

# CLID: Controlled-Length Image Descriptions with Limited Data

Elad Hirsch      Ayellet Tal  
 Technion – Israel Institute of Technology  
 {eladhirsh@campus, ayellet@ee}.technion.ac.il

## Abstract

*Controllable image captioning models generate human-like image descriptions, enabling some kind of control over the generated captions. This paper focuses on controlling the caption length, i.e. a short and concise description or a long and detailed one. Since existing image captioning datasets contain mostly short captions, generating long captions is challenging. To address the shortage of long training examples, we propose to enrich the dataset with varying-length self-generated captions. These, however, might be of varying quality and are thus unsuitable for conventional training. We introduce a novel training strategy that selects the data points to be used at different times during the training. Our method dramatically improves the length-control abilities, while exhibiting SotA performance in terms of caption quality. Our approach is general and is shown to be applicable also to paragraph generation. Our code is publicly available<sup>1</sup>.*

## 1. Introduction

*Image Captioning* refers to the task of generating human-like image descriptions [58]. It relies on supervision, utilizing large datasets of image and text pairs [10, 25, 68]. *Controllable Image Captioning (CIC)* aims at generating a caption while satisfying a constraint or a user request. The constraint may relate to the content [12, 74], the style [42], the structure [7], or the length [14] of the caption.

Our work focuses on controlling the description length and consequently, the amount of detail in it. This may suit different people/applications at different scenarios. For instance, visually-impaired observers may prefer a short and concise description when in a hurry and a long and detailed one at other times. Other applications, such as text-based image retrieval, may also benefit from this flexibility.

Captioning training datasets have length limitations, as human-annotated captions tend to be short, concise and omit some visual information [18]. For example, in MS-

COCO [10] 95% of the training captions contain less than 15 words. This limits the length of the generated captions, making the generation of long captions a challenge. The straightforward solution of collecting image descriptions of varying lengths is infeasible. This paper shows how to address this shortage of data.

We propose a novel approach for generating varying-length image descriptions with inadequate training data. Our approach is based on solving two sub-problems. At first, we automatically generate long “synthetic” captions, termed *self-generated* captions. While this procedure addresses the lack of long training captions, the generated captions might be of low quality (*domain gap*). This leads to the second challenge of how to use this low-quality dataset during training. We present a novel training method that makes use of the trusted dataset of short captions jointly with the large varying-quality caption dataset, in an informed manner.

To address the first challenge, we propose a method that parses scene graphs and generates varying-length sentences. Scene graphs contain the essential descriptive components: objects, objects’ attributes, relations, and activities [18, 29]. Image saliency [27] is utilized for generating reasonable sentences. We end up with a large dataset of self-generated varying-length & varying-quality descriptions.

The second challenge is how to use the *extended dataset*, which is composed of the trusted short captions and the self-generated captions, during training. Since this dataset might contain inaccuracies (that exist in the scene graphs), repetitions, and different linguistic styles, we cannot consider both caption types similarly. We introduce a training procedure in which the use of the self-generated data is guided by the (small and short-caption) trusted dataset. In particular, our method progressively filters out low-quality captions during training. It assures that even though the very long captions might be eliminated early on and the remaining high-quality captions are mostly short, the model remembers enough of the long captions, to be able to generate ones. For this strategy to work, we should be able to measure caption quality, a topic that is addressed as well.

We demonstrate the benefits of our approach on the MS-

<sup>1</sup><https://github.com/eladhi/CLID>



LaBERT [14]	Level 1	a young man plays with a videorest.
	Level 3	a young man sitting in a living room with a wii remote and a white rug on his head.
	Level 6	a boy sitting in a chair next to a table with a towel on it and a surfboard on the floor next to a book shelf with a book case on the table and a surfboard on top of the table and a plant on the wall.
CLID (ours)	Level 1	a young boy is playing a video game.
	Level 3	a boy is playing a video game in a living room with a planted plant in the background.
	Level 6	a boy in a black jacket with dark brown hair is holding a wii remote in a room area with a stack of shelves and a chair with a blue table and a bookcase next to a bed and a cord on to the side of the bed.
(a) Length		(b) Generated Caption

Figure 1. **Length-controlled image captioning.** People describe a given image briefly or in length. Most previous works generate short captions, which are prevalent in existing datasets. We propose a method that generates captions of sought-after lengths. Our method generates long captions, which hardly exist in training datasets, and achieves comparable results to SoTA methods for short captions.

COCO Caption dataset [10], in which the mean caption length is 10.47 words. For long captions, which barely appear in the dataset, our work highly improves the control precision (by up to 27%), while maintaining the caption quality, as demonstrated in Figure 1. It even enables generating captions of lengths that do not appear in the dataset at all. The short captions contain only the essence of the image, whereas the long captions reveal many more details.

Our method is general and may be applied also to the task of *descriptive paragraph generation* from images [33], which is a related, yet distinct task. While both tasks generate a coherent natural-language image description, the linguistic structure and the amount of details differ. Captioning refers to a single descriptive sentence, whereas paragraphs consist of multiple sequential sentences. The existing paragraph dataset [33] is small and contains mostly long descriptions. In Section 4 we show that despite the very small dataset, our method manages to generate length-control results for this task as well.

Hence, our contributions are as follows:

1. We introduce a novel approach for length-controllable image captioning. It manages to generate long, even out-of-distribution, descriptions.
2. Since our approach is general and unified, it can be used for paragraph generation. This is the first work to introduce length-control abilities for this task.
3. We present a method for generating varying-length diverse-quality captions from scene graphs, without the need of a ground-truth dataset.
4. We propose a training procedure that learns both from high-quality data and from low-quality data, such that the essence of the low-quality data is not forgotten.

