CIC-BART-SSA: Controllable Image Captioning with Structured Semantic Augmentation

Kalliopi Basioti[1,2*], Mohamed A. Abdelsalam[2], Federico Fancellu[3†], Vladimir Pavlovic[1†], and Afsaneh Fazly[2]

[1] Rutgers University, New Jersey, USA {kalliopi.basioti, vladimir}@rutgers.edu
[2] Samsung AI Centre - Toronto, Toronto, Canada {m.abdelsalam, a.fazly}@samsung.com
[3] Solventum ffancellu@solventum.com

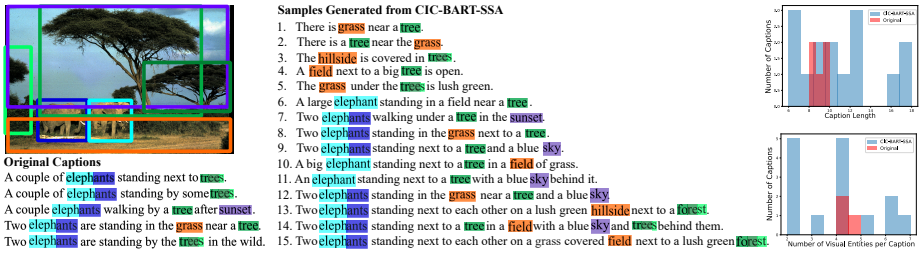

Fig. 1: Existing captioning datasets contain captions that describe the entirety of an image. This is reflected in the narrow distributions of the entities that appear in those captions and the caption lengths (the red-colored histograms). CIC aims to generate diverse descriptions by controllably re-focusing on different spatiosemantic aspects of an image, such as the semantically coherent subsets of image objects. Our proposed CIC-BART-SSA is designed to produce diverse, controlled captions ranging from brief and concise to detailed and comprehensive. Sentences 1-15 are example outputs of our approach where the highlighted text indicates the focus of a controllable caption. The histograms demonstrate that our approach generates high-quality descriptions for a wider range of scene focus (number of visual entities) and caption length compared to the original captions. Image is licensed under CC BY-SA 2.0.

Abstract. Controllable Image Captioning (CIC) aims at generating natural language descriptions for an image, conditioned on information provided by end users, e.g., regions, entities or events of interest. However, available image–language datasets mainly contain captions that describe the entirety of an image, making them ineffective for training CIC models that can potentially attend to any subset of regions or relationships. To tackle this challenge, we propose a novel, fully automatic method to sample additional focused and visually grounded captions using a unified structured semantic representation built on top of the existing set of captions associated with an image. We leverage Abstract Meaning Representation (AMR), a cross-lingual graph-based semantic formal-

*Work done during an internship at Samsung AI Centre - Toronto
†Work done while at Samsung AI Centre - Toronto

ism, to encode all possible spatio-semantic relations between entities, beyond the typical spatial-relations-only focus of current methods. We use this Structured Semantic Augmentation (SSA) framework to augment existing image–caption datasets with the grounded controlled captions, increasing their spatial and semantic diversity and focal coverage. We then develop a new model, CIC-BART-SSA, specifically tailored for the CIC task, that sources its control signals from SSA-diversified datasets. We empirically show that, compared to SOTA CIC models, CIC-BART-SSA generates captions that are superior in diversity and text quality, are competitive in controllability, and, importantly, minimize the gap between broad and highly focused controlled captioning performance by efficiently generalizing to the challenging highly focused scenarios. Code is available at `https://github.com/SamsungLabs/CIC-BART-SSA`.

1 Introduction

Image captioning refers to the task of providing an AI system with an input image, and asking the system to describe the visual content in natural language. This process requires the captioning system to understand what objects are present, in what context (e.g., event or scene), and how they relate. Recent deep learning approaches to this task [14,26,29,31,37,39,40,49,50,57] surpass human performance in standard image captioning metrics. However, these models tend to generate general captions that describe the entirety of an image, and are often of limited diversity; see Original Captions in Fig. 1.

