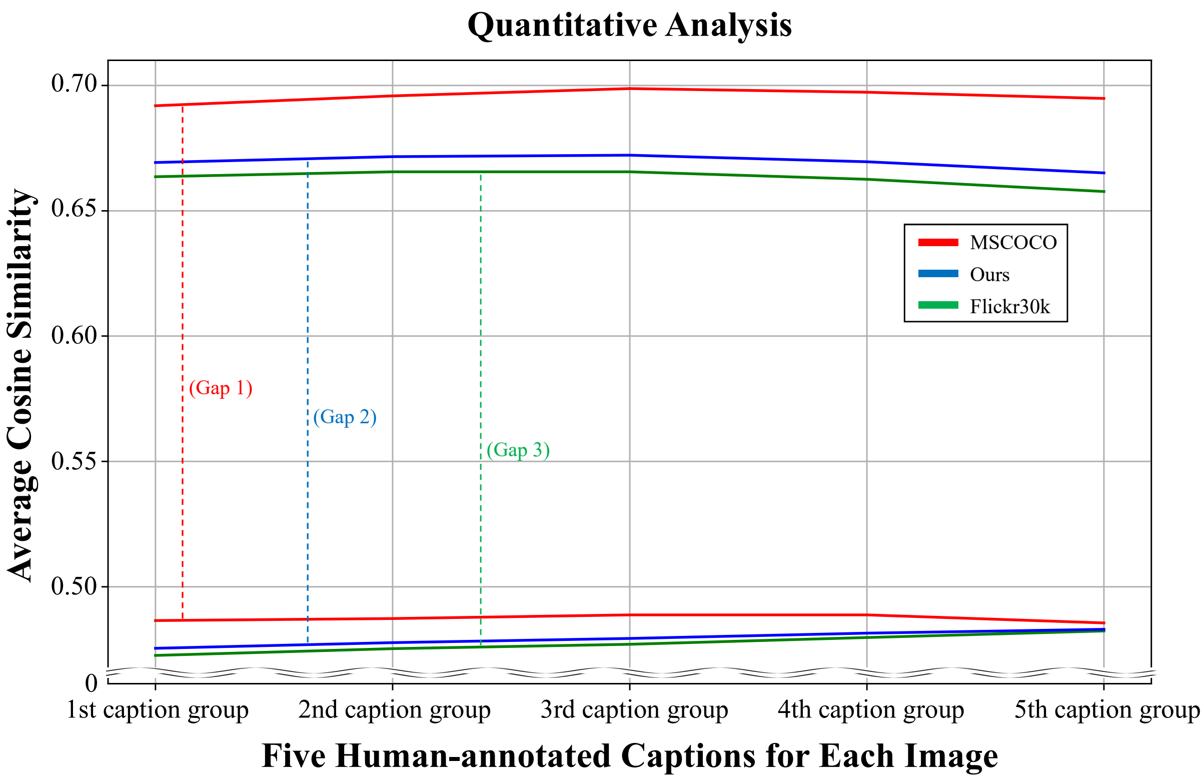
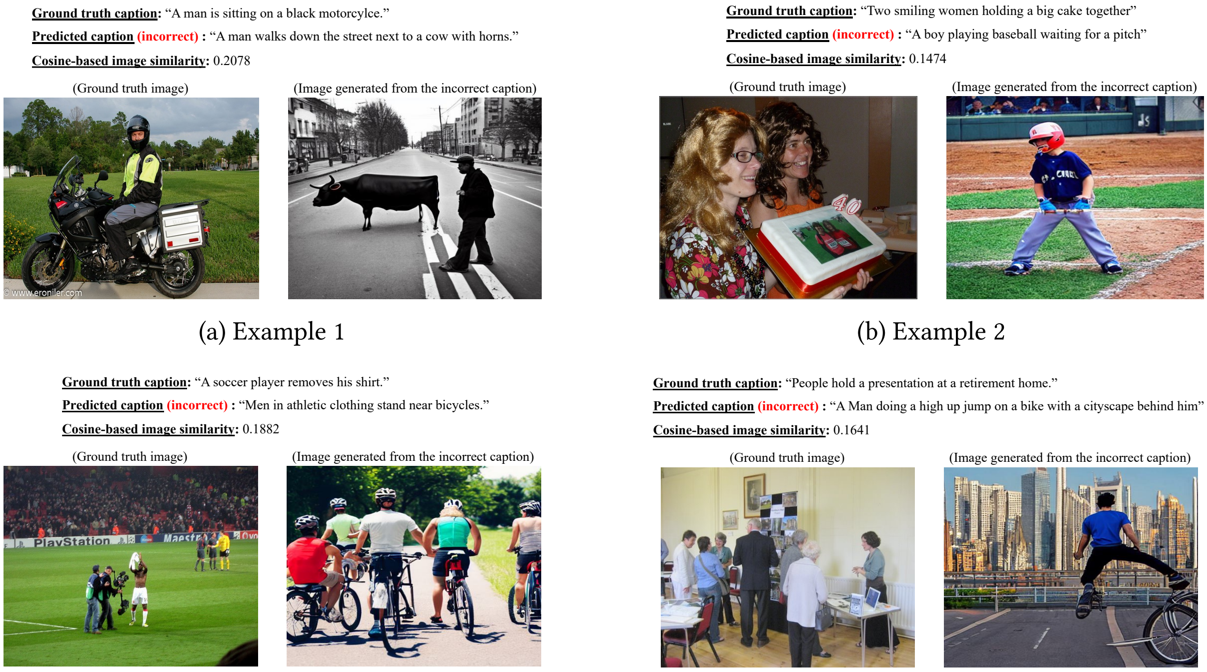