## 2. Related work

**Image captioning.** Image captioning is a core task in scene understanding. The common approach in recent years combines a visual encoding model, which extracts visual features, with a language model that learns to generate text

from both the visual features and the input text. The visual encoding model extracts global image features [17, 58], patch features [62, 65] or region features [3, 50]. Other image representation structures contain the visual features jointly with additional relevant information, such as scene graphs [46, 67]. The language models also vary, where RNNs [39, 58] and transformers [13, 24, 71] are common choices.

**Controllable image captioning (CIC).** This task takes captioning to the next level by adding constraints, which are usually user-defined. Such constraints may relate to caption linguistic style [1, 9], content [8, 12, 32, 74] or structure [7, 15]. Our work focuses on length control, also addressed in [14] and partially in [31]. In [14] an interesting architecture, *LaBERT*, is proposed for the problem. This is a transformer-based captioning model. Its input consists of visual and word embeddings, where the word embedding also represents the caption length. The length is measured in *tokens* (sub-words), which appear in the BERT vocabulary [16]. This model manages to control the length of captions up to 30 tokens well, with a precision rate of above 90%. Such captions are 99.8% of the training dataset. For longer lengths, which are the remaining 0.2%, the precision drops by 10-43%. The work of [31] aims to control both the image caption length and imageability (i.e., the clarity of the mental image). In the context of length, they use [14]. Our method addresses the limitation of long caption generation.

**Descriptive image paragraphs.** Paragraph generation aims at describing an image by a sequence of sentences. Common works use hierarchical RNNs to guide the sentence topics (high level) and word sequence (low level) [6, 33, 41, 59, 69]. Other solutions further guide the training of RNN models with reinforcement learning [40], scene graph hierarchy information [66] and adversarial training [36]. Despite the resemblance to single-sentence captioning, the tasks are considered distinct and thus the training datasets differ. Length-controllable paragraph generation has not yet been addressed.

**Scene graphs.** A scene graph represents the content of an image as a graph [29]. It is defined as a tuple  $G = (O, E)$ , where the set of vertices  $O$  represents the image objects and the set of edges  $E$  represents relationships between objects. Each object contains the object type and relevant attributes; each edge contains the relationship type. Scene graphs are used in numerous applications in computer vision, such as image retrieval [29, 49, 60], image captioning [19, 21, 67], VQA [20, 52] and image synthesis [28, 43, 54]. A widespread dataset of scene graphs is *Visual Genome* (*VG*) [34], which contains scene graphs of 108K images. Furthermore, there are various algorithms for scene graph generation (SGG), such as [22, 53, 64, 70].

**Small high-quality datasets and large low-quality datasets.** Having access to a small trusted dataset, as well as to a large untrusted dataset, is a typical scenario in classification with label noise [23, 35, 57, 63]. Another domain is *neural machine translation (NMT)*, in which the trusted dataset contains expert translations. The lack of data in this domain is treated by data mining or other automatic methods that provide new data in varying quality. This data is considered noisy, therefore used for training jointly with the smaller trusted dataset [4, 44, 61, 72].

### 3. Method

Given an image and a desired length level, our goal is to generate a description that satisfies the length constraint. Existing captioning models are trained on the same datasets, thus they share the same length statistics. We aim to generate descriptions of varying lengths, even when there are very few examples of a certain length. Towards this end, in Section 3 we propose a novel approach that handles the scarcity of long-captions. We assume that there exists some base model that is capable of generating short captions, for which there is ample data. We show that our approach manages to preserve this model’s performance for short captions, while dramatically improving that performance for long ones. In Section 4 we also show that our approach is as beneficial for paragraph generation.

Our approach consists of two key ideas, which lead to a two-phase method (Fig. 2). First, to solve the shortage of long captions, we enrich the dataset with varying-length captions, by utilizing a different image representation, a scene graph. Second, given a *self-generated* dataset, which contains captions of various lengths and varying quality, our novel training procedure benefits from the varying lengths and is barely affected by the varying qualities.

To realize the first phase, any length-aware base model can be utilized. We use *LaBERT* [14], since it is currently the only model that controls the caption length. This transformer-based model’s inputs are visual and word embeddings, where the word embeddings also represent the caption length. This phase is described in Section 3.1.

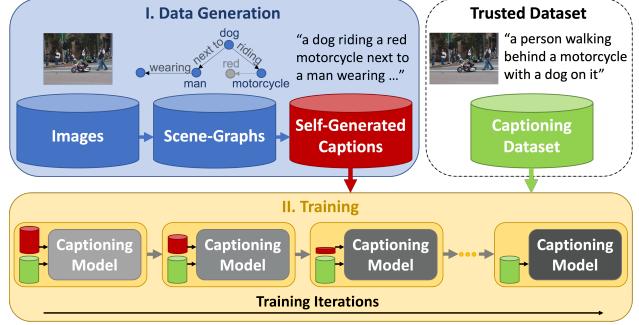


Figure 2. **Outline.** (I) To overcome the shortage in long captions in trusted datasets (green), a new dataset is self-generated (red) using scene graphs, creating an extended dataset. (II) During training, the low-quality data is gradually filtered out, while remembering the information learned from it. This improves length control and preserves captioning quality.

For the second phase, we propose a training procedure that makes use of the *extended dataset*, which consists of the mixed-quality *self-generated* dataset and the *trusted* dataset. Our strategy gradually tunes the model, to account for the diverse quality of the data, in a manner in which late iterations are exposed mostly to high-quality data, while not forgetting valuable information learned from the low-quality data. This phase is described in Section 3.2.

