Bright as the Sun: In-depth Analysis of Imagination-driven Image Captioning

Huyen Thi Thanh Tran[1] and Takayuki Okatani[1,2]

[1] RIKEN Center for AIP
[2] Graduate School of Information Sciences, Tohoku University
{tran,okatani}@vision.is.tohoku.ac.jp

Abstract. Existing studies on image captioning mainly focus on generating "literal" captions based on visual entities in images and their basic properties such as colors and spatial relationships. However, to describe images, humans use not only literal descriptions but also "imagination-driven" descriptions that characterize visual entities by some different entities; they are often more vivid, precise, and visually comprehensible by readers/hearers. Nonetheless, none of the existing studies seriously consider captions of this type. This study presents the first comprehensive analysis of the generation and evaluation of imagination-driven captions. Specifically, we first analyze imagination-driven captions in existing image captioning datasets. Then, we present the comprehensive categorizations of imagination-driven captions and their usage, discussing the (potential) issues with the current image captioning models to generate such captions. Next, compiling these captions extracted from the existing datasets and synthesizing fake captions, we create a dataset named *IdC-I* and *-II*. Using this dataset, we examine nine existing metrics of image captioning about how accurately they can evaluate imagination-driven caption generation. Last, we propose a baseline model for imagination-driven captioning. It has a built-in mechanism to select which to generate between literal and imagination-driven captions, which existing image captioning models cannot do. Experimental results demonstrate that our model performs better than six existing models, especially for imagination-driven caption generation. Dataset and code will be publicly available at: `https://github.com/TranHuyen1191/Imagination-driven-Image-Captioning`.

Keywords: Image Captioning · Imagination-driven Image Captioning.

1 Introduction

Image captioning is the task of automatically generating a description (also known as caption) in natural language for a given image. It is one of the fundamental tasks of computer vision, which has broad applicability in various areas of biomedicine, commerce, and web searching [1]. For instance, image captioning can help visually-impaired people understand the content of images and, to some extent, form similar images in their minds.

Fig. 1: Examples of image captioning. *GT* denotes human-generated descriptions: (a): "Literal" description; (b) (c) & (d): "Imagination-driven" descriptions. SAT denotes captions generated by Show-Attend-Tell [5] that is trained on the *MS COCO* dataset [6]. (a) & (b) are from *MS COCO* [6], (c) & (d) are from *ArtEmis* [7] and *Flickr30K* [8], respectively.

Existing studies on image captioning mainly focus on generating "literal" captions, which are based on visual entities in images and their basic properties (e.g., colors, spatial relationships) [2,3,4]. As an example in Fig. 1(a), the caption *"a sign that is on the side of a road"*, which is generated by an existing model of Show-Attend-Tell (SAT) [5], is based on three visual entities of *sign*, *side*, and *road*, and their relationships of *on* and *of*.

To describe images, humans actually use not only literal descriptions but also "imagination-driven" descriptions. Unlike literal descriptions, which directly describe visual entities, imagination-driven descriptions characterize visual entities by other entities, which we call *imaginary entities* in this paper. Imaginary entities are usually not presented in images, but typically share one or several common properties with visual entities. Examples of imagination-driven descriptions are shown in Fig. 1(b),(c), and (d). When observing the image in Fig. 1(b), rather than only thinking about the cupcakes, *a visual entity*, the annotator associates them with sheep, *an imaginary entity*, since they are similar in shape and color. Similarly, in the example of Fig. 1(c), the bright colors in the image evoke the Sun in the annotator's mind; the activity of the man in the image of Fig. 1(d) makes the annotator liken a tennis racket to a guitar.

We tend to generate imagination-driven descriptions for images that are capable of spontaneously stimulating our imagination to produce similar images in our minds. It is hard, possibly impossible, for us to give concise and understandable literal descriptions of such images. This is because of the limited expression power of literal language; humans can perceive many more colors and shapes than concepts or words we have in language to describe them [9]. In addition, forcibly generated literal descriptions for such images are usually unnatural and complicated, making them visually incomprehensible for humans. For instance, it is quite challenging to create literal descriptions for the example images in Fig. 1(b)(c) and (d). Instead, the annotators provide imagination-driven descriptions, which are vivid and easy to be visualized by readers/hearers.

In this paper, we study the generation and evaluation of such imagination-driven captions, which have been overlooked in previous studies. Although a lot of efforts have been made to provide various datasets [6, 8, 10, 11, 12] and methods [13, 14, 15, 16] for image captioning, there has been no prior study seriously considering imagination-driven captions. There are three main challenges to enable the generation of imagination-driven captions. The first challenge is the lack of datasets fit for the study. The second challenge is the absence of proper methods for generating such captions. The existing image captioning models are not designed to properly handle imagination-driven captions, as discussed later. The third challenge is the lack of the methods that can evaluate imagination-driven caption generation. Existing metrics for image captioning does not perform well, as we will show later.

Towards conquering the above challenges, in this paper, we first analyse four existing datasets, showing the statistics of imagination-driven captions in them. We then show the comprehensive categorization of imagination-driven captions and their usages. We also discuss the issues with existing image captioning models to generate such captions. By collecting imagination-driven captions from the four source datasets and synthesizing fake captions, we create a dataset, named *IdC-I* and *-II*. Using this dataset, we assess the accuracy of nine state-of-the-art evaluation metrics. Experimental results show that UMIC [17] shows high accuracy in evaluating imagination-driven captions of natural images. Meanwhile, for art images, all the considered metrics are not very effective.