Figure 4: Human evaluation results. The outcomes are derived from three datasets: MSCOCO (highlighted in red), Flickr30k (highlighted in green), and our dataset (highlighted in blue). The top three lines represent scenarios where the provided caption aligns with the correct human-annotated description, while the bottom three lines represent scenarios where the caption is incorrect. "Gap 1", "Gap 2", and "Gap 3" signify the disparities in average cosine similarity scores. We observe that these gaps are approximately 0.2, indicating the influence of the provided captions on the cosine similarity scores. A larger gap indicates a substantial mismatch between the human-annotated image description and the provided or model-generated caption, whereas a smaller gap suggests a higher degree of alignment.

text, and BP stands for brevity penalty. The geometric mean of the adjusted n-gram precisions p_n is calculated using n-grams up to a length of N, with positive weights w_n that sum to 1.

5. Conclusion

In this study, we have introduced a novel framework for evaluating automatically generated image descriptions, aiming to overcome the limitations of existing evaluation metrics like BLEU, ROUGE, METEOR, and CIDEr. Our framework leverages advancements in LLMs such as GPT-4 or Gemini to utilize image descriptions generated by an image captioning model for creating corresponding images. By quantifying the cosine similarity between the representation of the original input image in the image captioning model and the representation of the LLM-generated image, we can effectively assess the model's performance without relying on human-annotated reference captions. Through extensive experiments on the established datasets like Flickr30k and MSCOCO, we have demonstrated the effectiveness of the proposed evaluation framework. Our experimental results suggest that the proposed framework's performance closely correlates with human judgment, offering a valuable method



Ground truth caption: "A man is sitting on a black motorcycle."

Predicted caption **(incorrect)** : "A man walks down the street next to a cow with horns."

Cosine-based image similarity: 0.2078

(a) Example 1

Ground truth caption: "Two smiling women holding a big cake together"

Predicted caption **(incorrect)** : "A boy playing baseball waiting for a pitch"

Cosine-based image similarity: 0.1474

(b) Example 2

Ground truth caption: "A soccer player removes his shirt."

Predicted caption **(incorrect)** : "Men in athletic clothing stand near bicycles."