#### 3.1. Self-generating varying-length captions

Our goal is to generate a large dataset of image captions of various lengths, where we are willing to compromise on the quality of the captions. We hereby describe an algorithm that given an image, will generate such image captions. It is designed to satisfy three requirements: (1) The length of the captions should vary. (2) Each caption should include the important objects & relations. (3) The linguistic style should be as natural as possible.

At the base of the algorithm is the use of an image representation comprising the image’s objects, their attributes, their importance, and the relationships between them. This representation is parsed and captions of different lengths are generated, so as to take object importance into account.

**Extraction of image information.** Scene graphs allow us to extract the above information, with the exception of object importance. Briefly, in this directed graph, the vertices are objects and the edges are relationships. The objects may contain additional descriptive attributes, such as adjectives. Scene graphs are available for some datasets [34] or may be automatically generated [22, 53, 64, 70].

The only essential information that scene graphs do not hold is objects’ importance. This information is necessary, as captions of all lengths describe the salient objects. To gain this information, we compute an image saliency map, utilizing the method of [27], which is accurate and easy to run. It assigns each pixel a saliency score, which corre-

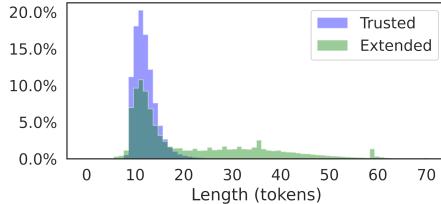


Figure 3. **Length of captioning datasets.** The average caption length in the trusted (MS-COCO) dataset is 11.95 tokens with standard deviation of 2.58, whereas in our extended dataset these are 21.3 and 13.56, respectively. (Overlaps cause the third color.)

sponds to human fixation. The object saliency is computed as the normalized sum of the pixel saliency scores of the object’s bounding box. In our case, the bounding box is provided by the scene graph. We create a vector whose entries are saliency-based weights of the image objects.

**Varying-length caption generation.** This is the core of our method. The idea is to explore the scene graph in a *Depth First Search (DFS)*-like manner, with a few modifications, which account for caption diversity and length diversity.

The algorithm proceeds as follows. The first object is chosen among the nodes of the scene graph, according to the distribution of the saliency vector. The more salient an object is, the more likely it is to be chosen. Once the initial node (object) is selected, DFS is applied with three differences. First, we do not necessarily visit all the children of each node. Instead, we sample up to  $k$  children, based on their saliency distribution (after re-normalization). Furthermore, we allow re-visiting nodes, since captioning “re-visits” objects, by mentioning them multiple times. For example, the object “boy” in the description “A boy is feeding a dog and a girl is smiling at the boy”, is revisited, where the second time is preceded by “the”. While revisiting nodes is allowed, loops are prohibited, in order to prevent (infinite) repetitions. This is done by avoiding propagation through already-visited vertices. Finally, if the exploration reaches an end, but less than  $T_{sal}\%$  of the sum of the image saliency map was explored, we “jump” to a new unexplored node, selected again according to the saliency vector. This means that we do not necessarily visit the whole graph.

During this traversal, a caption is generated, creating a noun for every node (object), a verb for every edge (relationship) and adjectives for the objects (node attributes). As a common connection of phrases in captioning, we add an “and” when there is a jump in the DFS (or a dot in the case of paragraphs). To increase the diversity of the captions and their lengths we add two random procedures: First, we cut out a random number of the last visited objects. Second, for every object in the generated caption we add a random number of attributes (up to a limit  $n_a$ ). The algorithm is summarized as pseudo-code in the supplemental materials.

**Linguistic style.** The generation algorithm described above

successfully depicts the objects, their order, and the relationship between them. However, it might generate unnatural linguistic style. A notable example is *grouping*: multiple objects (e.g. “two kids” or “animals”) appear in scene graphs as distinct nodes and are thus described as separate objects. To improve the grammar and the style, we use an off-the-shelf language paraphraser [26], which preserves the original meaning and produces a fluent and correct language. The paraphraser itself has no access to the image. Note that the paraphraser is applied only here and not on the output of the model described in Section 3.2.

**Results.** Fig. 3 compares the datasets. The captions in the original *trusted* dataset are relatively short, where 98% of the captions have less than 20 tokens. Differently, in our *extended* dataset only 53% are shorter than 20 tokens and the number of long captions is no longer negligible. We set  $T_{sal} = 80\%$ ,  $k = 2$  and  $n_a = 4$ , which suffice to explore both the essence of the image and many details.

### 3.2. Training with data selection

We are given a small high-quality (*trusted*) dataset and a large *self-generated* dataset. Our key assumption is that although the self-generated dataset is of low-quality, it does contain beneficial information for training. However, it should be utilized thoughtfully, in order not to be harmful.

A similar challenge is addressed in other domains, such as training classifiers with noisy labels [38, 63, 73] or using synthetic data [11, 45, 51]. The most related domain to ours is machine translation (NMT), where expert translations are rare, while noisy translations can be automatically generated. Our work, which is the first to explore mixed-quality datasets for image description, is inspired by [61].

We propose a training strategy that gradually tunes the model, to account for the diverse quality of the data. The core idea is to expose the model to the diverse data early on, and to expose it mostly to high-quality data in late stages of the training. This is so since the low-quality data provides a mass of fundamental information (objects, attributes, length control signals) that is essential for learning. The high-quality data is important later, since it fine-tunes the model on more relevant domain data. The benefit of the entire data is supported by our experiments in Section 4.

