

# On Decoding and Discourse Structure in Multi-Modal Text Generation

Nikolai Ilinykh      Simon Dobnik

Centre for Linguistic Theory and Studies in Probability (CLASP)  
Department of Philosophy, Linguistics and Theory of Science (FLoV)  
University of Gothenburg, Sweden  
{nikolai.ilinykh, simon.dobnik}@gu.se

## Abstract

This paper describes insights into how different inference algorithms structure discourse in image paragraphs. We train a multi-modal transformer and compare 11 variations of decoding algorithms. We propose to evaluate image paragraphs not only with standard automatic metrics, but also with a more extensive, “under the hood” analysis of the discourse formed by sentences. Our results show that while decoding algorithms can be unfaithful to the reference texts, they still generate grounded descriptions, but they also lack understanding of the discourse structure and differ from humans in terms of attentional structure over images.

## 1 Introduction

What are the properties of the well-generated text? This question has been in the centre of many debates in the natural language generation community (Dale and White, 2007; Gatt and Krahmer, 2018). While human evaluation has always been the gold standard in the quality assessment of generated texts, the field is often reluctant to run such evaluation due to the lack of standardisation in evaluation reports and generally high cost (Howcroft et al., 2020). Therefore, a number of simpler and cheaper *automatic metrics* were introduced, specifically in the field of machine translation, although their validity has been questioned (Reiter and Belz, 2009).

As computer vision and NLP started to merge, automatic metrics became an important part of the evaluation process of image descriptions. In general, image descriptions are evaluated with means of BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), ROUGE (Lin, 2004), CIDER (Vedantam et al., 2015) and WMD (Kusner et al., 2015). However, Kulkarni et al. (2011) and Elliott and Keller (2014) have demonstrated that such metrics only weakly correlate with human judgements in the context of image description generation task. The discrepancy between human and automatic

evaluation is deeply rooted in the differences between the fields of machine translation which originally introduced aforementioned metrics and image captioning, which adopted them. In principle, text-only evaluation is highly constrained: the key requirement for high-quality translation is the perseverance of semantics between two parallel texts. In comparison, evaluation of texts generated in multi-modal tasks is influenced by many factors as the generated texts might mention a different set of objects, attributes and relations which are not described in reference texts. Such generations would cause low values from reference-based metrics, although they could be completely plausible and truthful to the image. As such, the tasks of machine translation and image captioning are inherently dissimilar in terms of evaluation. To mitigate this problem, metrics that directly compare texts against image objects have been proposed (Jiang et al., 2019; Madhyastha et al., 2019; Wang et al., 2021; Hessel et al., 2021). They are typically better than BLEU in that they assign a more accurate score to image-correct descriptions. A relatively recent trend has been to develop a set of metrics that would evaluate goal-oriented captions, produced with specific communicative intent (Inan et al., 2021) or for a specific group of users (Fisch et al., 2020), for example, if an image of a snowdrop is described as “the spring flower”.

A notable feature of the aforementioned metrics is their sole focus on evaluation of image captions. Different from captions, **multi-sentence** image descriptions impose additional challenges for generation systems including understanding of the textual *discourse* in the multi-modal context. Analysis of discourse has been in the focus of both text-only (Poesio et al., 2004; Poesio, 2004) and language-and-vision tasks (Takmaz et al., 2020; Dobnik et al., 2022). However, given a huge interest in generation of longer image descriptions, e.g. image paragraphs (Kong et al., 2014; Krause

et al., 2017; Ilinykh and Dobnik, 2020), recipes (Nishimura et al., 2019) and stories (Huang et al., 2016), we believe it is important to gain a deeper insight into how humans and models *structure* and *realise* discourse in such descriptions. In this paper, we understand discourse as a match between linearisation of the semantic knowledge (e.g., a fit of non-linear concepts into linguistic linear order) and underlying planning (Reiter and Dale, 1997). We build on previous intuitions about evaluation in NLG and look under the hood of how different decoding algorithms build discourse in image paragraphs. We compare a number of decoding strategies for correspondence with how humans distribute and describe objects in longer texts. The main purpose of this study is to gain insights into whether decoding strategies generate texts *similar* to humans and whether these texts exhibit the corresponding discourse structure. There is a limitation on what and how things can be communicated and decoding algorithms have a direct control over it. The choice of decoding algorithm also has an effect on how information is expressed in the communicative channel (Shannon, 1948) and how successful its reconstruction by the perceiver will be (Lazari-dou et al., 2017). Our results shed more light on the differences between decoding algorithms in terms of (i) the discourse structure, (ii) faithfulness to the reference texts, (ii) groundedness into the image and (iv) attentional structure.

## 2 On the importance of decoding

It is impossible to neglect the impact of the choice of the decoding on the structure of the generated texts<sup>1</sup>. Discourse in multi-modal descriptions can be affected by many factors, including scene structure (Linde and Goguen, 1980), the desire to have more accurate or more diverse texts (Massarelli et al., 2020; Zhang et al., 2021) and aspects of the task (Kiddon et al., 2016; Narayan et al., 2022). Other constraints include adherence to a specific topic as in poetry generation by controlling for content and form (Hopkins and Kiela, 2017) and incorporating pragmatic reasoning when describing images with text (Cohn-Gordon et al., 2018; Vedantam et al., 2017) or optimising model’s predictions for a specific metric (Rennie et al., 2017; Gu et al., 2017; Zarri   and Schlangen, 2018) in the spirit of reinforcement learning. Notably, Balakrishnan

et al. (2019) have shown that using tree-structured semantic representations, similar to those used in traditional rule-based NLG systems, helps to evaluate generated texts during decoding for the specific discourse. In this work, we describe analysis on *what* and *when* different algorithms generate, comparing their outputs with the human gold standard.

## 3 Task and model

As our modelling task, we choose the task of image paragraph generation and the Tell-me-more corpus described in (Ilinykh et al., 2019). In this task, a human is given an image and five (5) text fields. The describer writes sentences about the image so that they help a potential listener to identify it within a set. The describer is also asked to write sentences in a sequence, keeping in mind that after each sentence the listener needs more information to identify the image, e.g. thus, tell-me-more. Ilinykh et al. (2019) show that collected multi-sentence descriptions have a fixed intentional structure, in the sense of Grosz et al. (1995), but attention structure demonstrates a different behaviour as supported by the analysis in (Dobnik et al., 2022).

As our model, we use the architecture of the object relation transformer proposed by Herdade et al. (2019)<sup>2</sup>. This is a two-stream multi-modal transformer, which consists of three self-attention blocks, operating on the image, text and across modalities. Each block has the standard parts of the transformer (Vaswani et al., 2017): multi-head self-attention followed by a feed-forward network, residual connection and layer normalisation.