Controllable image captioning (CIC) overcomes these challenges by generating different descriptions for the same image in a user-controlled fashion. That is, a CIC model receives as input an image paired with a user-specified *control signal* (e.g., entities or regions of interest), and generates a caption conditioned on the control signal. CIC models are thus capable of generating a diverse set of captions by varying the control signal for the same image; see CIC generated captions 1–15 in Fig. 1.

In realistic applications, the easiest way for the user to control the generation of captions is to limit the focus of the desired captions by selecting different entities (objects) using their bounding boxes, as shown in Figs. 1 and 2. Most previous work focuses on such spatial control signals [18, 23, 28, 47, 55, 56]. To improve performance, more recent studies supplement this spatial signal with additional information on the desired length, style, or syntactic and semantic structure of the generated text [12,13], increasing the richness and complexity of control signals. However, for the CIC approach to succeed, the CIC models need to be trained on equally rich datasets that incorporate, explicitly or implicitly, those control signals. Unfortunately, most image captioning datasets today, such as Flickr30k [36] or MS-COCO [18], lack this necessary diversity of controls and corresponding captions.

Our goal is to achieve SOTA performance in CIC without the need for new, increasingly rich, yet also costly, and impractical-to-collect datasets, where human workers would face the burden of having to provide multitudes of control

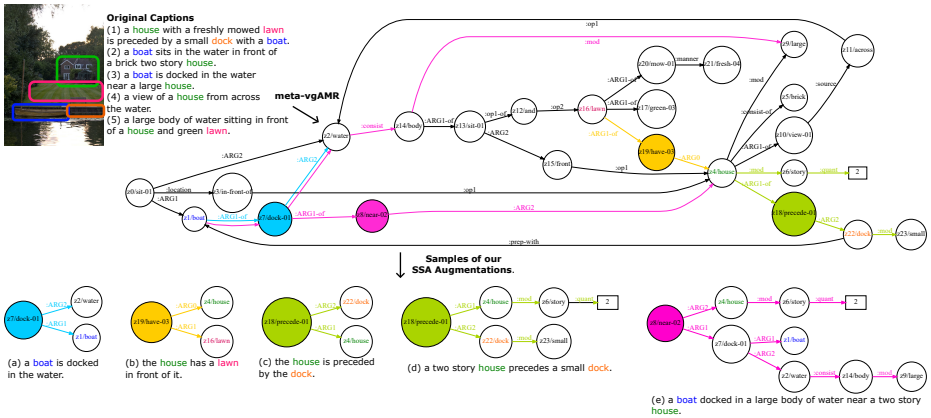

Fig. 2: An example of our Structured Semantic Augmentation (SSA) approach. Visually-grounded captions (1)-(5) are used to create a meta-vgAMR graph, which includes all available image information in one representation. Sub-graphs of meta-vgAMR are then sampled to generate a new and diverse set of captions, such as the sentences (a)-(e). Image is licensed under CC BY-SA 2.0.

signals and corresponding descriptive captions. To achieve this goal, we propose a novel Structured Semantic Augmentation (SSA) method, which automatically generates an augmented set of captions and the corresponding control signals with diverse spatiosemantic focus starting from only the core set of "original" uncontrolled captions. The method takes advantage of a detailed visual-linguistic semantic graph (illustrated in Fig. 2) constructed from the original captions and their image groundings. To build these semantic graphs, we use Abstract Meaning Representation (AMR) [6], a semantic formalism that can capture fine-grained linguistic relations beyond the exclusively spatial relationships present in the common scene graphs [24]. The availability of robust AMR parsers [5,9] allows us to generate semantic graphs for individual captions *automatically*, which we then merge into a rich meta-AMR graph for the joint image–language pair. From this meta-graph, we sample diverse connected subgraphs that represent semantically coherent combinations of image-anchored entities, events, and their relations, which we then turn into controlled captions automatically via existing AMR-to-text models [9]. Fig. 2 depicts an example of our meta-graph inferred from the original uncontrolled captions associated with an image. Filled nodes in the meta-graph indicate image entity groundings. Five semantically coherent subgraphs (a)–(e) of variable complexity are then sampled from the meta-graph, which are subsequently used to generate novel captions, shown below each subgraph. These new captions augment the original caption set by providing both image focus, through node groundings, and increased semantic diversity induced by the sampled subgraphs. Building upon SSA, we introduce a new CIC model, CIC-BART, suitable for generating focused controlled captions. Alongside the regions of interest, CIC-BART also makes use of the length of the desired cap-