To gradually filter out data, we shall find a measure that rates the data according to its quality. Intuitively, the quality of a data point can be determined by its domain relevance, measured by the distance to in-domain data. Similarly to [61], this is done by the ratio between the probability to appear in the trusted training dataset to the probability to appear in the extended (trusted & self-generated) dataset. The higher the ratio, the more likely the data point to contain high-quality information to learn from.

Formally, assume for now that a captioning model, parameterized by  $\theta$ , can output the probability  $p(y_j|x_j; \theta)$  of

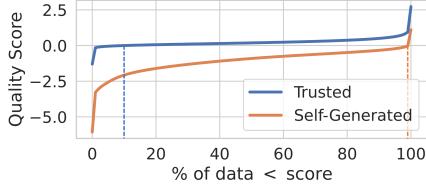


Figure 4. **Data separation.** Quality scores, computed by Eq. 1, manage to separate the trusted data from the self-generated data. The trusted data points have mostly (90%) positive scores (dashed blue line), while the self-generated data points have mostly (99%) negative scores (dashed orange line).

a caption  $y_j$  to match an image  $x_j$  (this assumption will be later removed). Given two models:  $M_\theta$ , which is trained on the trusted dataset  $D$ , and  $M_{\tilde{\theta}}$ , which is trained on the extended dataset  $\tilde{D}$ , we approximate the quality of each data point  $j$  in the self-generated dataset by:

$$u_j = \text{quality}(x_j, y_j; \theta, \tilde{\theta}) = \log \left( \frac{p(y_j|x_j; \theta)}{p(y_j|x_j; \tilde{\theta})} \right) = \log p(y_j|x_j; \theta) - \log p(y_j|x_j; \tilde{\theta}). \quad (1)$$

By Eq. 1, a positive value means that a data point is more likely according to the model trained on the trusted data than that trained on the extended data, while a negative value means the opposite. Therefore, this score rates the data points according to their likelihood of being in-domain.

Fig. 4 shows that our approach successfully separates most of the trusted data from the self-generated data on MS-COCO. While 90% of the trusted data has a positive score (high quality), it holds for only 1% of the self-generated data.

Luckily,  $p(y_j|x_j; \theta)$  can be estimated by language models that learn to predict tokens. We extend [55] to consider the additional input image, as follows. Given a caption  $y_j$ , which is a sequence of  $m$  tokens  $[w_1, \dots, w_m]$ , and an image  $x_j$ , the likelihood of the caption depends on the likelihood of each token to appear, based on the previous tokens and the image. For a token  $w_t$  the probability is  $P_\theta(w_t|w_{1:t-1}, x_j)$  and the log-likelihood of the caption is

$$\log p(y_j|x_j; \theta) = \frac{1}{m} \sum_{t=1}^m \log (P_\theta(w_t|w_{1:t-1}, x_j)). \quad (2)$$

Our data selection algorithm uses this estimation to gradually filter out data points below a *quality threshold*. As the training proceeds, the threshold increases, eliminating more and more self-generated data points. Fig. 4 shows that low-quality data is filtered at early iterations, while higher-quality data remains until later iterations.

A unique challenge that differentiates our length-control task from other tasks that use diverse-quality datasets is that

almost all long captions tend to have low-quality scores. Therefore, long captions are prone to catastrophic forgetting, as they are eliminated early on. This could reduce the control abilities for long captions. To avoid this situation, we make sure that low-quality captions (e.g., the long ones) will appear in late training iterations. This, however, might contradict the core idea of our training, which requires only high-quality captions to appear at late training stages. We propose to allow a small number of low-quality self-generated captions to appear as the training progresses, by adding some randomness, as follows.

Without adding the above randomness, we may view our filtering procedure at the  $i^{th}$  iteration as sampling according to a probability function, where data points with a quality score  $< T_i$  get probability 0 and all other points get the same positive probability. This function is essentially a step function, centered at  $T_i$ , which maps quality scores to probabilities. We add randomness by smoothing the step function and allowing a small number of low-quality data points to be selected accordingly.

In particular, given a threshold,  $T_i$ , we create a smooth step function in accordance with  $T_i$  and perform data selection, as follows. The caption quality  $u_j$  of a data point  $j$  (Eq. 1) and  $T_i$  are used to compute the probability function needed for data selection. We use the following smooth step function, centered at the threshold  $T_i$ :

$$f(u_j; T_i, s) = \frac{1}{2} \cdot \left[ 1 + \tanh \left( \frac{u_j - T_i}{s} \right) \right]. \quad (3)$$

Here,  $s$  is a tunable parameter, which controls the smoothness of the function. For  $s \rightarrow 0$ , we get an ideal step function, which means that values smaller than  $T_i$  have 0 probability to be sampled. The larger  $s$  is, the greater the probability of low-quality data points to be selected. Since self-generated long captions are usually assigned low quality scores (Fig. 4), a small  $s$  might cause long captions to be eliminated early on in the training process. Reversely, if  $s$  is too high, too much of out-of-distribution data will appear at late training iterations. We note that an additional benefit of our scheme is that it controls all low-quality captions and not only long ones.