On the vision side, the model takes the set of pre-extracted visual features of detected objects, which we receive by using the object detector released by Anderson et al. (2018)<sup>3</sup> and pre-trained on Visual Genome (Krishna et al., 2016). Specifically, every object  $o_j$  in the the set of detected image objects  $\mathbf{O} = (o_1, \dots, o_{|\mathbf{O}|})$  has a visual feature  $v_n \in \mathbb{R}^{1 \times D}$ , where  $|\mathbf{O}| = 36$  and  $D = 2048$ . In addition, we store other outputs of the object detector, including object labels, attributes and confidence scores. They will be used in later stages to link paragraphs with objects in the image. The benefit of the object relation transformer is its ability to encode complex geometric relations between bounding boxes. Thus, we also extract the set of

<sup>1</sup>For a broader overview of the factors that influence inference in generation we refer the reader to Zarri   et al. (2021).

<sup>2</sup>[https://github.com/yahoo/object\\_relation\\_transformer](https://github.com/yahoo/object_relation_transformer)

<sup>3</sup><https://github.com/peteanderson80/bottom-up-attention>

geometric features  $\mathbf{G} = \{x, y, w, h\}$ , which are fused with visual features inside the model<sup>4</sup>.

On the textual side, the model generates a paragraph word by word in auto-regressive fashion. Specifically, it takes the current token  $w_j$  and constructs its representation based on previously generated tokens  $w_1, \dots, w_{j-1}$ . All the future tokens in the paragraph  $w_{j+1}, \dots, w_{|\mathbf{W}|}$  are replaced with the MASK token, framing the task as the classic next word prediction task. The generation starts with the START token and ends when either the maximum length of the paragraph  $\mathcal{L}$  is reached or when the END token is generated. As the last step, representation from two self-attention blocks are processed by the cross-attention which outputs the probability of all tokens from the vocabulary  $\mathcal{V}$ .

In terms of model’s parameters, we keep all of them untouched, thus they correspond to the original set of parameters described in [Herdade et al. \(2019\)](#). We train the model on the full Tell-me-more dataset, consisting of 3590 image-paragraph pairs in the train set and 410 pairs in both validation and test sets. The analysis in this paper is performed on the test set only.

#### 4 Decoding algorithms

Given the model vocabulary  $\mathcal{V}$  and  $\mathcal{L}$  as the maximum length of the generated sequence, the space of possible sequences has  $|\mathcal{V}|^{\mathcal{L}}$  members, thus, becoming intractable. Rather than traversing through such space, a number of different decoding methods are used to find the most likely sequence. The most straightforward heuristics is to take the most probable word  $w$  at timestamp  $j$  until either the maximum length of the generated sequence  $\mathbf{w}$  is reached ( $\mathcal{L} = 100$ ) or the END token is generated. We employ standard **greedy search**:

$$w_j = \operatorname{argmax}_{w'_j} \log p(w'_j | \mathbf{w}_{<j}, \mathbf{O}; \theta), \quad (1)$$

where  $\mathbf{w}_{<j} = (w_1, \dots, w_{j-1})$  is the sequence of previously predicted words,  $\mathbf{O} = (o_1, \dots, o_{|\mathbf{O}|})$  is the set of detected image objects and  $\theta$  is the set of model parameters. Despite its simplicity and low complexity, greedy search is known for its sub-optimality on the global sentence level ([Gu et al., 2017](#); [Chen et al., 2018](#)), often leading to generation problems such as the garden path sentence issue ([Gibson, 1991](#)).

<sup>4</sup>We refer the reader to ([Herdade et al., 2019](#)) for more details.

A more popular and standardized approach is to use **beam search**, a version of the breadth-first search, that tracks multiple candidate sequences  $\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_k)$  and chooses the one with the highest cumulative probability score, frequently computed as summation of word scores in each sequence. Typically, the most probable sequence is picked as the final one, but other sequences can also be considered. The search starts with the word sequence  $\mathbf{w}_1 = \{\text{START}\}$  and continues until the length of every predicted sequence reaches the maximum length  $\mathcal{L}$  or all of them are completed with the END token:

$$\mathbf{w}_j = \operatorname{argmax}_{\substack{\mathbf{w}'_j \subseteq \mathcal{B}_j, \\ |\mathbf{w}'_j| = k}} \log p(\mathbf{w}'_j | \mathbf{w}_{j-1}, \mathbf{O}; \theta). \quad (2)$$

In beam search, the parameter  $k$  denotes the number of desired sequence candidates and  $\mathcal{B}$  stands for the set of sequences currently under generation. Beam search is computationally more expensive, but it is also more efficient in finding the optimal sequence due to more sophisticated exploration of the word space. However, bigger  $k$  often leads to “safe” and generic texts and candidate generations themselves can resemble each other a lot, lacking diversity ([Li et al., 2016](#)) or becoming repetitive ([Holtzman et al., 2020](#)).

The problems of beam search have been addressed by many different approaches, mostly focused on increasing intra-set diversity of generated sequences ([Kulikov et al., 2019](#); [Meister et al., 2021](#)). In one of such approaches, [Vijayakumar et al. \(2018\)](#) propose to extend beam search by incorporating a *dissimilarity* term in the objective function. Specifically, **diverse beam search** splits beam sets into  $G$  groups  $W^1, \dots, W^G$  and at each word generation timestamp  $j$  for every sequence in the current group  $\mathbf{w}_j^g \in W^g$ , it encourages diversity with sequences from previous groups  $W^h, h \leq g$  using a metric of dissimilarity  $\Delta$ :

$$W_j^g = \operatorname{argmax} \sum_{k \in [B']} \log p(\mathbf{w}_{k,[j]}^g) + \lambda \sum_{h=1}^{g-1} \Delta(\mathbf{w}_{k,[j]}^g, W_{[j]}^h), \quad (3)$$

where  $B'$  is the number of beams in each group,  $\lambda$  is the parameter that controls the diversity,  $\Delta$  is

the Hamming distance, which negatively penalises sequences sharing identical n-grams. Diverse beam search has been specifically designed to boost diversity in the multi-modal description generation task, where focus is to mimic human texts with shifts between many objects, relations and specific details. However, as reported by the authors, the best results in terms of diversity are achieved by using a simple n-gram-based heuristics, which does not take the multi-modal nature of the task into account. In addition, diversity is encouraged between beam sets on the group level rather than between sentences within a single group, limiting the scope of diversity on the sentence level. Finally, the look-up over groups is constrained to the current word position at each generation step, shrinking the context window for the currently generated word and possibly capping the number of satisfactory generations at this timestamp.

A very different method to encourage more diverse output is to sample from the word distribution. For obvious reasons pure sampling leads to incoherent and grammatically incorrect texts. Therefore, **top- $k$  sampling** has been proposed by Fan et al. (2018): the method cuts the probability distribution and keeps the distribution  $p'$  consisting of top  $k$  tokens with the highest probability:

$$w_j \sim \log p'(w_j \mid \mathbf{w}_{<j}, \mathbf{O}; \theta). \quad (4)$$

A known issues with the top- $k$  sampling algorithm is that it is hard to find the optimal value for the parameter  $k$  since setting it too low could remove highly probable words or, on the contrary, keep the less probable words if it is too high.