Finally, we propose a novel image captioning method that can adequately handle imagination-driven caption generation. We design it to address a particular difficulty that existing models have. It is that they do not distinguish between visual and imaginary entities, leading to the inability to generate good imagination-driven captions and select the appropriate caption type fit for the input image. In the above example of Fig. 1(b), *sheep* in the imagination-driven caption does not appear as real entities in the image. Thus, the models need to be able to differentiate between real and imaginary sheep. Moreover, they must adequately judge which type of captions to generate for a given image; literal captions are sufficient for some images, and imagination-driven ones are better or necessary for others. To address these, our model generates literal and imagination-driven captions separately and selects the one that best fits the input image. For the latter, we build a scorer that scores the quality of the generated imagination-driven caption based on the content of the images. The scorer is built based on CLIP [18], which is a pre-trained vision-language model. Experimental results demonstrate the effectiveness of the proposed model over six existing methods developed for standard image-captioning.

In summary, the main contributions of this work are as follows:

- We analyze the existing image captioning datasets and present the comprehensive categorization of imagination-driven captions and their usages.
- We introduce a dataset for the generation and evaluation of imagination driven image captioning.

– Using this dataset, we examine nine state-of-the-art metrics of image captioning to measure the accuracy of imagination-driven caption generation.
– We propose a new image captioning model and experimentally examine its effectiveness.

2 Related Work

Image Captioning Datasets In the literature, a lot of efforts have been made to build image captioning datasets [6, 7, 8, 10, 11, 19]. Some of them are domain-generic datasets with images of various scenes and objects such as *MS COCO* [6], *Flickr30K* [8], and *CC12M* [19]. Because of the broad coverage of scenes and objects, these datasets are usually considered as standard benchmarks to build and evaluate image captioning methods [20].

Besides, there are some domain-specific datasets that are constructed for several specific tasks [1]. For instance, the *CUB* dataset consists of 117,880 captions of bird images [21]. Considering linguistic aspects, the authors in [22] focus on captions including negations. In [23], an attempt has been made to build a corpus of commonly used phrases that are repeated almost verbatim in captions of different images. To the best of our knowledge, there has been no previous research constructing datasets specific to imagination-driven caption generation.

Image Captioning Models Most existing image captioning models are based on an encoder-decoder paradigm. In this paradigm, an image encoder is used to project images to visual representations, which are then fed into a text generator to generate captions [1]. Many models use CNN as an image encoder [5,13,14,15]. However, CNN usually results in loss of granularity [1]. To address this problem, an attention mechanism over visual regions is exploited [16, 24, 25, 26]. Typically, Faster R-CNN [27] is used to extract bounding boxes of concrete objects, whose representations are then fed to a text generator to output captions. Recently, thanks to computational efficiency and scalability, transformer architectures based on self-attention mechanisms [28] are also adopted as image encoders [29, 30].

Regarding text generators, LSTM [31] has become a predominant architecture for a long time due to its ability to learn dependencies in captions. However, it faces issues about training speed and the ability to learn long-term dependencies [32]. Recent studies employ transformer architectures [28] as an alternative, since it can better learn long-term dependencies [24, 25, 33]. Besides, to enrich generated captions, some studies adopt graphs to detect spatial relationships of visual entities [3,25]. Existing image captioning studies mainly focus on detecting visual entities and their spatial relationships. So far, there has been no previous study that aims at generating imagination-driven captions, where forming imaginary entities is also of crucial importance.

3 Analyzing Imagination-driven Captions

We first analyze imagination-driven captions to answer the following three questions: 1) how frequently imagination-driven captions are used in existing image captioning datasets; 2) how many types exist, and 3) when humans (annotators) use such captions. For this purpose, we first examine imagination-driven captions in four image captioning datasets. We then classify imagination-driven captions in terms of their types and usages. We finally introduce datasets, named *IdC-I* and *IdC-II*, for the study of imagination-driven caption generation, which are the collection of extracted captions from the above datasets and synthesized fake captions to assess evaluation metrics of image captioning.

3.1 Analysis of Existing Datasets

To answer the above questions, we consider the following existing image captioning datasets: *MS COCO* [6], *Flickr30K* [8], *VizWiz* [11], and *ArtEmis* [7]. We adopt a filtering strategy to extract imagination-driven captions from them. *MS COCO* and *Flickr30K* are domain-generic datasets while *VizWiz* and *ArtEmis* are domain-specific datasets. *MS COCO* contains 995,684 captions from 164,062 images, which were gathered by searching for 80 object categories and different scene types on Flickr [34]. *Flickr30K* consists of 158,920 captions from 31,784 images of everyday activities and scenes. *VizWiz* is constructed to study image captioning for people who are blind. It includes 195,905 captions from 39,181 images taken by blind photographers in their daily lives. *ArtEmis* containing 454,684 captions from 80,031 art images is created to investigate affective human experiences evoked by artworks.

As in the examples in Fig. 1(b)(c) and (d), it can be seen that the annotators use the phrases of *look like, remind me,* and *as if* to associate visual entities (i.e., *cupcakes*, *colors*, and *tennis racket*) with imaginary entities (i.e., *sheep*, *Sun*, and *guitar*). Taking advantage of this feature, a list of 34 keywords as given in Table 1 is exploited to automatically extract imagination-driven captions from the source datasets. A caption will be considered imagination-driven if it includes at least one of these keywords.