**Training—putting it all together.** Hereafter we recap our training procedure with data selection. We are given a trusted dataset  $D$ , an extended dataset  $\tilde{D}$ , and a step smoothness value  $s$ . Two captioning models are trained, one for the trusted dataset and the other for the extended dataset,  $M_\theta$  and  $M_{\tilde{\theta}}$ , respectively. The quality score  $u_j$  of each data point in the self-generated dataset is estimated using the two models (Eq. 1). Then, the captioning model is trained with iterative data selection until all the self-generated data is filtered out, as follows. At iteration  $i$ ,

1. A new threshold  $T_i$  is computed. Specifically, it is determined by the amount of the generated data to be re-

Level Length (tokens)	1 1-9	2 10-19	3 20-29	4 30-39	5 40-49	6 50-59	7 60-69	Average
% in trusted dataset	11.8%	86.6%	1.4%	0.08%	0.02%	0.004%	0%	
% in extended dataset	9.1%	51.8%	13.1%	13.5%	7.7%	3.8%	0.1%	
LaBERT [14]	<b>100%</b>	98.03%	<b>93.75%</b>	83.78%	49.92%	51.19%	0%	68.10%
CLID (ours)	<b>100%</b>	<b>98.64%</b>	92.6%	<b>84.17%</b>	<b>76.99%</b>	<b>69.18%</b>	<b>22.13%</b>	<b>77.67%</b>

Table 1. **Captioning control precision.** For each level, the second row shows the range of tokens for this level. The next two rows show the percentage of captions of each level in the training datasets. The two bottom rows compare the precision results of [14] to ours. We outperform [14] both on average and for most of the levels, with the exception of Level 3 (which is a prevalent level in the trusted dataset). As expected, our benefit is mostly evident for the high levels. The results are the mean of 3 independent runs.

moved, according to the  $c \cdot i$  percentile. For instance, at the 5<sup>th</sup> iteration and for  $c = 2\%$ ,  $T_5$  is set such that 10% of the data will be below the threshold.

2. A weight for each data point,  $f(u_j; T_i, s)$ , is computed, depending on its quality score, the threshold and the smoothness value (Eq. 3).
3. The self-generated dataset is randomly sampled, based on these weights.
4. The model is trained on the trusted & the sampled self-generated datasets (for duration  $\eta$ ).

See the supplementals for the model and pseudo-code.

## 4. Experimental results

**Datasets.** We use *MS-COCO Caption* [10], which is the most common captioning dataset in general, and the only dataset used in a previous length-control work [14]. It contains 5 captions per image, with 113, 287 training images, 5000 validation images and 5000 testing images, according to the widely-used Karpathy’s split [30]. As seen in Table 1, Level 2 contains the majority of the captions (87%). There are hardly any longer captions and no captions in level 7.

The *Visual Genome (VG)* [34] dataset contains scene graphs for images. Its intersection with MS-COCO is around 50%. For the other MS-COCO images, we use a scene graph generator [22].

*Flickr-30k* [48] is evaluated in the supplementary material. For this dataset, we utilize only automatically generated scene graphs.

**Evaluation metrics.** In addition to checking whether a description meets the length constraint, we measure its quality. We use the *control precision* metric [14] to measure the percentage of descriptions of the desired length. There are several possible NLP metrics for measuring quality that calculate the similarity between a generated description and the ground truth, including BLEU [47], ROUGE [37], METEOR [5], CIDEr [56], and SPICE [2]. As in [14], we focus on SPICE, which is more robust to caption length and has a high correlation with human judgement.

**Length-control results.** Table 1 demonstrates the effectiveness of our approach in terms of control. The control

precision is excellent for captions of Levels 1-3 and then starts dropping. Our approach significantly outperforms the average results of [14]. Specifically, it improves the results in all levels, except for Level 3, and vastly improves the results of the high levels.

**Quantitative results.** Fig. 5 shows that our method generates good-quality results, yet with higher control. The quality of our captions (orange) is comparable to those of [14], trained on the trusted data (green). It is also compared to the quality of the results of two other possible solutions. In the first the model is trained on a length-balanced version of the trusted dataset (purple), and in the second on the extended dataset (gray). These approaches have high control precision, but on the expense of quality. The supplemental material extends Fig. 5 to all the levels and exhibits similar results. It also shows results of the other metrics, demonstrating similar improvements. Furthermore, it includes additional results on Flickr-30k that exhibit a similar behavior.

In order to demonstrate the impact of high-quality information from short captions on long caption generation, we observe Level 7, which is absent from the trusted captioning dataset. When training a captioning model on the extended dataset without any modifications, the average SPICE score for captions at this level is 11.49. However, by employing our training procedure, which does not involve the addition of any ground-truth caption at this length level, the same model achieves a significantly improved score of 19.54.

**Qualitative results.** Fig. 1 illustrates representative outputs generated by our model. Notably, our descriptions for higher levels demonstrate greater elaboration and encompass more scene-specific details compared to the descriptions produced by [14]. However, short descriptions are comparable for both models. See supplementary materials for more qualitative results.

To assess the quality of our captions, as rated by humans, we conducted an experiment on Amazon Mechanical Turk. The experiment involved presenting a triplet consisting of an image and two captions: one generated by [14] and one generated by our model. Workers were asked to determine which caption better described the image, or if both captions were equally good. We randomly selected 300 image-

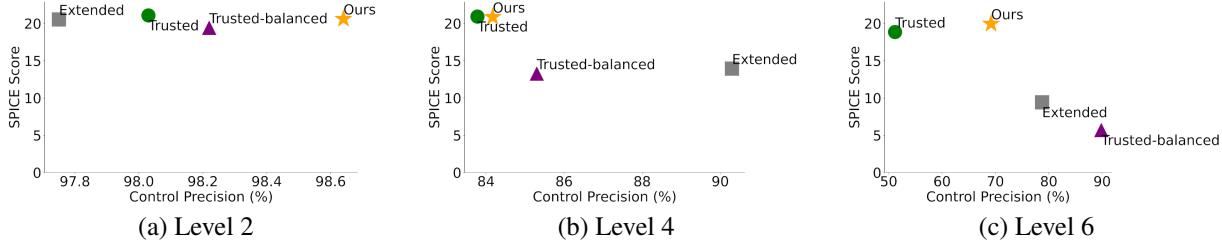


Figure 5. **Captioning performance.** In terms of the SPICE quality measure (vertical axis), our results (orange star) are similar to [14]’s (green circle), which is trained on the trusted dataset. The quality of other solutions (gray/purple) is dramatically degraded. While comparable to [14] quality-wise, our model improves the control precision (horizontal axis). In both measures, higher is better. The figure shows 3 length levels; the other levels appear in the supplements.