Instead of relying on pre-defined number of tokens, **nucleus sampling** (Holtzman et al., 2020) takes words from the subset of the vocabulary in which the defined probability mass is concentrated:

$$p' = \sum_{w_j \in \mathcal{V}'} \log p(w_j \mid \mathbf{w}_{<j}, \mathbf{O}; \theta) \geq p, \quad (5)$$

where  $\mathcal{V}'$  is the top- $p$  part of the vocabulary  $\mathcal{V}$ , in which only the words that accumulate most of the probability mass are kept. Parameter  $p$  is typically used to define the maximum value of accumulated probability. The original distribution is then re-scaled and the next word is sampled from the new distribution  $P$ :

$$P = \begin{cases} \log p(w_j \mid \mathbf{w}_{<j}, \mathbf{O}; \theta) / p' & \text{if } w_j \in \mathcal{V}' \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

The main advantage of nucleus sampling is its ability to track the shape of the probability distribution, allowing for dynamic control of the number of candidates at each timestamp. A different, but related method to introduce controlled randomness is to use **temperature scaling**. The diversity is achieved by controlling the peaks in the distribution and dividing it by the parameter  $\tau$ :

$$p(w_j \mid \mathbf{w}_{<j}, \mathbf{O}; \theta) = \frac{\exp(\varphi_j / \tau)}{\sum_{w_j \in \mathcal{V}} \exp(\varphi_j / \tau)}, \quad (7)$$

where  $\varphi_j$  is the logit for a word  $w_j$  in the vocabulary. Lower temperatures are known to enforce the high probability events and choosing a proper value for this parameter can lead to better texts in terms of quality and diversity (Caccia et al., 2020).

We note that in this work we mainly focus on the most frequently used decoding strategies, excluding analysis of the result of more direct manipulations with texts such as length normalisation and coverage penalty (Wu et al., 2016), n-gram blocking or introduction of the noise model (Hill et al., 2016; Lample et al., 2018).

For our experiments with decoding algorithms, we set the following set of parameters. We set the beam size  $k = 2$ . Vijayakumar et al. (2018) argue that setting  $G = k$  leads to the best results in terms of generation with diverse beam algorithm, therefore, we set  $G = k = 2$  and  $\lambda$  equals 0.5. For top- $k$  sampling, we try multiple values for  $k$ , aiming to investigate the impact of this parameter on generation. Specifically, we generate texts with  $k$  being the value from the following set:  $\{25, 50, 75, 100\}$ . For nucleus sampling, we set  $p$  to one of the following values:  $\{25, 50, 95\}$ . We also run pure sampling with  $k = 100$  and temperature scaling with  $\tau = 0.5$ . Our parameters for different inference algorithms are chosen based on experiences from the corresponding research that introduces these algorithms. They also reflect our goal of evaluating how results generated by different searches can be affected by a single hyperparameter.

## 5 Linking

In the context of the image paragraph generation task, discourse structure in texts is affected by both



Metric	g	b2	s25	s50	s75	s100	st50	n25	n50	n95	db2
BLEU-1	<b>37.16</b>	30.79	33.82	34.57	33.84	33.91	36.48	34.11	34.36	33.61	<i>37.08</i>
BLEU-2	<b>23.90</b>	19.86	18.54	19.20	18.54	18.29	22.20	18.70	19.07	18.46	<i>23.85</i>
BLEU-3	<b>15.53</b>	13.13	10.07	10.77	10.09	9.99	13.67	10.30	10.67	10.25	<i>15.51</i>
BLEU-4	<b>9.54</b>	8.02	4.81	5.40	4.95	5.00	7.89	5.29	5.62	5.15	<i>9.52</i>
METEOR	<b>14.22</b>	12.97	12.53	12.79	12.53	12.46	14.00	12.67	12.80	12.58	<i>14.20</i>
ROUGE-L	<i>30.64</i>	<b>30.71</b>	23.86	23.77	23.56	23.29	28.48	23.15	23.75	23.79	<i>30.55</i>
CIDEr	<i>16.62</i>	12.30	10.48	11.30	9.78	9.54	16.51	10.54	10.76	10.56	<b>16.80</b>
WMD	39.80	39.10	38.40	38.41	38.17	38.06	<b>40.26</b>	38.28	38.34	38.33	<i>39.84</i>

Table 1: Scores of automatic metrics for different inference algorithms. The best scores per metric are in **bold**, while second best scores are in *italics*. The notation for searches should be read as follows throughout the paper: “g” - greedy, “b2” - beam search with the width  $k = 2$ , “sk” - sampling, where  $k$  is the top tokens from which the prediction is sampled, “st50” - sampling from the full probability distribution with temperature scaling  $\tau = 0.5$ , “np” - nucleus sampling with  $p$  denoting the part of the vocabulary with the most probability mass, “db2” - diverse beam search with the width  $k = 2$ .

text and image. To evaluate such structure, we require a mapping between object descriptions and objects in the image. While images in the Tell-me-more corpus were originally annotated with objects as part of the ADE20k corpus of house environments (Zhou et al., 2017), the descriptions were collected separately, hence, there are no annotations between texts and images. We decided to map noun phrases and image objects automatically, using *linking*, which is based on similarity between object labels and noun phrases in texts<sup>5</sup>.

Primarily, linking is performed by taking both attribute and object label from the object detector and merging them into a single string, e.g. “white couch”. Next, spaCy (Honnibal et al., 2020) is used to extract noun phrases from image paragraphs, and we seek to connect each noun phrase with one of the objects in the image  $o_n \in \mathbf{O}$  by embedding them both with a Sentence Transformer (Reimers and Gurevych, 2019) and comparing them based on the cosine similarity with the threshold of 0.5<sup>6</sup>. If there are multiple similarity values that exceed these threshold for a single noun phrase, we map this phrase with the object that has the highest similarity value. Otherwise, if the noun phrase is in plural form, we map multiple objects that also share the same lemma. We perform linking for both reference texts and texts generated by each of the decoding algorithms.

## 6 Automatic evaluation

Table 1 shows scores for the most common metrics in multi-modal automatic evaluation. As we

can see, greedy search and diverse beam perform the best. The worst performance is demonstrated by a variety of sampling algorithms and, somewhat surprisingly, nucleus sampling. Beam performs relatively well, achieving the highest score in ROUGE-L. When looking at the example generations in Table 2, we see that beam search generates very short sentences with fewer mentions of different objects, which definitely has an effect on the performance with n-gram-based metrics. Top- $k$  sampling generally performs worse when the sampling size is increasing: CIDEr score drops to 9.54 with sampling from the full distribution. Interestingly, setting  $k$  to 50 improves the performance, indicating that this value might be the optimal one for this parameter. Nucleus sampling has a very stable performance with n50 showing the best scores. We note that temperature scaling has a huge positive impact on the scores of sampling algorithm, pushing it towards the performance of greedy and diverse beam search. This might indicate that sampling and its randomness can be successfully controlled with the proper value for temperature.

The reason for a high performance of greedy search could be its ability to generate the “safest” words combined with the simplicity of the images and the lack of surprisal in them. For example, images in the dataset correspond to standard room types and thus contain standard objects. This is a blessing if we care about model’s ability to generalise over the house environments, but also a curse since occasionally a highly salient and surprising object might appear in the images and searches will not be able to describe it. We believe that our results show the inadequacy of automatic metrics in measuring the sensitivity of inference algorithms

<sup>5</sup>We use the linking from Dobnik et al. (2022).