Table 2 shows the statistics of imagination-driven captions extracted by the above procedure; *ratioImg* (*ratioCapt*) is defined as the ratio of the extracted images (captions) over the total number of images (captions) in each source dataset. It can be seen that imagination-driven captions account for only a small portion of captions (i.e., $< 1\%$) in *MS COCO*, *Flickr30K*, and *VizWiz*. There are multiple possible reasons. One is that these datasets mainly include natural images with typical objects/activities that do not trigger human imagination. In addition, the annotators of these datasets are requested to be as objective as possible to provide captions [10, 11]; in other words, they tend to avoid imagination-driven captions that are likely to be subjective. Thus, note that just because their percentages are small does not mean that they are unimportant.

Meanwhile, for *ArtEmis*, 70.86 percent of the included images are described by imagination-driven captions. Also, imagination-driven captions account for 22.08 percent of all the captions. This is because this dataset is created to investigate affective human experiences; the included images (i.e., artworks) are

Table 1: Keywords used to extract imagination-driven captions.

Keywords						
'looks like'	'look like'	'look as'	'looks as'	'looks likely'	'look likely'	'is likely'
'are likely'	'is like'	'are like'	'looks almost like'	'look almost like'	'shaped like'	'shapes like'
'shape like'	'is almost as'	'are almost as'	'seems to be'	'seem to be'	'seems like'	'thinks of'
'think of'	'as if'	'as though'	'seems as'	'seem as'	'seem like'	'calm like'
'feels like'	'feel like'	'resemble'	'resembling'	'reminds me'	'remind me'	

Table 2: Statistics of imagination-driven captions extracted from source datasets.

Source dataset	**MS COCO**	**Flickr30K**	**VizWiz**	**ArtEmis**
Number of extracted images	1489	577	1133	56,707
Number of extracted captions	1699	595	1160	100,393
ratioImg(%)	1.21	1.82	3.63	70.86
ratioCapt(%)	0.28	0.37	0.74	22.08

mostly abstract and have a tendency to evoke human emotion as well as human imagination.

3.2 Classifying Imagination-driven Captions and Their Usages

Through the analyses of the above datasets and others, we found that imagination driven captions can be categorized into two types, which we name object-based and action-based. We also found that there are three scenarios of their usage. We show their details below, along with the challenges for image captioning models to generate each type in each scenario.

Object-based Caption Type Object-based captions are created based on similarities of characteristics between visual and imaginary objects. Imaginary objects in these captions are either "common" or "proper". Figures 2(a), (b), and (c) show three examples of this caption type. In the first example, the clocks in the image evoke cats in the annotator's mind, *a common object/animal*. Interestingly, the clocks do not have identical shapes and colors to real cats; they are modified/deformed based on the designer's imagination and creativity. This raises the first challenge for image captioning models: how they can connect visual entities that are the products of human imagination and creativity with the right imaginary entities. On the other hand, the cake in Fig. 2(b) looks so similar to a real dog that they are indistinguishable even from humans without considering the context of the image. Thus, the second challenge is how models distinguish real objects and their "look-alike" objects, generating good captions. In the third example, the vanity style makes the annotator imagine a Victorian house, *a proper object*, which refers to a popular architectural revival style during the reign of Queen Victoria (1837-1901). To generate such captions, the model needs to retain special knowledge they may not be able to acquire through learning using generic image captioning datasets.

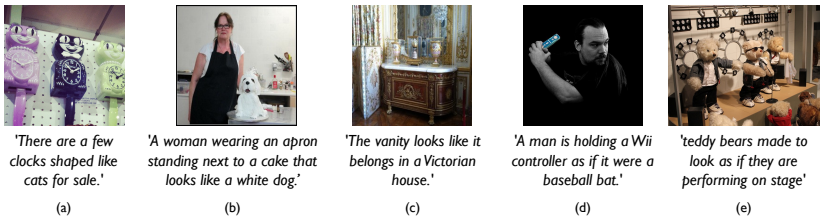

Fig. 2: Examples of imagination-driven caption types. (a),(b)&(c): Object-based captions; (d)&(e): Action-based captions. Examples are from *MS COCO* [6].

Fig. 3: Examples of imagination-driven description usages. *LC* and *IdC* denote literal and imagination-driven descriptions. Examples are from *MS COCO* [6].

Action-based Caption Type Unlike object-based captions, imaginary entities in action-based captions are triggered by the poses of visual humans/animals or the contexts of the images. Figures 2(d) and (e) show two examples of this caption type. When observing the image in Fig. 2(d), the man's pose makes the annotator think about a baseball bat. It is challenging for models to detect poses and link them to actions as humans do regardless of irrelevant objects in images. In the second example of Fig. 2(e), based on the context of the image (i.e., microphone, stage, audience, etc.), teddy bears are personified as humans performing on stage. It will be hard for existing captioning models to understand contexts to generate such personification captions.

Usage Scenarios There are three typical usage scenarios for imagination-driven descriptions. The first scenario is when objects/actions in an image remind us of imaginary entities due to their similarities in key characteristics. The second scenario is when annotators want to express their emotions when observing the images or the expressions of humans/animals in images. Finally, the third scenario is when annotators attempt to use their imagination to describe a visual entity that they cannot accurately recognize.