tion as a control signal proxy for the verbosity of the caption. CIC-BART can be trained on SSA-augmented versions of standard VL datasets such as MS-COCO or Flicker30k to accommodate the CIC task. Our experiments show that, compared to several SOTA models, the captions generated by our model have superior text quality and diversity, while being comparable in terms of faithfulness to control signals.

In summary, our contributions are:

1. We propose a novel data augmentation technique, SSA, that draws on a structured semantic formalism (AMR) to automatically generate focused captions suitable for training of CIC models. We empirically show that our SSA technique enables CIC models to generate captions with high controllability, diversity, and text quality.

2. We propose CIC-BART, a model designed for CIC, that does not require overly descriptive and complex control signals that SOTA models often require to achieve high performance. We show a superior overall performance, compared to SOTA, while relying on simple control signals (i.e., regions of interest and preferred caption length).

3. We present an extensive evaluation of our model, compared with existing SOTA. Specifically, we report results on different aspects of generated captions, including controllability (faithfulness to control signal), diversity, and text quality (linguistic well-formedness). To account for the trade-off among these metrics, we propose an overall performance score based on their harmonic mean. This metric helps us identify models that perform well in all these aspects.

2 Related Work

Controllable Image Captioning (CIC). Various types of control have been used for CIC, including visual entities, a type of region-based control [18, 23, 28, 47, 55, 56], where generated captions should learn to focus on the regions of interest. Others draw on complex control signals where additional knowledge about the generated caption structure is provided. For example, some recent work provides the complete skeleton of the desired sentence in the form of a number of objects or attributes or object-relation-object templates [12, 13]. Additional control signals that CIC draws on include different caption styles, e.g., positive, negative, humorous, or romantic tone [20, 21, 32, 33, 45, 47, 53, 54], user personality [17, 41], or the length of the generated captions [19, 22, 45, 47, 48]. The use of complex control signals aims at improving the diversity of captions and the quality of the text in CIC models. However, it requires the users to provide a detailed description of the control signal, which is not realistic in practical settings where such models are to be deployed (e.g., a self-driving car or personal assistant). We instead draw on two simple control signals (regions of interest and desired caption length) and show that we can achieve competitive performance on CIC, while keeping the control signals simple and practical.

Recent SOTA models that draw on spatial control include the SCT model [18] that also uses the Faster R-CNN feature vectors and object tags (corresponding

GloVe vectors [35]) of the entities of interest, as well as models that include skeleton-based control, namely ASG2Caption [13] and VSR [12]. ASG2Caption uses an abstract scene graph (ASG) to express the desired structure of a caption. ASG contains three types of unlabeled *abstract nodes* (object, attribute, relationship) that are grounded in the image by extracting features from the corresponding bounding boxes (for objects and attributes) or from the union of bounding box pairs (for a relationship node). ASG2Caption shows improved controllability (by conditioning on ASGs), and diversity (by automatically sampling diverse ASGs as control signals). The VSR model [12] draws on GloVe embeddings of Faster R-CNN object tags for visual entities (as in SCT). It also uses a skeleton control signal (like ASG2Caption), but one that includes more detailed information and richer semantics. Specifically, the VSR control signal follows the form of a fine-grained PropBank entry[4] — i.e., specifying the exact verb(s) expressing action(s) depicted in the image, and their visually grounded arguments (e.g., subject, object, location, manner). Thus, VSR uses the most descriptive control signal among the SOTA models. Refer to our supplementary material for an illustrative example of the control signal used for each method.