Cosine-based image similarity: 0.1882

(c) Example 3

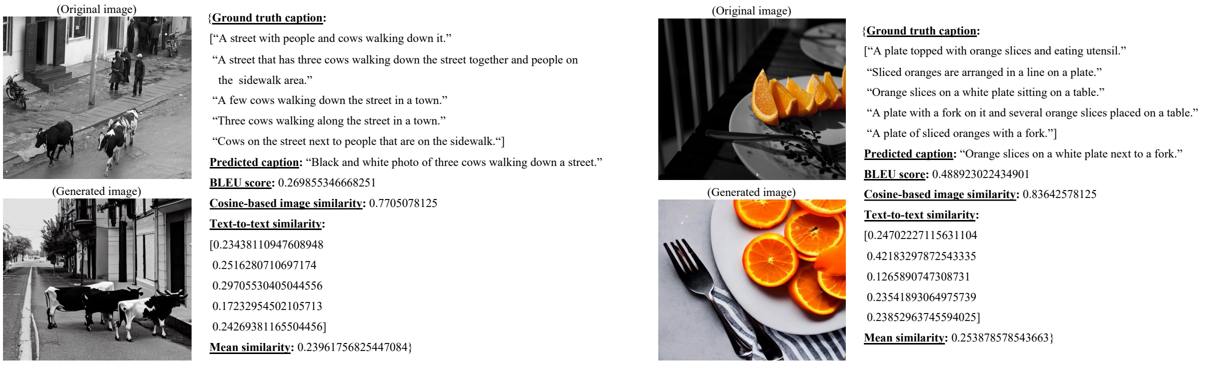
Ground truth caption: "People hold a presentation at a retirement home."

Predicted caption **(incorrect)** : "A Man doing a high up jump on a bike with a cityscape behind him"

Cosine-based image similarity: 0.1641

(d) Example 4

Figure 5: Impact of incorrect image captions. Due to LLMs' proficiency in generating images accurately from provided text prompts, inconsistencies between model-generated image captions and human-annotated ground truth descriptions can lead to discrepancies in the generated images. Leveraging this observation, we propose an evaluation framework to assess the performance of image captioning models without using human-annotated ground truth captions.



{**Ground truth caption**:
["A street with people and cows walking down it."
"A street that has three cows walking down the street together and people on the sidewalk area."
"A few cows walking down the street in a town."
"Three cows walking along the street in a town."
"Cows on the street next to people that are on the sidewalk."]
Predicted caption: "Black and white photo of three cows walking down a street."
BLEU score: 0.269855346668251
Cosine-based image similarity: 0.7705078125
Text-to-text similarity:
[0.23438110947608948
0.2516280710697174
0.29705530405044556
0.17232954502105713
0.24269381165504456]
Mean similarity: 0.23961756825447084}

(a) Example 1

{**Ground truth caption**:
["A plate topped with orange slices and eating utensil."
"Sliced oranges are arranged in a line on a plate."
"Orange slices on a white plate sitting on a table."
"A plate with a fork on it and several orange slices placed on a table."
"A plate of sliced oranges with a fork."]
Predicted caption: "Orange slices on a white plate next to a fork."
BLEU score: 0.488923022434901
Cosine-based image similarity: 0.83642578125
Text-to-text similarity:
[0.24702227115631104
0.42183297872543335
0.1265890747308731
0.23541893064975739
0.23852963745594025]
Mean similarity: 0.253878578543663}

(b) Example 2

Figure 6: Limitations of text-based evaluation metrics in image captioning. See Equation (2) for the calculation of the BLEU score. "Predicted caption" refers to the caption generated by the InstructBLIP model. "Text to text similarity" indicates the cosine similarity between the human-annotated ground truth caption and the model-generated caption using text-based CLIP embeddings. "Mean similarity" represents the average of the five values of "Text to text similarity".

for evaluating the effectiveness of image captioning models. Additionally, human evaluations conducted on our introduced dataset validate the framework's efficacy in capturing various aspects such as grammaticality, coverage, correctness, and truthfulness in automatically generated image descriptions. Moving forward, the proposed framework presents new opportunities for evaluating image captioning models, offering a more efficient and reliable alternative to traditional human evaluations and existing automated evaluation metrics. It is designed to complement, rather than replace, human judgment. In summary, our work contributes to the ongoing development of robust evaluation frameworks for image captioning models, bridging the gap between automated metrics and human judgment, and driving advancements in this field.

References

[1] M. Mitchell, X. Han, J. Hayes, Midge: Generating descriptions of images, in: INLG 2012 Proceedings of the Seventh International Natural Language Generation Conference, 2012, pp. 131–133.

[2] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, B. Schiele, Translating video content to natural language descriptions, in: Proceedings of the IEEE international conference on computer vision, 2013, pp. 433–440.

[3] Y. Yang, C. Teo, H. Daumé III, Y. Aloimonos, Corpus-guided sentence generation of natural images, in: Proceedings of the 2011 conference on empirical methods in natural language processing, 2011, pp. 444–454.