To illustrate these three usage scenarios, Fig. 3 shows four examples of the pairs of literal and imagination-driven descriptions. In the first example, to depict the color of the fire hydrant, the first annotator uses a literal adjective phrase of

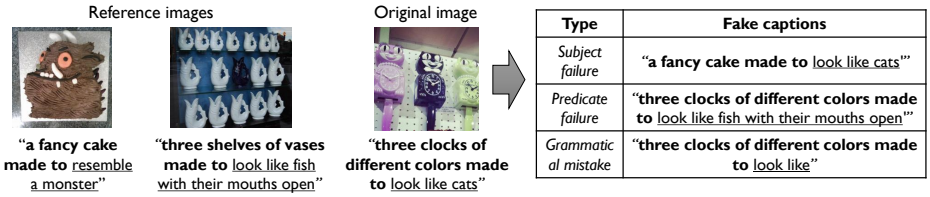

Fig. 4: Example of real and fake captions in *IdC-II*. The bold texts indicate *subjects* while the underlined texts indicate *predicates*.

red and white. Meanwhile, the second annotator utilizes the imagination-driven phrase of *look like a dalmatian dog*. Compared to the literal description, the imagination-driven description is more vivid and easier to be visualized by humans.

In the second example, unlike the literal description of *"A giraffe standing next to a fence a green field"*, which describes the action and position of the giraffe, the imagination-driven description, *"A giraffe who looks like he need friends"*, emphasizes the affective expression of the giraffe. Similarly, the feeling of the annotator when observing the image is conveyed through the imagination-driven description in the third example, *"Reminds me of my trip to Venice. Going under the bridge in a boat is so romantic."*. In the final example, while the literal description uses *something* to explain the unrecognizable object in the image, the imagination-driven caption associates the object to *"a piece of cheese or bread"*.

These examples illustrate that imagination-driven captions can be used in various scenarios to make descriptions more vivid and effectively express the feelings of annotators and the expressions of visual entities in images. However, imagination-driven captions are often unnecessary or inadequate for some types of images that do not evoke human imagination. How to select which to generate between literal and imagination-driven captions for individual images is challenging for existing models.

3.3 Dataset: *IdC-I* and *IdC-II*

In this section, we present a dataset that will be useful for studying imagination driven image captioning, named *IdC-I* and *-II*. *IdC-I* contains all the imagination driven captions extracted from the existing datasets as above, which we call "real" captions in what follows. *IdC-II* contains *IdC-I* and additionally "fake" captions we create by combining irrelevant visual and imaginary entities. Using *IdC-II*, we can assess existing metrics for image captioning; how accurately they can evaluate imagination-driven captions. This is done by checking if each metric yields higher scores with input images for their real captions than fake captions.

To create fake captions, we first split the sentence of each real caption into two parts: the set of words before keywords, called *subject*, and the set of the remaining words, called *predicate*. Then, we select the pairs of real captions satisfying: 1) they are of different images, and 2) the number of overlapping

words of the two subjects is highest. Next, through an analysis of failures of image captioning models, we exploit three typical error types to create fake captions, namely subject failures (SF), predicate failures (PF), and grammatical mistakes (GM). The first type refers to failures at detecting visual objects that are usually included in subjects. For the second type, predicates are not aligned with subjects considering image content. Related to grammatical mistakes, we generate incomplete captions by randomly excluding some ending words in the real captions. Figure 4 shows an example of fake captions corresponding to the three error types. A good metric should give the highest score to real captions and lower scores to fake captions.

4 A Method for Generating Both Types of Captions

4.1 Overview of the Proposed Method

As discussed earlier, existing image captioning models cannot generate imagination driven captions properly. This is because it requires higher-level skills than literal caption generation. For instance, it requires the ability to interpret a visual entity in two ways (i.e., literal and imaginary) and then link them to generate a meaningful description. An image captioning model may not be able to acquire such an ability, if not impossible, through pure learning using generic image-captioning datasets.

To cope with this difficulty, we build a model that internally generates literal and imagination-driven captions for the input image and then selects the one best fit for the image. Specifically, it has two text generators, a literal caption generator (LCGen) and an imagination-driven caption generator (IdCGen); each is designed to generate one caption type. Figure 5(b) depicts the model's architecture, which consists of four main modules: an image encoder, LCGen, IdCGen, and a selector. Given an image, the image encoder extracts visual representation $\mathbf{V}$ from the input image, which is then fed into LCGen and IdCGen to produce two captions. To select one of them, the selector scores the imagination-driven caption using a CLIP-based scorer, named *CScorer*. If the score is higher than a threshold θ, the model selects the imagination-driven caption. Otherwise, it selects the literal caption. The image encoder is designed based on Vision Transformer (ViT) [29] while the text generators are built based on the transformer decoder architecture [28]. We will explain the image encoder, the text generator, and the scorer in this order.

4.2 Image Encoder

Figure 5(a) shows the image encoder architecture. First, an image with the resolution of $H \times W$ is spatially divided into patches with a fixed resolution of $P \times P$. Consequently, there are totally $N = HW/P^2$ patches. Next, a linear projection is applied to generate the patch embeddings $\mathbf{X}^p$. The embeddings are then concatenated with a learnable embedding $\tilde{\mathbf{x}}_0$. To retain positional information, the concatenated embeddings are added to positional encodings $\mathbf{P}^{image}$.