Compared to ASG2Caption and VSR, our control signal is kept minimal and only specifies the bounding boxes and desired caption lengths. To improve the diversity of captions, we draw on a structured semantic graph (AMR) that expresses the semantics of a sentence based on PropBank semantics. Notably, we do **not** use these rich graphs to express detailed and overly descriptive control signals (as in VSR), but we use these semantic structures to augment our training data with richer and diverse captions, which will result in the model learning to generate more diverse captions. Additionally, we include a length control signal to further increase diversity without needing to specify detailed information about the structure of the output (e.g., number of attributes per object, etc.). This way, we can generate a variety of captions for a fixed image sub-region by simply controlling the desired length of the output.

Abstract Meaning Representation (AMR). AMR [7] is a rich semantic formalism for expressing the meaning of natural language sentences as a formal graph. AMR draws on PropBank, which is a rich lexical semantic resource encoding predicates expressing an action or state, as well as the number and nature of the participating entities (arguments and other semantic roles, such as location, manner, etc.). AMR is a widely researched semantic formalism for which highly accurate automatic Text-to-AMR and AMR-to-Text models are developed [5,9]. We rely on these models to augment original image–caption datasets with newly generated captions (as explained in Sect. 4).

AMRs vs. Scene Graphs. Recent studies [1,15,16,52] have shown that AMRs better capture the semantic relations of an image as compared to the scene graphs [11]. Existing scene graph annotations mainly capture geometric or possessive relations, which account for more than 90% of the relations captured, whereas more than 1/3 of the captured entities refer to clothing, object, or body parts information [1,52]. This difference is crucial for high-quality image caption-

[4] `https://propbank.github.io`

ing, as we use higher-level semantic relations in our everyday language rather than geometric ones. For instance, during a soccer game, we would probably describe a goal save as 'the player kicks the ball away from the goal' or 'the goalkeeper defends his team by saving a goal' and not by using mainly geometric and possessive relations like 'a person wearing a white shirt, standing with his right leg lifted, close to a ball which is above the ground'. In the supplementary, we provide a detailed comparison of AMR and scene graph representations, particularly focusing on their applications in data augmentation.

3 Model

We propose CIC-BART, specifically designed to generate controlled image captions. Specifically, it can generate descriptions of particular areas within a scene with a desired level of detail. Our model, based on VL-BART [14], utilizes a transformer-based encoder-decoder architecture, as shown in Fig. 3. CIC-BART extends VL-BART encoder to the CIC task by modifying the encoder input to include: a) a global image embedding that provides the context of the full image to the model; b) the visual control signal, including the visual embedding of the regions that contain the entities of interest; c) the text control signal, containing length control (indicating the desired length range of the output caption) and an *optional* verb signal that indicates the action we want the generated caption to concentrate on.

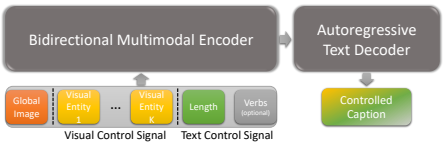

Fig. 3: The architecture diagram of our model, CIC-BART.

The visual embeddings of the regions are position-aware embeddings from a Faster R-CNN model [38] trained for visual object and attribute classification [3] on Visual Genome. The global image feature vector is extracted as well from Faster R-CNN. For the length control signal, we add to our vocabulary L tokens for the L different caption length levels; for instance, level one represents sentences between one and nine words, and level two, ten to nineteen. These tokens describe our coarse levels, for a finer sentence size accuracy, we accompany the tokens with the desired number of words. This choice gives our model the capacity to generate diverse captions for a particular length level. Finally, the output of the decoder generates the desired, controlled image caption.

4 Structured Semantic Augmentation (SSA)

The goal of our SSA method is to augment existing image captioning datasets with new focused captions along with their control signals (i.e., regions corresponding to entities). We rely on datasets where visual entities in the captions

are annotated with their corresponding regions (see Sect. 5 for details on the datasets). The SSA process consists of four main steps, as described below. For more details, refer to our supplementary material, which also includes a step-by-step example of our SSA methodology.