<sup>6</sup>Different methods of linking noun phrases and object labels have been evaluated in Ilinykh and Dobnik (2022).



<b>ref</b>	there is a very large well made bed on the left. this looks like a hotel room. two lamps are on a table on the right side. a window that you can barely see out of can be seen if you look straight ahead. there is a picture on the wall to the left and another on the right side wall.
<b>g</b>	this is a living room with a white couch and love seat. there is a white couch with a white and white pillow. there is a coffee table in front of the couch. there is a glass coffee table in the middle of the room. there is a picture hanging on the wall.
<b>b2</b>	this is a living room. there is a white couch. there is a white coffee table. there is a couch.
<b>s25</b>	this is a tv in UNK. the couches match the couches and bedspread is white. there are two green UNK and there is a sofa table with pillows. there are two different decorative beige feet couch.
<b>s50</b>	a large room with a sofa for a coffee table. a four drawer rectangular coffee table sits full of glass roses loveseat. a fireplace is on the left hand wall. a large telephone is near the couch. there are two brown chairs sitting at the far end of the couch.
<b>s75</b>	its a formal living room. a glass sofa and ottoman. a glass table is behind the couch. the tv is behind the sofa. a patterned sofa is on the right side of the one sofa.
<b>s100</b>	living room with a black couch with white print pillows underneath. coffee table piece of 2 candles hang on the back wall. green coffe coffee table lobby is sitting in the middle of the couch by chair and books. decorative and an condition painted on the floor in color. light brown seat at the back end of the room.
<b>st50</b>	this is a living room. the couches are white. there is a white couch. there is a coffee table. there is a white couch.
<b>n25</b>	this room is very neat and space. the couches have orange felt UNK around the mantel. on the right side of the picture contains red flowers. there is a picture on the wall facing the sofa. there is a sofa on the wall.
<b>n50</b>	its a living room with a grey couch and upholstered sofa. there is a tv on the right side next to the couch. two dining chairs UNK facing the couch. there is a coffee table in front of the couch and chair. there is a glass coffee table cloth and robe hanging in the middle.
<b>n95</b>	there is a stone seat in the photo. the sofa is white with UNK upholstery. a beige chair and orange chair chair a round coffee table but its not sailboat. the couches fabric cover match the white accent pillows with a picture on the wall alongside them as the black chair and tan carpet.
<b>db2</b>	this is a living room with a white couch and love seat. there is a white couch with a white and white pillow. there is a coffee table in front of the couch. there is a glass coffee table in the middle of the room. there is a picture hanging on the wall.

Table 2: Example of the image and paragraphs generated with different inference algorithms.

		<b>g</b>			<b>b2</b>			<b>s50</b>			<b>st50</b>			<b>n50</b>			<b>db2</b>		
		<i>P</i>	<i>S</i>	<i>K</i>	<i>P</i>	<i>S</i>	<i>K</i>	<i>P</i>	<i>S</i>	<i>K</i>	<i>P</i>	<i>S</i>	<i>K</i>	<i>P</i>	<i>S</i>	<i>K</i>	<i>P</i>	<i>S</i>	<i>K</i>
<b>R</b>	BLEU_1	0.23	0.18	0.13	0.3	0.28	0.22	-0.01	-0.06	-0.03	-0.06	-0.03	-0.02	0.25	0.21	0.15	0.27	0.19	0.15
	BLEU_2	0.21	0.17	0.12	<b>0.34</b>	0.28	0.2	-0.04	-0.16	-0.1	-0.13	-0.15	-0.11	0.14	0.1	0.06	0.3	0.19	0.14
	BLEU_3	0.14	0.16	0.1	0.29	0.22	0.17	-0.05	-0.12	-0.07	-0.15	-0.2	-0.14	0.11	0.1	0.07	0.27	0.21	0.16
	BLEU_4	0.01	0.1	0.07	0.26	0.24	0.18	0.04	-0.1	-0.04	-0.12	-0.16	-0.12	0.19	0.11	0.08	0.2	0.22	0.16
	METEOR	-0.21	-0.18	-0.13	0.14	0.12	0.09	-0.21	-0.22	-0.16	-0.22	<b>-0.32</b>	-0.22	-0.05	-0.09	-0.06	-0.26	-0.24	-0.19
	ROUGE_L	0.18	0.15	0.11	0.22	0.23	0.16	0.06	0.02	0.02	-0.19	-0.22	-0.15	0.19	0.16	0.11	0.28	0.21	0.15
	CIDER	0.02	0.15	0.1	<b>0.33</b>	0.17	0.12	-0.06	-0.17	-0.1	-0.15	-0.18	-0.11	0.23	0.23	0.17	0.16	0.19	0.14
<b>C</b>	WMD	-0.0	0.0	-0.0	0.2	0.16	0.1	-0.14	-0.09	-0.06	-0.12	-0.14	-0.1	-0.09	-0.09	-0.06	-0.14	-0.12	-0.09
	BLEU_1	0.14	0.13	0.09	0.11	0.12	0.09	0.02	0.01	0.0	0.06	0.11	0.07	0.22	0.19	0.13	0.19	0.18	0.12
	BLEU_2	0.12	0.08	0.06	0.15	0.15	0.12	-0.05	-0.12	-0.09	0.09	0.13	0.09	0.13	0.1	0.06	0.21	0.18	0.12
	BLEU_3	0.02	0.05	0.03	0.09	0.11	0.08	-0.09	-0.12	-0.09	0.11	0.13	0.08	0.12	0.1	0.06	0.18	0.18	0.13
	BLEU_4	-0.12	-0.02	-0.03	0.02	0.09	0.06	-0.0	-0.13	-0.1	0.1	0.14	0.09	0.22	0.15	0.11	0.17	0.22	0.16
	METEOR	-0.15	-0.13	-0.08	0.09	0.08	0.06	-0.19	-0.17	-0.13	-0.09	-0.1	-0.07	-0.16	-0.24	-0.16	-0.27	-0.27	-0.2
	ROUGE_L	0.13	0.19	0.14	0.06	0.08	0.05	-0.07	-0.09	-0.07	-0.02	0.02	0.02	0.22	0.19	0.14	0.16	0.17	0.11
<b>F</b>	CIDER	0.03	0.12	0.09	0.14	0.1	0.05	-0.0	-0.01	-0.01	-0.07	0.09	0.08	0.22	0.26	0.17	0.12	0.21	0.16
	WMD	-0.02	-0.03	-0.02	0.16	0.13	0.1	-0.22	-0.17	-0.12	-0.09	-0.07	-0.05	-0.22	-0.28	-0.21	-0.1	-0.09	-0.08
	BLEU_1	<b>0.41</b>	<b>0.37</b>	<b>0.27</b>	<b>0.42</b>	<b>0.4</b>	<b>0.31</b>	-0.22	-0.24	-0.19	0.01	0.0	0.01	0.13	0.08	0.06	<b>0.32</b>	<b>0.32</b>	<b>0.24</b>
	BLEU_2	<b>0.39</b>	<b>0.36</b>	<b>0.28</b>	<b>0.38</b>	<b>0.29</b>	<b>0.23</b>	-0.18	-0.27	-0.21	-0.01	-0.04	-0.03	0.07	0.05	0.03	<b>0.32</b>	<b>0.31</b>	0.22
	BLEU_3	0.29	<b>0.32</b>	<b>0.23</b>	<b>0.35</b>	0.25	0.19	-0.22	-0.25	-0.18	0.01	-0.0	0.0	0.12	0.07	0.05	0.3	0.3	0.22
	BLEU_4	0.15	0.24	0.18	0.23	0.2	0.14	-0.01	-0.17	-0.12	0.03	0.05	0.03	0.19	0.06	0.04	0.22	0.24	0.17
	METEOR	-0.07	-0.07	-0.08	0.12	0.09	0.06	-0.11	-0.14	-0.1	-0.0	-0.06	-0.03	-0.12	-0.2	-0.16	-0.01	-0.01	-0.01
<b>F</b>	ROUGE_L	<b>0.31</b>	0.29	0.22	0.24	0.24	0.18	-0.06	-0.1	-0.08	-0.08	-0.07	-0.05	0.16	0.12	0.09	0.28	0.29	0.19
	CIDER	0.16	0.27	0.2	<b>0.36</b>	0.27	0.21	<b>-0.35</b>	<b>-0.37</b>	<b>-0.28</b>	0.01	0.02	0.02	0.02	0.1	0.07	0.13	0.29	0.23
	WMD	0.04	0.03	0.02	0.19	0.22	0.14	-0.14	-0.16	-0.12	-0.03	-0.02	-0.03	-0.02	-0.03	-0.03	0.1	0.05	0.03