[4] D. Elliott, F. Keller, Image description using visual dependency representations, in: Proceedings of the 2013 conference on empirical methods in natural language processing, 2013, pp. 1292–1302.

[5] M. Yatskar, M. Galley, L. Vanderwende, L. Zettlemoyer, 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), 2014, pp. 110–120.

[6] R. Linkert, A technique for measuring attitude scale, Psychometrical 140 (1932) 40–55.

[7] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[8] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.

[9] R. Vedantam, C. Lawrence Zitnick, D. Parikh, Cider: Consensus-based image description evaluation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566–4575.

[10] S. by Saheel, Baby talk: Understanding and generating image descriptions (????).

[11] C. Callison-Burch, M. Osborne, P. Koehn, Re-evaluating the role of bleu in machine translation research, in: 11th conference of the european chapter of the association for computational linguistics, 2006, pp. 249–256.

[12] M. Hodosh, P. Young, J. Hockenmaier, Framing image description as a ranking task: Data, models and evaluation metrics, Journal of Artificial Intelligence Research 47 (2013) 853–899.

[13] D. Chan, S. Petryk, J. E. Gonzalez, T. Darrell, J. Canny, Clair: Evaluating image captions with large language models, arXiv preprint arXiv:2310.12971 (2023).

[14] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[15] G. Team, R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al., Gemini: a family of highly capable multimodal models, arXiv preprint arXiv:2312.11805 (2023).

[16] T. Wu, S. He, J. Liu, S. Sun, K. Liu, Q.-L. Han, Y. Tang, A brief overview of chatgpt: The history, status quo and potential future development, IEEE/CAA Journal of Automatica Sinica 10 (2023) 1122–1136.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, arXiv preprint arXiv:1706.03762 (2017).

[18] X. Xiao, L. Wang, K. Ding, S. Xiang, C. Pan, Deep hierarchical encoder–decoder network for image captioning, IEEE Transactions on Multimedia 21 (2019) 2942–2956.

[19] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3156–3164.

[20] T. J. Jun, J. Kweon, Y.-H. Kim, D. Kim, T-net: Nested encoder–decoder architecture for the main vessel segmentation in coronary angiography, Neural Networks 128 (2020) 216–233.

[21] J.-H. Huang, M. Alfadly, B. Ghanem, Robustness analysis of visual qa models by basic questions, VQA Challenge and Visual Dialog Workshop, CVPR (2018).

[22] J.-H. Huang, M. Alfadly, B. Ghanem, Vqabq: Visual question answering by basic questions, VQA Challenge Workshop, CVPR (2017).

[23] J.-H. Huang, Robustness analysis of visual question answering models by basic questions, King Abdullah University of Science and Technology, Master Thesis (2017).

[24] J.-H. Huang, C. D. Dao, M. Alfadly, B. Ghanem, A novel framework for robustness analysis of visual qa models, in: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, 2019, pp. 8449–8456.

[25] J.-H. Huang, M. Alfadly, B. Ghanem, M. Worring, Improving visual question answering models through robustness analysis and in-context learning with a chain of basic questions, arXiv preprint arXiv:2304.03147 (2023).

[26] J.-H. Huang, M. Alfadly, B. Ghanem, M. Worring, Assessing the robustness of visual question answering, arXiv preprint arXiv:1912.01452 (2019).

[27] J.-H. Huang, C.-H. H. Yang, F. Liu, M. Tian, Y.-C. Liu, T.-W. Wu, I. Lin, K. Wang, H. Morikawa, H. Chang, et al., Deepopht: medical report generation for retinal images via deep models and visual explanation, in: WACV, 2021, pp. 2442–2452.

[28] J.-H. Huang, T.-W. Wu, C.-H. H. Yang, Z. Shi, I. Lin, J. Tegner, M. Worring, et al., Non-local attention improves description generation for retinal images, in: WACV, 2022, pp. 1606–1615.

[29] J.-H. Huang, T.-W. Wu, M. Worring, Contextualized keyword representations for multi-modal retinal image captioning, in: ICMR, 2021, pp. 645–652.

[30] J.-H. Huang, T.-W. Wu, C.-H. H. Yang, M. Worring, Deep context-encoding network for retinal image captioning, in: ICIP, IEEE, 2021, pp. 3762–3766.