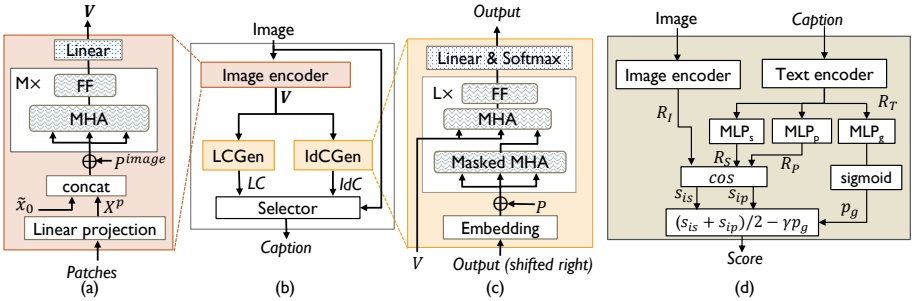

Fig. 5: (a) Image encoder architecture, (b) General architecture of the proposed model, (c) Text generator architecture, (d) CScorer architecture.

The result is then fed into a transformer encoder [28]. The transformer encoder is constructed from a stack of M identical layers, each contains two sub-layers of multi-head self-attention (MHA) and position-wise fully connected feed forward (FF) in sequence. The FF sub-layer is composed of two linear transformations with a GELU non-linearity in between. Note that each sub-layer is preceded by layer normalization and followed by a residual connection [35]. Layer normalization and a linear transformation are then employed to map the output of the transformer encoder to the visual representation $\mathbf{V}$.

4.3 Text Generator

Figure 5(c) shows the text generator architecture. Given the visual representation $\mathbf{V}$, the output is a sequence of T tokens one at a time. Similar to existing studies on text generation [36, 37], the proposed model does also work in an auto-regressive manner [24]. To retain positional information, the token embeddings are added to positional encodings $\mathbf{P}$. The result is then fed into a stack of L identical layers, each includes three alternating sub-layers of masked MHA, MHA, and FF. Each sub-layer is in between layer normalization and a residual connection [35]. To map the result to the token probability, we use layer normalization, a linear transformation, and a softmax transformation. The loop to generate tokens ends when either the number of the output tokens reaches the maximum length T or a special token $<eos>$ (end of sequence) is produced.

4.4 CLIP-based Scorer

Figure 5(d) illustrates the architecture of the scorer. By using the image encoder and the text encoder of CLIP, the image and the caption are projected to the corresponding representations of R_I and R_T. With the input of R_T, three multi-layer perceptrons (MLP) modules compute the grammatical penalty p_g, the subject representation R_S, and the predicate representation R_P. Each MLP module includes two linear transformations with a GELU non-linearity in between. The image-caption alignment score is then computed as the average of

the two cosine similarity values: $s_{is} = cos(R_I, R_S)$ and $s_{ip} = cos(R_I, R_P)$. The final score is calculated by $s = (s_{is} + s_{ip})/2 - \gamma p_g$.

Similar to [18], we train CScorer by using the multi-class N-pair loss [38]. However, because each image is paired with not only one real caption but also three fake captions, an asymmetric loss is adopted instead of a symmetric loss. The loss function is $\mathcal{L} = \mathcal{L}_{is}^{\mathcal{CE}}(s_{is}) + \alpha\mathcal{L}_{ip}^{\mathcal{CE}}(s_{ip}) + \beta\mathcal{L}_{s}^{\mathcal{CE}}(s) + \mu\mathcal{L}_{g}^{\mathcal{BCE}}(p_g)$, where $\mathcal{L}^{\mathcal{CE}}$ and $\mathcal{L}^{\mathcal{BCE}}$ denote cross entropy loss and binary cross entropy loss, respectively. Note that image-caption pairs are used to calculate $\mathcal{L}_{is}$, $\mathcal{L}_{ip}$, and $\mathcal{L}_s$, while $\mathcal{L}_g$ requires only captions.

5 Experiments

In this section, we show the results of two experiments. The first one assesses the accuracy of nine existing metrics in evaluating imagination-driven image captions. The second experiment evaluates the effectiveness of the image captioning model presented in Sec. 4. To indicate the source dataset of captions, we will use names with the prefix of the source dataset. For instance, *MS COCO-IdC-I* denotes the set of captions in *IdC-I* that originate from the *MS COCO* dataset.

5.1 Evaluation of Metrics

Experimental Settings We use *IdC-II* to evaluate the accuracy of nine existing metrics, namely TIGEr [39], VIFIDEL$_{\text{no-refs}}$ (nrVIFIDEL) [40], VIFIDEL [40], BERTScore (BERTS) [41], ViLBERTScore (ViLBERTS) [42], UMIC [17], SMURF [43], CLIPScore (CLIPS) [44], and RefCLIPScore (RefCLIPS) [44]. Among these metrics, nrVIFIDEL, UMIC, and CLIPS are unreferenced metrics [17], which do not require human-generated annotations, whereas the others are referenced metrics.

As mentioned in Section 3.3, *IdC-II* includes caption tuples, each consisting of one real caption and three fake captions. For each tuple, a good metric should give the highest score to the only real caption and lower scores for the fake captions. Based on this idea, we regard the scoring of a tuple as accurate if and only if the real caption has the highest score. We use the ratio of the number of accurately scored caption tuples over the total number of caption tuples as the evaluation measure of the metrics.

For the three sets of *MS COCO-IdC-II*, *VizWiz-IdC-II*, and *Flickr30K-IdC-II*, we evaluate only the three unreferenced metrics, due to the deficient numbers of human-generated annotations per image. For *ArtEmis-IdC-II*, to enable the evaluation of all the metrics, we use only 2497 images that have at least four human-generated annotations per image.