	Level 2	Two men in a white t-shirt are playing frisbee.
LaBERT [14]	Level 5	A man is playing a park. He is reaching for a frisbee. The man is wearing a white shirt with white stripes down the side. The frisbee is white and has a lot of grass on it.
	Level 8	Two men are playing frisbee on a grassy field. Both of the men are wearing black t-shirts with white stripes down the side of the shirts. One man is holding a frisbee in his hand. The field is covered in green grass. Part of green trees can be seen behind the men. It appears to be daylight outside.
	Level 2	A man in a white shirt and black shorts holds a frisbee.
CLID (ours)	Level 5	A man is standing outside on a grassy field. He is holding a white frisbee in his hand. The man is wearing a white t-shirt and black shorts. There are trees behind the man in the background.
	Level 8	A man is playing frisbee on a field. He is holding a frisbee in his hand. The man is wearing black and white shirt and shorts. The man has black and white stripes on the shorts. He has dark hair. The field is green with green grass behind the trees in the field. The man is standing in the field.
	(a) Length	(b) Generated Caption

Figure 6. **Qualitative comparison (paragraphs).** Comparing our outputs to those of [14] demonstrates similar quality for the short descriptions. However, our long descriptions are more coherent and detailed than [14]’s.

Method	Level					
	1	2	3	4	5	6
LaBERT [14]	40%	44%	39%	47%	41%	43%
CLID (ours)	<b>60%</b>	<b>56%</b>	<b>61%</b>	<b>53%</b>	<b>59%</b>	<b>57%</b>

Table 2. **User study results.** Percentage of votes for captions generated by each method, per length level. Human evaluators tend to prefer our captions over those of [14].

caption triplets and assigned each to three different workers.

Table 2 presents the percentage of worker votes for each model, per caption length (ranging from 1 to 6, the levels which both models can generate caption in). The table reveals that our captions were preferred by human evaluators across all length levels. For instance, in level 1, 60% of workers preferred our captions, while in level 6, 57% of workers favored our captions. An interesting observation is that even though our focus was not primarily on the lower length levels (levels 1-3) due to the abundance of available data, the human evaluation demonstrated a quality advantage in these levels as well.

**Paragraph generation.** As our proposed method is gen-

eral, we address the task of paragraph generation similarly to captioning. A paragraph consists of consecutive sentences separated by dots. We consider them as a sequence of text tokens, as in captioning. To create the self-generated dataset, we used the algorithm described in Section 3.1. The only difference is that a dot was added instead of an “and” when jumping during the DFS traversal.

We use the *Stanford Descriptive Paragraphs* [33] dataset, which is the only available dataset for the task. The images in this dataset are taken from VG. It contains 14,575 training images, 2,487 validation images and 2,489 test images. Each image is associated with a single paragraph. Hence, this dataset is much smaller than the captioning dataset (by a factor of 30). 2% of the paragraphs have less than 29 tokens; 32% have 30-59 tokens; 52% have 60-89 tokens; and 14% have 90-129 tokens. We set 13 uniform length levels, which range from 1-129 tokens.

Since there are no works on length-control paragraph generation, we trained [14] to generate paragraphs, in order to allow comparisons. Our method improves the control abilities at all length levels, compared to training only on

$T_{sal}$	Level						
	1	2	3	4	5	6	7
20%	100%	98.67%	92.55%	83.24%	74.44%	39.88%	0.1%
50%	100%	97.95%	90.9%	84.2%	78.54%	66.3%	5.3%
80%	100%	98.64%	92.60%	84.17%	76.99%	69.18%	22.13%

Table 3. **Graph coverage vs. control precision (captioning).** Higher values of  $T_{sal}$  allow more diverse descriptions in the self-generated dataset, enhancing the control over long descriptions.

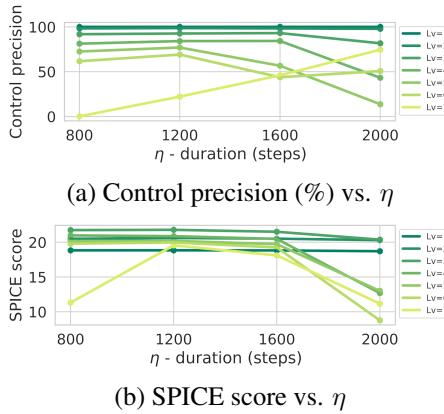


Figure 7. **Ablation study of the  $\eta$  parameter.** (a) Control precision and (b) SPICE score per length level (lighter color = longer captions) as a function of the duration  $\eta$  at  $s = 1$ .

the trusted data. It achieves an average of 94% in control precision, which is a 14% gain. In terms of the the common NLP scores (e.g., BLEU), our model achieves comparable results to [14]’s across all levels. Additionally, Fig. 6 shows typical outputs of our model. While our short descriptions are comparable to those of [14], our long descriptions are more coherent and detailed. Please refer to the supplemental materials for further details and examples.