[31] J.-H. Huang, T.-W. Wu, C.-H. H. Yang, M. Worring, Longer version for" deep context-encoding network for retinal image captioning", arXiv preprint arXiv:2105.14538 (2021).

[32] J.-H. Huang, L. Murn, M. Mrak, M. Worring, Gpt2mvs: Generative pre-trained transformer-2 for multi-modal video summarization, in: ICMR, 2021, pp. 580–589.

[33] J.-H. Huang, M. Worring, Query-controllable video summarization, in: ICMR, 2020, pp. 242–250.

[34] T.-W. Wu, J.-H. Huang, J. Lin, M. Worring, Expert-defined keywords improve interpretability of retinal image captioning, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 1859–1868.

[35] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, K. Murphy, Generation and comprehension of unambiguous object descriptions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 11–20.

[36] C. Wang, H. Yang, C. Bartz, C. Meinel, Image captioning with deep bidirectional lstms, in: Proceedings of the 24th ACM international conference on Multimedia, 2016, pp. 988–997.

[37] M. Pedersoli, T. Lucas, C. Schmid, J. Verbeek, Areas of attention for image captioning, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 1242–1250.

[38] J. Laserson, C. D. Lantsman, M. Cohen-Sfady, I. Tamir, E. Goz, C. Brestel, S. Bar, M. Atar, E. Elnekave, Textray: Mining clinical reports to gain a broad understanding of chest x-rays, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11, Springer, 2018, pp. 553–561.

[39] B. Jing, P. Xie, E. Xing, On the automatic generation of medical imaging reports, arXiv preprint arXiv:1711.08195 (2017).

[40] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[41] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, Advances in neural information processing systems 32 (2019).

[42] Y. Li, X. Liang, Z. Hu, E. P. Xing, Hybrid retrieval-generation reinforced agent for medical image

report generation, Advances in neural information processing systems 31 (2018).

[43] D. M. Tierney, J. S. Huelster, J. D. Overgaard, M. B. Plunkett, L. L. Boland, C. A. St Hill, V. K. Agboto, C. S. Smith, B. F. Mikel, B. E. Weise, et al., Comparative performance of pulmonary ultrasound, chest radiograph, and ct among patients with acute respiratory failure, Critical Care Medicine 48 (2020) 151–157.

[44] J. Chen, Y. He, E. C. Frey, Y. Li, Y. Du, Vit-v-net: Vision transformer for unsupervised volumetric medical image registration, arXiv preprint arXiv:2104.06468 (2021).

[45] S. Banerjee, A. Lavie, 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, 2005, pp. 65–72.

[46] P. Anderson, B. Fernando, M. Johnson, S. Gould, 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, Springer, 2016, pp. 382–398.

[47] S. Sharma, L. E. Asri, H. Schulz, J. Zumer, Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation, arXiv preprint arXiv:1706.09799 (2017).

[48] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, arXiv preprint arXiv:1904.09675 (2019).

[49] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, Y. Choi, Clipscore: A reference-free evaluation metric for image captioning, arXiv preprint arXiv:2104.08718 (2021).

[50] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[51] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[52] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, T. Duerig, Scaling up visual and vision-language representation learning with noisy text supervision, in: International conference on machine learning, PMLR, 2021, pp. 4904–4916.

[53] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, S. Hoi, Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023, arXiv preprint arXiv:2305.06500 (????).

[54] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, Y. Cao, Eva: Exploring the limits of masked visual representation learning at scale, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19358–19369.

[55] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).

[56] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[57] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. A. Li, P. Fung, S. C. H. Hoi, Instructblip: Towards general-purpose vision-language models with instruction tuning, ArXiv abs/2305.06500 (2023). URL: https://api.semanticscholar.org/CorpusID:258615266.

[58] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. Gonzalez, I. Stoica, Judging llm-as-a-judge with mt-bench and chatbot arena, ArXiv abs/2306.05685 (2023). URL: https://api.semanticscholar.org/CorpusID:259129398.

[59] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10684–10695.

[60] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, 2021. URL: https://api.semanticscholar.org/CorpusID:231591445.

[61] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, A. Komatsuzaki, Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, ArXiv abs/2111.02114 (2021). URL: https://api.semanticscholar.org/CorpusID:241033103.

[62] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755.

[63] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, Transactions of the Association for Computational Linguistics 2 (2014) 67–78.