Results Table 3 shows the accuracy of the metrics on *IdC-II*. For *MS COCO-IdC-II*, *VizWiz-IdC-II*, and *Flickr30K-IdC-II*, which include natural images, UMIC generally achieves the highest accuracy. In terms of average accuracy,

Table 3: Accuracy of the metrics on *IdC-II*.

		MS COCO-IdC-II			*VizWiz-IdC-II*			*Flickr30K-IdC-II*		
Metric		**nrVIFIDEL**	**UMIC**	**CLIPS**	**nrVIFIDEL**	**UMIC**	**CLIPS**	**nrVIFIDEL**	**UMIC**	**CLIPS**
Error type	*SF*	0.87	**0.98**	0.97	0.65	0.86	**0.87**	0.63	**0.97**	0.97
	PF	0.67	**0.91**	0.89	0.65	0.81	**0.88**	0.61	0.89	**0.89**
	GM	0.47	**0.83**	0.80	0.35	**0.69**	0.68	0.37	**0.78**	0.64
Average		0.67	**0.91**	0.88	0.55	0.78	**0.81**	0.54	**0.88**	0.83

	ArtEmis-IdC-II								
Metric	**TIGEr**	**nrVIFIDEL**	**VIFIDEL**	**BERTS**	**ViLBERTS**	**UMIC**	**SMURF**	**CLIPS**	**refCLIPS**
Error type *SF*	0.68	0.60	0.71	0.78	0.77	0.72	0.55	0.78	**0.78**
Error type *PF*	0.66	0.59	0.63	0.70	0.62	0.70	0.64	0.77	**0.77**
Error type *GM*	0.54	0.33	0.26	**0.66**	0.28	0.60	0.57	0.55	0.55
Average	0.63	0.51	0.54	**0.71**	0.56	0.68	0.59	0.70	0.70

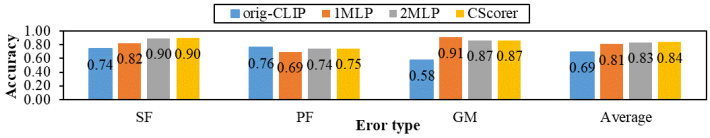

Fig. 6: Accuracy of CScorer and baselines on *ArtEmis-IdC-II* test set.

the results of UMIC and CLIPS are rather high (i.e., ≥ 0.78) while that of nfV-IFIDEL is quite low (i.e., ≤ 0.67). When comparing the three error types, the accuracies are highest for the subject failure type, followed by the predicate failure type. With the accuracies higher than 0.81, UMIC and CLIPS can be used to evaluate the degree of image-text alignment of imagination-driven captions. Meanwhile, for grammatical mistakes, the accuracies range from 0.64 to 0.83, suggesting that there is still room for improvement in evaluating this error type.

Regarding *ArtEmis-IdC-II*, which includes art images, all the nine considered metrics are not very effective. Their average accuracies are from 0.51 to 0.71. In particular, all of them show poor performance for grammatical mistakes (i.e., ≤ 0.66). These results imply that it is essential to build dedicated metrics for evaluating imagination-driven captions for art images.

5.2 Evaluation of Image Captioning Models

Experimental Settings In this experiment, we use the *ArtEmis* dataset, which has a sufficient number of imagination-driven captions for training and testing. *ArtEmis* consists of *ArtEmis-IdC-I* and the set of literal captions named *ArtEmis-LC*. Regarding the data splitting, *ArtEmis* is divided into the training, validation, and test sets, including 68,028 images, 6000 images, and 5497 images. To make a comprehensive evaluation of the models, we additionally consider two subsets of the test set: 1) a set of all the literal captions with 4019 images (*ArtEmis-LC-TS*) and 2) a set of all the imagination-driven captions from 2497 images (*ArtEmis-IdC-I-TS*). By using the same data split, *ArtEmis-IdC-II*, which is used to train CScorer, is also divided into the training, validation, and test sets.

Table 4: Performances of the models. The bold number indicate the highest performance.

Models	*Whole test set*						*ArtEmis-LC-TS*						*ArtEmis-IdC-I-TS*					
	B1	**B2**	**B3**	**B4**	**M**	**R**	**B1**	**B2**	**B3**	**B4**	**M**	**R**	**B1**	**B2**	**B3**	**B4**	**M**	**R**
NN	40.8	17.6	7.6	3.4	10.9	22.8	40.3	16.8	6.9	2.9	10.7	22.7	38.6	16.2	7.0	3.2	10.6	21.3
SAT	57.5	34.4	20.5	12.5	15.2	31.5	58.4	32.8	17.3	8.6	14.7	31.5	54.2	31.9	19.0	12.1	14.9	30.1
$\mathcal{M}^2$	57.1	33.4	19.6	11.7	14.8	31.3	57.1	31.9	17.3	8.9	14.5	31.6	51.0	28.7	16.2	9.4	13.6	28.0
CLIPCap	47.8	27.2	15.4	9.3	**16.0**	25.6	47.8	24.8	11.7	5.6	15.0	24.3	44.1	24.9	14.6	9.2	14.7	23.7
Oscar	44.1	24.0	13.0	7.2	15.5	28.2	54.0	29.0	14.9	7.4	14.5	30.3	50.6	28.6	16.7	10.3	14.5	28.8
OFA	59.9	35.1	18.8	9.9	15.8	32.3	59.4	34.6	18.1	9.3	**15.7**	32.7	53.4	28.8	14.4	6.8	13.4	27.6
1GEN	60.1	37.6	**22.8**	**14.1**	15.4	32.6	**61.2**	36.3	**19.9**	**10.2**	15.3	32.9	53.6	32.0	18.7	11.5	14.2	29.6
2GEN	**61.8**	**38.1**	22.5	13.6	15.9	**33.2**	61.2	**36.6**	19.9	10.1	15.3	**33.4**	**62.6**	**43.0**	**29.1**	**19.9**	**19.1**	**37.6**