## 5. Ablation study

**Graph coverage.** The amount of information extracted from a scene graph is determined by  $T_{sal}$  (%). A low value of  $T_{sal}$  results in overall short generated description datasets. Table 3 illustrates the impact of this parameter on the model’s control ability. While the control ability is comparable for levels 1-4, which are prevalent in all of these self-generated datasets, the advantage of using larger  $T_{sal}$  values becomes pronounced in longer levels.

**Data selection parameters.** Our approach has three parameters: (1)  $c$  that determines  $T_i$ , i.e. the filtration rate between consecutive iterations; (2)  $\eta$  that sets the duration (number of steps) of training the model at every iteration; and (3)  $s$  from Eq. 3, that determines the smoothness of the step function that sets the sampling weight of a data point.

The smaller  $c$  is, the more gradual filtration is. We observed a similar performance gain when dropping 1-5% of the data. Thus, we set  $c = 2\%$ .

If  $\eta$  is large, sampled data points at a certain iteration are

repeated many times, and therefore have more influence on the model. If  $\eta$  is small, each iteration is shorter and the self-generated data is sampled more often, potentially exposing the model to more diverse data, since at every iteration new data is selected.  $s$  determines the sampled data diversity at every iteration. A small  $s$  restricts the randomness of data sampling. Large values of  $s$  give more probability to data points with low-quality to appear at later stages, leading to more diverse data points between consecutive iterations.

We observed a connection between the values of  $\eta$  and  $s$ . Increasing  $\eta$  usually requires decreasing  $s$ . This is logical, since substantial changes in data (high  $s$ ) during consecutive long iterations (high  $\eta$ ) can resemble training the model repeatedly on different domains, potentially causing deviation from the desired domain represented by trusted data. The effect of high values of  $s$  is enhanced when replacing the smooth step function with a constant function (assigning the same probability to all data points at every iteration). This results in up to a 40% decline in output quality for levels 4-6, where data is limited.

To study the effect of  $\eta$  &  $s$ , we experiment with  $\eta = \{800, 1200, 1600, 2000\}$  and  $s = \{0.1, 0.5, 1, 1.4\}$ . Levels 1-3, which are 99.8% of the trusted captioning dataset, are hardly affected by these parameters. Figure 7 shows the influence of  $\eta$  on our results, for  $s = 1$ . As shown, short captions are barely affected by changing  $\eta$ . For the long captions that appear in the dataset (Levels 4-6), there is a sweet-spot at  $\eta = 1200$  in terms of control and quality. For level 7, which does not exist in the trusted dataset, we observe a low control precision for small  $\eta$ . Increasing  $\eta$  solves this problem. We set  $s = 1$  and  $\eta = 1200$  for the best overall performance ( $\eta = 200$  for paragraph generation).

**Limitations.** Levels that are completely out-of-distribution (7 or up) do not always produce sought-after caption quality.

## 6. Conclusions

We propose a novel, general and unified method to address the shortage of certain length descriptions in common image captioning & paragraph generation datasets. Our approach consists of two complementary ideas. First, we show how to enrich the existing dataset with self-generated varying-length descriptions, using scene graphs and saliency maps. Second, we introduce a training procedure that gets both a trusted (original) dataset and a self-generated one. It gradually trains the model so as to learn from the (varying-length) low-quality data, while not harming the information learned from the clean trusted data.

We show that the length-control abilities vastly improve. Our method achieves 10% (14%) improvement on average in caption (paragraph) control precision. This is done while preserving the quality of the output across all length levels.

**Acknowledgements.** We gratefully acknowledge the support of the Israel Science Foundation 2329/22.

## References

[1] Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone. Clue: Cross-modal coherence modeling for caption generation. *arXiv preprint arXiv:2005.00908*, 2020. 2

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016. 6

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 2

[4] Amitai Axelrod. Cynical selection of language model training data. *arXiv preprint arXiv:1709.02279*, 2017. 3

[5] Satanjeev Banerjee and Alon 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*, pages 65–72, 2005. 6

[6] Moitreya Chatterjee and Alexander G Schwing. Diverse and coherent paragraph generation from images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 729–744, 2018. 2

[7] Long Chen, Zhihong Jiang, Jun Xiao, and Wei Liu. Human-like controllable image captioning with verb-specific semantic roles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16846–16856, 2021. 1, 2

[8] Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9962–9971, 2020. 2

[9] Tianlang Chen, Zhongping Zhang, Quanzeng You, Chen Fang, Zhaowen Wang, Hailin Jin, and Jiebo Luo. “factual” or “emotional”: Stylized image captioning with adaptive learning and attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 519–535, 2018. 2

[10] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 1, 2, 6

[11] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1841–1850, 2019. 4

[12] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, control and tell: A framework for generating controllable and grounded captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8307–8316, 2019. 1, 2

[13] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587, 2020. 2

[14] Chaorui Deng, Ning Ding, Mingkui Tan, and Qi Wu. Length-controllable image captioning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 712–729. Springer, 2020. 1, 2, 3, 6, 7, 8

[15] Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander G Schwing, and David Forsyth. Fast, diverse and accurate image captioning guided by part-of-speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10695–10704, 2019. 2

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. 2

[17] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 2

[18] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*, pages 15–29. Springer, 2010. 1

[19] Lizhao Gao, Bo Wang, and Wenmin Wang. Image captioning with scene-graph based semantic concepts. In *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*, pages 225–229, 2018. 3

[20] Shalini Ghosh, Giedrius Burachas, Arijit Ray, and Avi Ziskind. Generating natural language explanations for visual question answering using scene graphs and visual attention. *arXiv preprint arXiv:1902.05715*, 2019. 3

[21] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, Handong Zhao, Xu Yang, and Gang Wang. Unpaired image captioning via scene graph alignments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10323–10332, 2019. 3