Table 3: Correlation scores between automatic metrics and human judgements across three criteria. **R**, **C** and **F** on the left side stand for relevance, correctness and composition (flow), corresponding to the type of questions that the crowdworkers were provided with. *P*, *S* and *K* stand for Pearson’s, Spearman’s and Kendall’s correlations. We report correlation scores per search and per correlation metric. The scores coloured in red have  $p < 0.05$ .

to the type of objects and their salience.

## 7 Human evaluation

To support our hypothesis that automatic metrics are not enough to measure fine-grained differences between various decoding algorithms, we conduct a human evaluation on Amazon Mechanical Turk. We randomly sample 10% of images from the test set, which equals 41 items. For each of these images, we take generated texts from the top-6 decodings based on the CIDER score. We get 287 different image-text pairs to evaluate. During the evaluation, we provide workers with an image and its description and ask them to answer 3 (three) dif-

ferent questions, aiming to evaluate (i) relevance: does the text describe relevant and essential objects, (ii) correctness: does the text describe objects correctly (e.g., using correct words), (iii) composition: do object descriptions naturally follow each other. The example item for human evaluation is shown in Appendix A. Each judgement is a score on a scale between 1 and 5, where 1 is the lowest rank. We collect three different judgements per item and average them. We pay 0.17 US dollars for a single assignment and restrict the location of the workers to the US, the UK, Canada, Ireland or Australia. We also ran our experiments with Master workers only (25 different human participants). We follow

	ref	g	b2	s25	s50	s75	s100	st50	n25	n50	n95	db2
<b>s1</b>	2.9	0.9	0.3	1.0	1.1	1.1	1.0	0.9	1.1	1.1	1.1	0.9
<b>s2</b>	1.6	1.7	1.5	1.7	1.7	1.8	1.8	1.7	1.8	1.8	1.7	1.8
<b>s3</b>	1.4	1.6	1.4	1.7	1.8	1.8	1.8	1.6	1.8	1.8	1.8	1.6
<b>s4</b>	1.3	1.6	1.4	1.8	1.8	1.8	1.9	1.6	1.8	1.9	1.8	1.6
<b>s5</b>	1.2	1.7	1.4	1.8	1.8	1.8	1.7	1.6	1.7	1.8	1.7	1.7

Table 4: Average number of noun phrases generated by different inference algorithms. The numbers are given per sentence.

Kilickaya et al. (2017) and compute three different correlation scores: Pearson’s correlation, Spearman’s rank correlation and Kendall’s correlation.

The correlation scores are presented in Table 3. In general, sampling-based methods do not significantly correlate with automatic metrics or correlate but negatively. More controlled decodings, such as greedy or beam search, correlate with automatic metrics more, especially for the composition question (F). This indicates that automatic metrics correlate more with decodings that introduce less randomness. Future work will need to examine whether randomness and diversity in such searches as top- $k$  sampling is a suitable type of diversity since it is unclear from correlation scores alone. In terms of the relevance of objects, sampling with temperature generally has negative scores (similar to other sampling-based methods). Still, a significant negative correlation is found only with Spearman’s rank correlation for METEOR. Beam, however, might produce more relevant objects as demonstrated by high correlation in terms of BLEU\_2 and CIDEr. We do not observe any correlation for the correctness criterion. On the contrary, text composition (flow) shows that more controlled decodings correlate considerably more with human judgements, especially when looking at n-gram metrics. This might demonstrate that more specific automatic metrics better reflect whether the object descriptions naturally follow each other. Overall, we show that while most of the automatic metrics are not sufficient in providing us with information about the salience and correctness of object descriptions for many different decoding algorithms, their scores, somewhat surprisingly, might still tell us about the sentence-level discourse and flow of object descriptions.

## 8 Non-grounded evaluation

Next, we will look at the surface level of noun phrases and examine faithfulness of generated texts to the reference ones. Noun phrases in image de-

	g	b2	s25	s50	s75	s100	st50	n25	n50	n95	db2
<b>s1</b>	200.0	175.0	227.9	215.0	233.3	231.0	208.0	227.9	209.2	228.2	200.0
<b>s2</b>	205.2	215.8	207.6	215.8	218.1	220.1	207.5	228.2	231.1	206.1	206.5
<b>s3</b>	210.8	205.2	213.8	215.0	233.3	203.8	210.3	202.6	219.3	207.1	207.2
<b>s4</b>	197.4	196.0	216.4	206.5	200.0	208.9	208.9	205.9	201.3	216.4	197.5
<b>s5</b>	198.0	212.5	202.6	205.1	211.3	214.4	197.3	209.6	215.8	208.7	200.0

Table 5: Average proportion of noun phrases (in percent) when *more* are generated than present in the references.

	g	b2	s25	s50	s75	s100	st50	n25	n50	n95	db2
<b>s1</b>	20.3	8.5	18.8	18.9	19.6	18.6	18.9	19.6	18.8	20.7	20.3
<b>s2</b>	45.3	44.2	42.6	42.9	40.4	43.1	45.9	39.8	40.0	39.7	45.5
<b>s3</b>	45.9	45.6	42.0	45.9	44.4	44.5	45.5	42.7	41.2	44.4	47.1
<b>s4</b>	46.5	46.8	43.9	45.8	43.3	43.4	47.0	45.0	43.3	42.9	46.6
<b>s5</b>	49.3	43.7	41.4	42.5	37.6	40.7	46.2	38.3	40.0	39.9	49.0

Table 6: Average proportion of noun phrases (in percent) when *fewer* are generated than present in the references.

scriptions typically depict image objects, thus we believe that direct comparison of noun phrases in different texts can help us to understand how much each decoding algorithm learns on the surface of descriptions. Table 4 shows the average number of noun phrases in each sentence across different searches and references. We see that there is a gradual decrease in the number of noun phrases in references throughout the paragraph. Such decrease is not observed in texts generated by all algorithms. On the contrary, the first sentence typically has the fewest number of noun phrases generated with other sentences containing mostly the same number. This could be a sign that on the surface level decoding algorithms do not capture discourse structure, reflected in gradual decrease of the number of noun phrases. Instead, search algorithms tend to generate the same number of noun phrases across sentences, treating each sentence equally.