Step 1: Image-level AMR graph generation. Our objective in this stage is to enclose all the information available from the visually grounded captions into a single representation. To accomplish this, we create a visually grounded AMR graph (vgAMR) for each caption of an image and then merge them into a single image-level graph, the meta-vgAMR. To create the vgAMRs of an image, we first convert each of its N captions to their AMR representation, using the Neural transition-based Text-to-AMR parser [5] which also aligns words in a caption with their respective nodes in the AMR graph. We utilize the alignment information of 1) caption words and AMR nodes (from Text-to-AMR parser) and 2) caption words to image bounding boxes (from existing dataset annotations) to visually ground the AMR nodes. After this step, we get the collection of nodes referring to visual entities, where each grounded meta-AMR node is linked with a non-empty set of bounding boxes. This extended representation, 'AMR + visual grounded nodes', is our vgAMR.

Our next step is to combine the N vgAMRs to form a single meta-vgAMR. To achieve this, we employ a pairwise strategy to merge the most similar vgAMRs first (we measure similarity with Smatch score [10]). We use the UPGMA hierarchical clustering algorithm [30,34] to find the optimal merge ordering starting from the most similar graphs. UPGMA creates a hierarchy where the bottom level consists of the N individual vgAMRs. By merging all vgAMRs using the UPGMA ordering, we obtain a single structure called meta-vgAMR.

When merging two vgAMRs, the main challenge is identifying which nodes correspond to the same concepts, such as entities, attributes, actions, and relations. We use three node properties to accomplish this: a) visual grounding information, b) semantic similarity of node labels, and c) node neighborhood semantic similarity. We derive two node-merging criteria from there: 1) visually grounded entity nodes are merged if they point to the same image-bounding boxes. When 1) does not hold, we check the second criterion: 2) for the remaining non-grounded nodes, including amr-specific, predicates, adjectives, and adverbs, we use a combination of node label semantic similarity (cosine similarity of the labels using their GloVe embeddings) and neighborhood similarity. Neighborhood similarity examines the similarity of parents for adjectives/adverbs nodes and children for predicate nodes, along with the similarity of connecting edge roles. When two nodes satisfy criterion 1) or 2), we merge them into a single node. Moreover, if they have different labels, we maintain both names by keeping a list of synonyms to increase representation diversity. In the special case when the two vgAMRs describe two totally different concepts, and hence they have no common nodes, we add an amr-specific node called 'multi-sentence' as the root with the two independent vgAMRs as its children. The final graph, meta-

vgAMR, includes all non-redundant[5] elements of the original N captions while preserving the visual grounding between the meta nodes and their respective image regions.

Remark: Meta-vgAMR efficiently compresses all available image information into a single structure. Following our approach, we can easily scale when new scene information becomes available by applying our pairwise merge procedure.

Step 2: Event-based graph sampling from image-level AMRs. We start from the predicate nodes, which mainly correspond to verbs, to sample subgraphs in meta-vgAMR graphs. Predicate nodes are identified by their label and the edges connected to them. The label of a predicate node typically follows the format 'predicate_name-xx,' where 'xx' represents the different senses a word can have regarding the concept it is used for. Predicate nodes have outward ARGy edges, where 'y' can take values from 0 to 5, connecting them to their arguments. We sample subgraphs from these nodes by following the outgoing argument edges, which are labeled as ARGn in an AMR graph, each defining a particular semantic role (e.g., ARG0 points to the agent, ARG1 to the patient, etc.). Finally, we add one more subgraph containing the remaining children branches of other non-ARG optional predicate edges (e.g., 'location', 'time'). We repeat this process until the leaves of the graphs are reached. During sampling, we randomly select one of the synonyms if a node is a list of synonym labels, as mentioned in the previous step. The output of this step is our more focused *event-focused* sub-graphs. In Fig. 2, we can see some instances of our event-based sampling (SSA samples), where the predicate nodes include z0/sit-01, z7/dock-01, z13/sit-01, and so on[6]. Although we cannot show all the sampled event-based subgraphs in the figure, we included five of them and used colored roots and edges for visualization purposes.