[22] Xiaotian Han, Jianwei Yang, Houdong Hu, Lei Zhang, Jianfeng Gao, and Pengchuan Zhang. Image scene graph generation (sgg) benchmark, 2021. 3, 6

[23] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. *Advances in neural information processing systems*, 31, 2018. 3

[24] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural*

*Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2

[25] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013. 1

[26] HuggingFace. Pegasus paraphraser. <https://huggingface.co/tuner007/pegasus-paraphrase>, 2020. 4

[27] Sen Jia and Neil D.B. Bruce. Eml-net: An expandable multi-layer network for saliency prediction. *Image and Vision Computing*, 95:103887, 2020. 1, 3

[28] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228, 2018. 3

[29] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015. 1, 3

[30] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 6

[31] Marc A Kastner, Kazuki Umemura, Ichiro Ide, Yasutomo Kawanishi, Takatsugu Hirayama, Keisuke Doman, Daisuke Deguchi, Hiroshi Murase, and Shin’ichi Satoh. Imageability- and length-controllable image captioning. *IEEE Access*, 2021. 2

[32] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Dense relational captioning: Triple-stream networks for relationship-based captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6271–6280, 2019. 2

[33] Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 7

[34] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 3, 6

[35] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1910–1918, 2017. 3

[36] Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P Xing. Recurrent topic-transition gan for visual paragraph generation. In *Proceedings of the IEEE international conference on computer vision*, pages 3362–3371, 2017. 2

[37] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 6

[38] Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. In *International Conference on Machine Learning*, pages 6226–6236. PMLR, 2020. 4

[39] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7219–7228, 2018. 2

[40] Yadan Luo, Zi Huang, Zheng Zhang, Ziwei Wang, Jingjing Li, and Yang Yang. Curiosity-driven reinforcement learning for diverse visual paragraph generation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2341–2350, 2019. 2

[41] Yuzhao Mao, Chang Zhou, Xiaojie Wang, and Ruifan Li. Show and tell more: Topic-oriented multi-sentence image captioning. In *IJCAI*, pages 4258–4264, 2018. 2

[42] Alexander Mathews, Lexing Xie, and Xuming He. Semstyle: Learning to generate stylised image captions using unaligned text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8591–8600, 2018. 1

[43] Gaurav Mittal, Shubham Agrawal, Anuva Agarwal, Sushant Mehta, and Tanya Marwah. Interactive image generation using scene graphs. *arXiv preprint arXiv:1905.03743*, 2019. 3

[44] Robert C Moore and William Lewis. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 conference short papers*, pages 220–224, 2010. 3

[45] Jiteng Mu, Weichao Qiu, Gregory D Hager, and Alan L Yuille. Learning from synthetic animals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12386–12395, 2020. 4

[46] Kien Nguyen, Subarna Tripathi, Bang Du, Tanaya Guha, and Truong Q Nguyen. In defense of scene graphs for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1407–1416, 2021. 2

[47] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6

[48] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 6

[49] Mengshi Qi, Yunhong Wang, and Annan Li. Online cross-modal scene retrieval by binary representation and semantic graph. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 744–752, 2017. 3

[50] Yu Qin, Jiajun Du, Yonghua Zhang, and Hongtao Lu. Look back and predict forward in image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8367–8375, 2019. 2

[51] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmenta-

tion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3752–3761, 2018. 4

[52] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8376–8384, 2019. 3

[53] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3716–3725, 2020. 3

[54] Subarna Tripathi, Sharath Nittur Sridhar, Sairam Sundaresan, and Hanlin Tang. Compact scene graphs for layout composition and patch retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 5

[56] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 6

[57] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 839–847, 2017. 3

[58] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 1, 2

[59] Jing Wang, Yingwei Pan, Ting Yao, Jinhui Tang, and Tao Mei. Convolutional auto-encoding of sentence topics for image paragraph generation. *arXiv preprint arXiv:1908.00249*, 2019. 2

[60] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1508–1517, 2020. 3

[61] Wei Wang, Taro Watanabe, Macduff Hughes, Tetsuji Nakagawa, and Ciprian Chelba. Denoising neural machine translation training with trusted data and online data selection. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 133–143. Association for Computational Linguistics, 2018. 3, 4

[62] Yufei Wang, Zhe Lin, Xiaohui Shen, Scott Cohen, and Garrison W Cottrell. Skeleton key: Image captioning by skeleton-attribute decomposition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7272–7281, 2017. 2

[63] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699, 2015. 3, 4

[64] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5419, 2017. 3

[65] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 2

[66] Xu Yang, Chongyang Gao, Hanwang Zhang, and Jianfei Cai. Hierarchical scene graph encoder-decoder for image paragraph captioning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4181–4189, 2020. 2

[67] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10685–10694, 2019. 2, 3

[68] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 1

[69] Zheng-Jun Zha, Daqing Liu, Hanwang Zhang, Yongdong Zhang, and Feng Wu. Context-aware visual policy network for fine-grained image captioning. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 2

[70] Ji Zhang, Kevin J Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11535–11543, 2019. 3

[71] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 2

[72] Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat, and Kevin Duh. Curriculum learning for domain adaptation in neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1903–1915. Association for Computational Linguistics, 2019. 3

[73] Zhihu Zhang and Mert R Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *32nd Conference on Neural Information Processing Systems (NeurIPS)*, 2018. 4

[74] Yiwu Zhong, Liwei Wang, Jianshu Chen, Dong Yu, and Yin Li. Comprehensive image captioning via scene graph decomposition. In *European Conference on Computer Vision*, pages 211–229. Springer, 2020. 1, 2