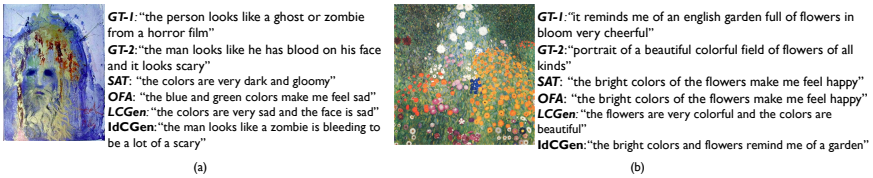

Fig. 7: Examples of human-generated descriptions (*GT-1* and *GT-2*) and captions generated by SAT, OFA, LCGen and IdCGen.

Regarding the image encoder in the proposed model, we use the pre-trained image encoder of CLIP with the ViT-B/16 backbone [18]. For LCGen and IdCGen, the number of layers and heads are set to 8. The maximum caption length is set to $K = 65$. The loss is calculated by the average loss over LCGen and IdCGen using cross-entropy loss functions. For CScorer, we use the pre-trained CLIP with the ResNet-50x16 backbone [18]; θ and γ are set to 0.5 and 0.2. The parameters of the loss function are set to $\alpha = \beta = \mu = 1$. The optimizer is Adam with (0.9,0.999) [45]. To assess the effectiveness of using the two text generators (called *2GEN*), we also evaluate the case of using only one text generator for both the caption types, called *1GEN*.

Our model is compared with six reference models of Nearest-Neighbor (NN) [7], Show-Attend-Tell (SAT) [5], Meshed-Memory transformer ($\mathcal{M}^2$) [26], Oscar [46], CLIPCap [47], and OFA [48]. Among the considered models, Oscar, CLIPCap, OFA, and our model are based on large-scale vision-language pre-training (VLP). For the remaining models, they are based on basic backbones: ResNet-34 pre-trained on ImageNet for NN and SAT, Faster R-CNN [27] pre-trained on Visual Genome for $\mathcal{M}^2$. For a fair comparison, we use beam search for the reference models with the beam size of 2 since our model generates two captions simultaneously from LCGen and IdCGen. We use six commonly used evaluation metrics of BLEU 1-4 (*B1-B4*) [49], ROUGE-L (*R*) [50], and *METEOR* (*M*) [51].

Results By using *ArtEmis-IdC-II*, we evaluate the accuracy of CScorer and three baselines, namely 2MLP, 1MLP, and orig-CLIP. These baselines excludes one, two, or three MLP modules, respectively, from CScorer. Figure 6 shows

the accuracy of CScorer and baselines. It can be seen that CScorer is the most effective scorer with high and stable accuracies (i.e., ≥ 0.75). These results also demonstrate the effectiveness of adding MLP modules in increasing the accuracy of CScorer, compared to orig-CLIP.

Table 4 shows the obtained results of the models. Among the reference models, SAT and OFA usually perform best. Comparing all the models, we see that 1GEN and 2GEN generally achieve the highest performance for the whole test set and *ArtEmis-LC-TS*. Also, the gap between them is small. However, for *ArtEmis-IdC-TS*, 2GEN outperforms all the other models with a significant margin, suggesting that 2GEN can effectively generate literal and imagination-driven captions. This result also implies the advantage of separating the generation of these two caption types.

Figure 7 shows two examples of captions generated by SAT, OFA, and LCGen/IdCGen of our model[3]. Compared with SAT and OFA, our model generally produces captions closer to human-generated descriptions. Particularly, our model accurately detects visual entities and associates them with imaginary entities similar to the annotators. As an example in Fig. 7(a), LCGen and IdCGen detect *face* and *man* as visual entities; *man* is linked with *a zombie is bleeding* in the imagination-driven caption generated by IdCGen. SAT and OFA generate only literal captions about the image's colors; no visual or imaginary object is included in these captions. In Fig. 7(b), although all the captions include *flowers* as a visual entity, only IdCGen successfully associates *flowers* with *a garden* as in the first annotator's description.

In summary, it can be observed that the proposed model is more effective than the existing models in generating both literal and imagination-driven captions. Also, it is suggested to learn these two caption types separately.

6 Summary and Conclusion

In this paper, we have shed light on the previously overlooked problem of generating imagination-driven captions. Specifically, we have analyzed existing datasets, classified imagination-driven captions, and discussed their usage. In addition, we have introduced the dataset for generating and evaluating imagination-driven image captioning methods. By using this dataset, we have assessed the nine existing evaluation metrics. Also, we have proposed a model capable of generating literal and imagination-driven captions. By separately learning the two caption types, our model is experimentally found to be more effective than the six existing models, especially for generating imagination-driven captions. We hope this study will be the groundwork for future studies on imagination-driven captions.