Step 3: New caption generation from sampled AMRs We use the SPRING AMR-to-Text model [8] to generate new event-focused captions from the sampled vgAMR subgraphs. Because both vgAMR merging and sampling steps introduce noise, the output captions are not always of good quality. We automatically filter low-quality captions by using a linguistic well-formedness measure, GRUEN [58], which is a reference-free metric based on BERT contextual embeddings.

Step 4: Control signal generation. The last step is to create the control signal for the generated captions. The spatial control signal for a specific caption is extracted from the corresponding sampled vgAMR, by pulling the bounding boxes of the visual entity linked AMR nodes.

[5] A node may have different names for the same bounding box in different meta-vgAMRs, such as 'A male' and 'A person'. According to criterion 1), we merge the corresponding AMR nodes and keep both 'male' and 'person' in the names list to avoid redundancy. Therefore, criteria 1) and 2) ensure that multiple nodes don't describe the same concept in the meta-vgAMR.

[6] Note that in Fig. 2, the node z17/green-03 is also categorized as a predicate. This may seem an error because we usually think of 'green' as an attribute node rather than a predicate. However, in AMRs, when 'green' is paired with its argument, in this case, z16/lawn, it encapsulates a predicate/verb that can be expressed in natural language as 'the lawn is green.'

4.1 Mixing Strategies of Original and SSA Data

To analyze the impact of our SSA data, we explore various mixing strategies with the original training set. Assume $\mathcal{D}$ represents the training control-caption pairs in the original dataset, containing $N_{\mathcal{D}}$ samples, and SSA represents our SSA samples, containing N_{SSA} instances. The augmented dataset $\mathcal{D}_{SSA}$ is defined by combining $\mathcal{D}$ and SSA: $\mathcal{D}_{SSA} = \text{sam}_{\mathcal{D}}(\tau_{\mathcal{D}}, p_{\mathcal{D}}) \cup \text{sam}_{SSA}(\tau_{SSA}, p_{SSA})$, where the functions $\text{sam}_{\mathcal{D}}$ samples a subset of the original dataset, and sam_{SSA} a subset of our SSA data. Since we are interested in the effect of our SSA, we assume that $\text{sam}_{\mathcal{D}}(\tau_{\mathcal{D}}, p_{\mathcal{D}}) = \mathcal{D}$, with $\tau_{\mathcal{D}}$ = 'Random Sampling Strategy' and $p_{\mathcal{D}} = 100\%$, meaning that all original data are included in the mixed dataset. Depending on the sam_{SSA} parameter τ_{SSA} we have the cases:

Random Sampling Strategy. In this case, we randomly select a pre-specified number of examples from SSA. The parameter p_{SSA} expresses the percentage of SSA samples included in $\mathcal{D}_{SSA}$. With boundary cases $p_{SSA} = 100\%$ (all N_{SSA} samples are included), and $p_{SSA} = 0\%$ (no SSA data are added).

Uniform-Coverage Sampling Strategy. To mitigate the original dataset's bias (having mainly samples describing the entire image), we aim to create a new focus-unbiased dataset. By modeling the control signal focus as the image area percentage covered by the bounding boxes of the control signal, we split the original data into B coverage bins. Then, we will randomly add in each bin SSA samples, aiming to create a new uniform, coverage-unbiased $\mathcal{D}_{SSA}$ dataset. Here, p_{SSA} contains the range of each bin for the coverage histogram. For example, in the case where we choose ten uniform coverage bins, we have $p_{SSA} = \{[0\%, 10\%), [10\%, 20\%), \ldots, [90\%, 100\%]\}$.