We also observe that the algorithms generate more noun phrases per sentence than required rather than generate fewer of them. Specifically, across all image-paragraph pairs a fewer number of noun phrases is generated for 757 sentences, a bigger number for 955 sentences and the exact number as in the references was produced for 493 sentences. To closer identify the impact of over- and under-generation of noun phrases, we compute proportion of noun phrases for both cases. As Table 5 demonstrates, all searches tend to generate nearly two times more noun phrases than required in each sentence. The picture changes when the searches under-generate. According to Table 6, while most of the sentences lack at least half of the required noun phrases (in terms of quantity), the first sentence is affected the most by under-

	g	b2	s25	s50	s75	s100	st50	n25	n50	n95	db2
<b>s1</b>	0.18	0.10	0.10	0.13	0.11	0.08	0.14	0.12	0.10	0.13	0.18
<b>s2</b>	0.17	0.17	0.13	0.13	0.13	0.14	0.17	0.12	0.15	0.13	0.17
<b>s3</b>	0.13	0.12	0.10	0.10	0.09	0.11	0.11	0.10	0.11	0.10	0.13
<b>s4</b>	0.10	0.09	0.10	0.09	0.08	0.09	0.09	0.09	0.09	0.09	0.10
<b>s5</b>	0.10	0.10	0.07	0.07	0.08	0.07	0.08	0.07	0.07	0.08	0.10

Table 7: Dice similarity coefficient between the set of objects described in reference texts and texts generated by different decoding algorithms. The values are provided per sentence and averaged across all image-paragraph pairs.

generation. Coupled with the results in Table 4, we conclude that decoding algorithms do not learn the structure of discourse on the simplest surface level of descriptions reflected in the differences in the number of noun phrases. This result indicates that searches might generate a discourse that is different from the one observed in references. In the following analysis, we will move from the surface level to the grounding level, in which we will examine if the noun phrases that are generated can be linked with image objects. We will also compare whether the objects described by different searches overlap with the ones found in reference texts.

## 9 Grounded evaluation

Table 7 shows the degree of overlap between two object sets: the first set includes objects described in references, while the second set contains objects mapped with noun phrases in generated texts from different decodings algorithms. We use Sørensen–Dice coefficient  $\frac{2|A \cap B|}{|A| + |B|}$  to measure the overlap. The closer the result to 0, the less overlap is present. The results demonstrate that searches describe a very different set of objects rather than the one mentioned in the references. The highest overlap is observed with greedy search and diverse beam. The scores indicate that either a different and correct set of objects is described or the noun phrases cannot be linked with objects because they are incorrect (could also be because of high randomness, leading to the lack of grammaticality).

We examine whether noun phrases in generated texts can be linked with any of the objects in the image. Table 8 shows the proportion of successful linking once we link noun phrases with image objects using cosine similarity. We set the similarity threshold to 0.5: if the similarity between the object label and noun phrase is higher than this value, we decide that this noun phrase is faithful to the image and can be grounded.

	g	b2	s25	s50	s75	s100	st50	n25	n50	n95	db2
<b>s1</b>	69.5	72.7	46.5	46.3	43.2	46.1	66.5	51.1	45.2	50.5	69.5
<b>s2</b>	65.6	65.1	47.2	49.0	43.8	46.6	58.8	44.7	50.3	47.7	65.5
<b>s3</b>	61.6	59.5	43.6	46.9	40.5	45.1	53.1	40.1	40.6	44.7	60.7
<b>s4</b>	55.4	57.6	43.7	42.7	44.7	41.0	52.5	45.0	43.7	38.3	55.7
<b>s5</b>	60.5	57.4	47.6	43.2	43.4	43.7	53.3	39.2	38.9	44.4	59.3

Table 8: Average proportion of successful linking (in percent) between noun phrases in generated texts and image objects.

The results demonstrate that half and more of the generated noun phrases can be linked with objects in the image. In general, sampling algorithms generate fewer number of grounded noun phrases, possibly due to the increased randomness. Greedy search, beam and diverse beam generate the highest number of noun phrases which are truthful to the image. We believe that while reference-correctness of generated texts can get worse, inference algorithms are still able to generate alternative descriptions of images which can be grounded. However, the structure of discourse reflected on the surface level and the level of grounding might not necessarily correspond to the one observed in references. In the next experiment, we look at the problem under the angle of attentional structure and examine spatial arrangement of linked objects and how these arrangements differ between decoding algorithms.

## 10 Attentional structure of discourse

Figure 1 demonstrates a number of the attention heatmaps across areas in the image for different sentences. At first glance, different inference algorithms look at similar locations in the image and also focus on parts which are attended by humans. However, there are relatively more areas described in the first sentence of the references, while a much smaller and fewer areas are described in generated texts. This could be directly related to the fewer number of objects and under-generation discussed previously. The second and third sentences describe specific areas of the image in all cases, mostly central ones. Interestingly, greedy and diverse beam have highly similar attention across the image. In sentence 4, human attention disperses over the full scene, while it is unclear whether the same pattern happens in generated texts. This could signal a possible *topic shift*, happening in later parts of the paragraph and inability of searches to capture that. To understand the differences on the level of sentences better, we measure the correlation between flattened heatmaps pixel by pixel.