Acknowledgements This work was supported by JST [Moonshot Research and Development], Grant Number [JPMJMS2032] and by JSPS KAKENHI Grant Number 20H05952 and 19H01110.

[3] See the captions generated by the other models and more examples in the supplementary material.

References

1. Hossain, M.Z., Sohel, F., Shiratuddin, M.F., Laga, H.: A comprehensive survey of deep learning for image captioning. ACM Computing Surveys **51** (2019) 1–36
2. Nguyen, K., Tripathi, S., Du, B., Guha, T., Nguyen, T.Q.: In defense of scene graphs for image captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 1407–1416
3. Yao, T., Pan, Y., Li, Y., Mei, T.: Exploring visual relationship for image captioning. In: Proceedings of the European Conference on Computer Vision. (2018) 684–699
4. Gu, J., Joty, S., Cai, J., Zhao, H., Yang, X., Wang, G.: Unpaired image captioning via scene graph alignments. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 10323–10332
5. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning. (2015) 2048–2057
6. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
7. Achlioptas, P., Ovsjanikov, M., Haydarov, K., Elhoseiny, M., Guibas, L.J.: Artemis: Affective language for visual art. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 11569–11579
8. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: 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
9. Carston, R.: Figurative language, mental imagery, and pragmatics. Metaphor and Symbol **33** (2018) 198–217
10. Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research **47** (2013) 853–899
11. Gurari, D., Zhao, Y., Zhang, M., Bhattacharya, N.: Captioning images taken by people who are blind. In: European Conference on Computer Vision. (2020) 417–434
12. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL). (2018)
13. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 3156–3164
14. Chen, F., Ji, R., Sun, X., Wu, Y., Su, J.: Groupcap: Group-based image captioning with structured relevance and diversity constraints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 1345–1353
15. Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-rnn). In: The International Conference on Learning Representations (ICLR). (2015)
16. Liao, W., Rosenhahn, B., Shuai, L., Ying Yang, M.: Natural language guided visual relationship detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. (2019)
17. Lee, H., Yoon, S., Dernoncourt, F., Bui, T., Jung, K.: 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). (2021) 220–226
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. (2021) 8748–8763
19. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 3558–3568
20. Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., Cucchiara, R.: From show to tell: A survey on deep learning-based image captioning. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
21. Reed, S., Akata, Z., Lee, H., Schiele, B.: Learning deep representations of fine-grained visual descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 49–58
22. van Miltenburg, C., Vallejo, R.M., Elliott, D.: Pragmatic factors in image description: the case of negations. In: Proceedings of the Workshop on Vision and Language. (2016) 54–59
23. Chen, J., Kuznetsova, P., Warren, D., Choi, Y.: Déja image-captions: A corpus of expressive descriptions in repetition. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. (2015) 504–514
24. Xu, G., Niu, S., Tan, M., Luo, Y., Du, Q., Wu, Q.: Towards accurate text-based image captioning with content diversity exploration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 12637–12646
25. Chen, H., Wang, Y., Yang, X., Li, J.: Captioning transformer with scene graph guiding. In: IEEE International Conference on Image Processing (ICIP). (2021) 2538–2542
26. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 10578–10587
27. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 1440–1448
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)
29. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: The International Conference on Learning Representations (ICLR). (2021)
30. Shen, S., Li, L.H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.W., Yao, Z., Keutzer, K.: How much can clip benefit vision-and-language tasks? arXiv preprint arXiv:2107.06383 (2021)
31. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9** (1997) 1735–1780
32. Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., Cucchiara, R.: From show to tell: A survey on deep learning-based image captioning. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)

33. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. ACM Computing Surveys (2021)
34. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision. (2014) 740–755
35. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 770–778
36. Aneja, J., Deshpande, A., Schwing, A.G.: Convolutional image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 5561–5570
37. Li, Z., Tran, Q., Mai, L., Lin, Z., Yuille, A.L.: Context-aware group captioning via self-attention and contrastive features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 3440–3450
38. Sohn, K.: Improved deep metric learning with multi-class n-pair loss objective. Advances in neural information processing systems **29** (2016)
39. Jiang, M., Huang, Q., Zhang, L., Wang, X., Zhang, P., Gan, Z., Diesner, J., Gao, J.: 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). (2019) 2141–2152
40. Madhyastha, P.S., Wang, J., Specia, L.: Vifidel: Evaluating the visual fidelity of image descriptions. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. (2019) 6539–6550
41. Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. In: International Conference on Learning Representations. (2020)
42. Lee, H., Yoon, S., Dernoncourt, F., Kim, D.S., Bui, T., Jung, K.: Vilbertscore: Evaluating image caption using vision-and-language bert. In: Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems. (2020) 34–39
43. Feinglass, J., Yang, Y.: 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). (2021) 2250–2260
44. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. (2021) 7514–7528
45. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
46. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: European Conference on Computer Vision, Springer (2020) 121–137
47. Mokady, R., Hertz, A., Bermano, A.H.: Clipcap: Clip prefix for image captioning. arXiv preprint arXiv:2111.09734 (2021)
48. Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., Yang, H.: Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: The International Conference on Machine Learning (ICML). (2022)

49. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. (2002) 311–318
50. ROUGE, L.C.: A package for automatic evaluation of summaries. In: Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL), Workshop on Text Summarization, Spain (2004)
51. Denkowski, M., Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the Workshop on Statistical Machine Translation. (2014) 376–380