Due to space limitations, we present results from the Random Sampling Strategy for $p_{SSA} = 0\%$ and $p_{SSA} = 100\%$ in the main paper. Results from other scenarios can be found in the supplementary material.

5 Experimental Setup

5.1 Data

We use Flickr30k Entities (Flickr-Ent) [36] and MS-COCO Entities (COCO-Ent) [18] for training and evaluation. Flickr-Ent augments the original captions of Flickr30k [51] with manually-annotated region–phrase groundings. Flickr-Ent contains the original $31K$ images annotated with five captions each. COCO-Ent augments the original MS-COCO [27] ($120K$ images each annotated with around five captions) with semi-automatically collected grounding annotations; see [18] for details on the annotation process. For both datasets, we follow previous work and use the training and test splits by Karpathy *et al.* [25]. We apply our SSA algorithm on the aforementioned datasets to create their augmented variations, COCO-Ent-SSA and Flickr-Ent-SSA, containing about $800K$ and $250K$ training captions, respectively, of which 33% and 37%, are generated by our SSA algorithm.

For all four training sets, we automatically generate image–control–caption triplets to train our model on. For spatial control, we extract from the grounded captions the bounding boxes of the entities of interest using the annotations from COCO-Ent and Flickr-Ent. Note that since these datasets do not contain the text control signal, we use each ground-truth caption as a proxy for a controlled caption, from which we first generate the coarse- and fine-length control levels and then extract their verbs using part-of-speech tagging for the optional action control.

5.2 Models and Evaluation Metrics

We compare two variations of our model (with and without SSA augmentations) with SOTA models as our baselines: Show Control & Tell (SCT) [18] that uses region-based control (bounding boxes of visual entities of interest); ASG2Caption (ASG) [13] that draws on visually grounded abstract scene graphs as control signal; and VSR [12] that uses overly descriptive control signals that express verb(s) and fine-grained verb-specific semantic roles of the desired captions; ComPro [48] that learns a mapping from the bounding boxes of the entities of interest and caption length level to GPT-2 Large prompts aiming to retrieve controlled captions; and the LaBERT length-control-only model [19].

We report the performance of our models and baselines using a comprehensive set of metrics that evaluate different aspects of caption controllability, diversity, and quality. We also propose and report an overall performance metric that summarizes these different aspects in a meaningful way. To measure *diversity*, we compute n-gram diversity, D-n for $n = 1, 2$ [4], as well as self-CIDEr-based diversity (sC) [46]. For a fair comparison of the different CIC models, we measure diversity for the five generated captions for each test image (in COCO-Ent and Flickr-Ent), and report their average. To measure *content controllability* we design an extended version of the IoU metric of [18] that calculates the degree-of-match (faithfulness) between a control signal and the corresponding generated caption. For our control signal, we use the set of nouns $\mathcal{E}$ that represent the entities of interest, which are the names of the visual objects in the control. To extract nouns from the predicted sentences, we use the Stanford part-of-speech tagger [42,43]. We then find the semantic intersection of the two sets using Hungarian Matching, as in [18]. Finally, we calculate the semantic intersection over union of the control nouns and the nouns extracted from the controllable caption, which gives us our content controllability IoU. We further analyze IoU by introducing the Hallucinating Nouns (Hal) metric. Details can be found in the supplemental material. For *length controllability* (L), we measure the Mean Absolute Error (MAE) between the fine length control (number of words) and the size of the resulting $M = 10$ controlled captions, which are generated from M randomly created control signals. We also calculate the length precision (LP) [19] by determining the percentage of generated captions that match the desired coarse length level. We assess *text quality* of generated captions using GRUEN (G) [58], a reference free metric based on BERT contextual embeddings that measure the syntactic and semantic well-formedness of a text segment. Finally, we measure

Table 1: Performance of CIC models for the original test sets. **ASG-type dataset is not released for Flickr-Ent; therefore, we could not reproduce the ASG results.*