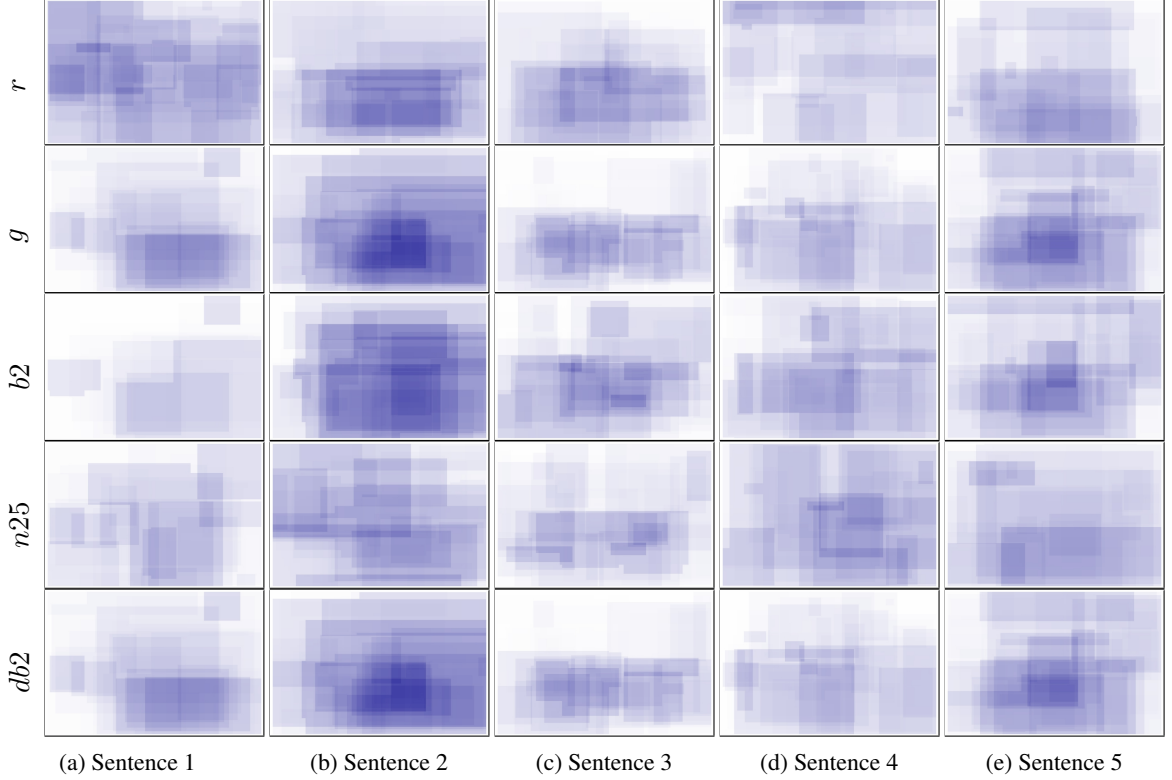


Figure 1: Attention heatmaps over objects, described in texts according to the results of linking. Results are shown per sentence and per search. The first row denotes attention in reference texts. We aggregate heatmaps across all images into the single image, therefore, darker colour denotes higher focus on the specific area in the image.

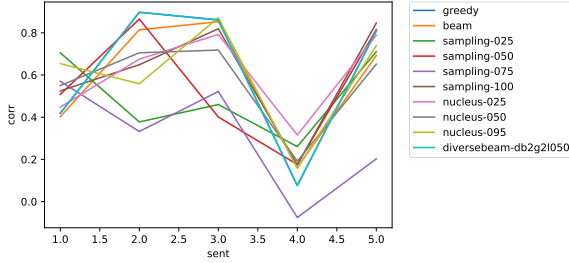


Figure 2: Correlation between heatmaps for different searches and reference paragraphs. X-axis is sentence in the paragraph (1-5), Y-axis is the correlation coefficient, pixel-by-pixel correlation between attention heatmaps.

We use Pearson product-moment correlation coefficient which can be applied to images across the channels. The results are shown in Figure 2. As we can see, in sentence 4 attentional structure on the image differs between searches and references, supporting the idea of topic shift. Sampling methods have the lowest correlation with the references, while nucleus with  $p = 25$  is affected the least in sentence 4. Note that the correlation in the first sentence is lower than in the second and the third one for most of the searches. This could be related to the importance of the first sentence and a big-

ger number of noun phrases in it, which are not generated during the decoding stage.

## 11 Conclusion

In this paper we described our analysis of how decoding strategies structure discourse in multi-modal longer image descriptions. We performed evaluation using intuitions from different evaluation perspectives: automatic, surface-based (non-grounded), image-based (grounded) and attention-based. The results suggest that for the task of image paragraph generation decoding algorithms diverge from humans in generating specific type of discourse. Although they might generate reference-incorrect but image-correct descriptions, it is unclear what kind of discourse is generated in the end. In general, algorithms which are less random construct discourse similar to the one in human references, while sampling-based methods generate a different type of discourse, which is hard to control for. We plan to use the insights described in this paper and build a metric that would evaluate the structure of longer image paragraphs, reflected in *both* object and relation descriptions as this is currently a much needed evaluation measure.

## 12 Limitations

There are several directions which can support the analysis in this paper. First, the automatic linking is not a perfect mechanism, prone to errors. The method that we use works better for shorter phrases which share the same lemmas and thus are less ambiguous. Second, using more models (Li et al., 2019) or more datasets (Krause et al., 2017) would potentially give us a broader picture of the type of discourses formed by humans and quality of representations used during decoding phase. We also consider our analysis preliminary with the opportunity of developing a separate metric to evaluate discourse in longer image descriptions.

## Acknowledgements

The research reported in this paper was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

## References

- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. [Bottom-up and top-down attention for image captioning and visual question answering](#). In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.
- Anusha Balakrishnan, Jinfeng Rao, Kartikeya Upasani, Michael White, and Rajen Subba. 2019. [Constrained decoding for neural NLG from compositional representations in task-oriented dialogue](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 831–844, Florence, Italy. Association for Computational Linguistics.
- Massimo Caccia, Lucas Caccia, William Fedus, Hugo Larochelle, Joelle Pineau, and Laurent Charlin. 2020. [Language gans falling short](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Yun Chen, Victor O.K. Li, Kyunghyun Cho, and Samuel Bowman. 2018. [A stable and effective learning strategy for trainable greedy decoding](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 380–390, Brussels, Belgium. Association for Computational Linguistics.
- Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 2018. [Pragmatically informative image captioning with character-level inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 439–443, New Orleans, Louisiana. Association for Computational Linguistics.
- Robert Dale and Michael White. 2007. [Shared tasks and comparative evaluation in natural language generation](#). In *Proceedings of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation*.
- Simon Dobnik, Nikolai Ilinykh, and Aram Karimi. 2022. [What to refer to and when? reference and re-reference in two language-and-vision tasks](#). In *Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, pages 146–159, Dublin, Ireland. SEMDIAL.
- Desmond Elliott and Frank Keller. 2014. [Comparing automatic evaluation measures for image description](#). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 452–457, Baltimore, Maryland. Association for Computational Linguistics.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. [Hierarchical neural story generation](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.
- Adam Fisch, Kenton Lee, Ming-Wei Chang, Jonathan Clark, and Regina Barzilay. 2020. [CapWAP: Image captioning with a purpose](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8755–8768, Online. Association for Computational Linguistics.
- Albert Gatt and E.J. Krahmer. 2018. [Survey of the state of the art in natural language generation: Core tasks, applications and evaluation](#). *Journal of Artificial Intelligence Research*, 61(1):65–170.
- Edward Albert Fletcher Gibson. 1991. *A Computational Theory of Human Linguistic Processing: Memory Limitations and Processing Breakdown*. Ph.D. thesis, Carnegie Mellon University, USA. UMI Order No. GAX91-26944.
- Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. [Centering: A framework for modeling the local coherence of discourse](#). *Computational Linguistics*, 21(2):203–225.
- Jiatao Gu, Kyunghyun Cho, and Victor O.K. Li. 2017. [Trainable greedy decoding for neural machine translation](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1968–1978, Copenhagen, Denmark. Association for Computational Linguistics.
- Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. [Image captioning: Transforming objects into words](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. [CLIPScore: A reference-free evaluation metric for image captioning](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. [Learning distributed representations of sentences from unlabelled data](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377, San Diego, California. Association for Computational Linguistics.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. [The curious case of neural text degeneration](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. [spaCy: Industrial-strength Natural Language Processing in Python](#).
- Jack Hopkins and Douwe Kiela. 2017. [Automatically generating rhythmic verse with neural networks](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 168–178, Vancouver, Canada. Association for Computational Linguistics.
- David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. [Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions](#). In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.
- Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. [Visual storytelling](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, San Diego, California. Association for Computational Linguistics.
- Nikolai Ilinykh and Simon Dobnik. 2020. [When an image tells a story: The role of visual and semantic information for generating paragraph descriptions](#). In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 338–348, Dublin, Ireland. Association for Computational Linguistics.
- Nikolai Ilinykh and Simon Dobnik. 2022. [Attention as grounding: Exploring textual and cross-modal attention on entities and relations in language-and-vision transformer](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4062–4073, Dublin, Ireland. Association for Computational Linguistics.
- Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. [Tell me more: A dataset of visual scene description sequences](#). In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 152–157, Tokyo, Japan. Association for Computational Linguistics.
- Mert Inan, Piyush Sharma, Baber Khalid, Radu Soricut, Matthew Stone, and Malihe Alikhani. 2021. [COSMic: A coherence-aware generation metric for image descriptions](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3419–3430, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. 2019. [TIGer: Text-to-image grounding for image caption evaluation](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2141–2152, Hong Kong, China. Association for Computational Linguistics.
- Chloé Kiddon, Luke Zettlemoyer, and Yejin Choi. 2016. [Globally coherent text generation with neural checklist models](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 329–339, Austin, Texas. Association for Computational Linguistics.
- Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2017. [Re-evaluating automatic metrics for image captioning](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 199–209, Valencia, Spain. Association for Computational Linguistics.
- Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. 2014. [What are you talking about? text-to-image coreference](#). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3558–3565.
- Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. [A hierarchical approach for generating descriptive image paragraphs](#). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 317–325.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. [Visual genome: Connecting language and vision using crowdsourced dense image annotations](#).




- Ilya Kulikov, Alexander Miller, Kyunghyun Cho, and Jason Weston. 2019. [Importance of search and evaluation strategies in neural dialogue modeling](#). In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 76–87, Tokyo, Japan. Association for Computational Linguistics.
- Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2011. [Baby talk: Understanding and generating simple image descriptions](#). In *CVPR 2011*, pages 1601–1608.
- Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. [From word embeddings to document distances](#). In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France. PMLR.
- Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2018. [Unsupervised machine translation using monolingual corpora only](#). In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.
- Alon Lavie and Abhaya Agarwal. 2007. [METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments](#). In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.
- Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. [Multi-agent cooperation and the emergence of \(natural\) language](#). In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.
- Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. 2019. [Entangled transformer for image captioning](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8928–8937.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. [A diversity-promoting objective function for neural conversation models](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Charlotte Linde and J.A. Goguen. 1980. [On the independence of discourse structure and semantic domain](#). In *18th Annual Meeting of the Association for Computational Linguistics*, pages 35–37, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Pranava Madhyastha, Josiah Wang, and Lucia Specia. 2019. [VIFIDEL: Evaluating the visual fidelity of image descriptions](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6539–6550, Florence, Italy. Association for Computational Linguistics.
- Luca Massarelli, Fabio Petroni, Aleksandra Piktus, Mylène Ott, Tim Rocktäschel, Vassilis Plachouras, Fabrizio Silvestri, and Sebastian Riedel. 2020. [How decoding strategies affect the verifiability of generated text](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 223–235, Online. Association for Computational Linguistics.
- Clara Meister, Martina Forster, and Ryan Cotterell. 2021. [Determinantal beam search](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6551–6562, Online. Association for Computational Linguistics.
- Shashi Narayan, Gonçalo Simões, Yao Zhao, Joshua Maynez, Dipanjan Das, Michael Collins, and Mirella Lapata. 2022. [A well-composed text is half done! composition sampling for diverse conditional generation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1319–1339, Dublin, Ireland. Association for Computational Linguistics.
- Taichi Nishimura, Atsushi Hashimoto, and Shinsuke Mori. 2019. [Procedural text generation from a photo sequence](#). In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 409–414, Tokyo, Japan. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Massimo Poesio. 2004. [Discourse annotation and semantic annotation in the GNOME corpus](#). In *Proceedings of the Workshop on Discourse Annotation*, pages 72–79, Barcelona, Spain. Association for Computational Linguistics.
- Massimo Poesio, Rosemary Stevenson, Barbara Di Eugenio, and Janet Hitzeman. 2004. [Centering: A parametric theory and its instantiations](#). *Computational Linguistics*, 30(3):309–363.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*



- and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Ehud Reiter and Anja Belz. 2009. [An investigation into the validity of some metrics for automatically evaluating natural language generation systems](#). *Computational Linguistics*, 35(4):529–558.
- Ehud Reiter and Robert Dale. 1997. [Building applied natural language generation systems](#). *Natural Language Engineering*, 3(1):57–87.
- Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. [Self-critical sequence training for image captioning](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1179–1195.
- Claude Elwood Shannon. 1948. [A mathematical theory of communication](#). *The Bell System Technical Journal*, 27:379–423.
- Ece Takmaz, Mario Giulianelli, Sandro Pezzelle, Arabella Sinclair, and Raquel Fernández. 2020. [Refer, Reuse, Reduce: Generating Subsequent References in Visual and Conversational Contexts](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4350–4368, Online. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.
- Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. [Context-aware captions from context-agnostic supervision](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1070–1079.
- Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. [Cider: Consensus-based image description evaluation](#). In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.
- Ashwin Vijayakumar, Michael Cogswell, Ramprasaath Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. [Diverse beam search for improved description of complex scenes](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).
- Sijin Wang, Ziwei Yao, Ruiping Wang, Zhongqin Wu, and Xilin Chen. 2021. [Faier: Fidelity and adequacy ensured image caption evaluation](#). In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14045–14054.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. [Google’s neural machine translation system: Bridging the gap between human and machine translation](#). *CoRR*, abs/1609.08144.
- Sina Zarrieß and David Schlangen. 2018. [Decoding strategies for neural referring expression generation](#). In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 503–512, Tilburg University, The Netherlands. Association for Computational Linguistics.
- Sina Zarrieß, Henrik Voigt, and Simeon Schüz. 2021. [Decoding methods in neural language generation: A survey](#). *Information*, 12(9).
- Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. 2021. [Trading off diversity and quality in natural language generation](#). In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 25–33, Online. Association for Computational Linguistics.
- Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. [Scene parsing through ade20k dataset](#). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

## A Appendix A

First, read the instructions ("Instructions" in the top-left corner).  
 Have a look at both short and detailed instructions.  
 If you don't follow them, we have a right to reject your submission.



this is a living room with tan walls and white ceiling . there is a large black couch and brick wall . there is a golden hutch in the front part of the couch . there is a brown night table in the back room with a white chair at the center of the room . there are plants on the ceiling by a chandelier hanging above the couch .

How well do you agree with the following statements?

Relevance: does the text describe relevant and important objects?

Correctness: does the text describe objects correctly (e.g., using correct words)?

Composition: do objects descriptions naturally follow each other?

Figure 3: The example item for the workers on AMT for human evaluation.