Model	$H\uparrow$	IoU↑	G↑	sC↑	D-1↑	D-2↑	L↓	$H\uparrow$	IoU↑	G↑	sC↑	D-1↑	D-2↑	L↓
	COCO-Ent							Flickr-Ent						
SCT [18]	55.8	67.3	64.4	42.8	27.0	35.5	-	54.6	50.7	79.8	44.0	29.3	36.5	-
ASG* [13]	74.2	72.6	72.0	78.3	37.8	56.6	-	-	-	-	-	-	-	-
VSR [12]	56.2	**77.6**	39.0	67.4	30.0	42.2	-	62.5	**60.2**	54.0	77.9	33.3	49.3	-
CIC-BART	75.9	76.2	73.0	78.7	38.0	56.2	.49	69.8	54.0	85.0	78.6	43.6	58.2	1.24
CIC-BART-SSA	**78.3**	77.2	**74.8**	**82.5**	**44.6**	**63.2**	**.11**	**71.3**	55.0	**86.0**	**81.7**	**47.0**	**62.6**	**1.05**

Table 2: Performance of CIC models, for the SSA-only samples.

Model	$H\uparrow$	IoU↑	G↑	sC↑	D-1↑	D-2↑	$H\uparrow$	IoU↑	G↑	sC↑	D-1↑	D-2↑
	COCO-Ent (SSA only)						Flickr-Ent (SSA only)					
SCT [18]	51.7	62.1	64.8	37.8	23.7	31.0	43.9	29.9	77.3	45.7	31.0	36.7
CIC-BART	69.2	61.4	73.9	74.0	44.2	57.0	68.5	53.0	80.5	79.8	52.9	62.9
CIC-BART-SSA	**75.6**	**65.2**	**80.7**	**83.7**	**53.8**	**67.8**	**72.0**	**55.6**	**82.9**	**86.1**	**56.5**	**69.3**

the *overall performance* of each model based on its ability to balance content controllability, diversity, and text quality. To calculate this, we use the harmonic mean of IoU, G, and sC. All of these metrics range between 0 and 1, with a higher value indicating better performance. The harmonic mean (H) helps us determine the model with the best overall performance. It prioritizes models that perform well across all metrics while penalizing those with poor performance, even in one metric. In our supplementary, we include comparisons of the CIC models on *standard captioning metrics* (like CIDEr [44] and Spice [2]).

6 Results

6.1 Overall Performance

We first compare the overall performance of our models with the SCT, ASG, and VSR baselines with respect to controllable captioning metrics. We do not include ComPro in this comparison due to the unavailability of the codebase. We also exclude LaBERT since this model solely focuses on length controllability.

Tab. 1 presents results for content and length controllability (IoU and L), text quality (G), and diversity (sC, D-1, D-2), as well as the harmonic mean (H). For both datasets, CIC-BART-SSA has the best performance in all metrics, except IoU, where it is the second best. Specifically, CIC-BART-SSA is superior to all other models with respect to diversity (sC, D-1, D-2) and text quality (G), but comparable to VSR in terms of content controllability (IoU). The length controllability (L) scores show that our SSA augmentation helps the model learn to generate high-quality output at the desirable length (compare CIC-BART and CIC-BART-SSA). This is due to the increased diversity in caption length provided by our SSA augmentations.

Importantly, we can see that model performance can vary depending on the metric. E.g., whereas VSR has the highest IoU, it falls behind in text quality and

diversity. In our qualitative analysis, we observe the poor quality of the captions generated by VSR. The best-performing model should be identified based on the H score that summarizes content controllability, text quality, and diversity into a single score. Based on this score, CIC-BART-SSA is better than SCT and VSR baselines by a large margin, and notably better than ASG. Nevertheless, ASG requires complex control signals in the form of scene graphs, in contrast to the simple control signal requirements of CIC-BART.

Next, we conduct a further evaluation of the CIC performance on our SSA samples from the test set images of COCO-Ent and Flickr-Ent. We present the results in Tab. 2 for our models CIC-BART, CIC-BART-SSA, and SCT. ASG and VSR are excluded since they need complex control signals (grounded abstract scene graphs for ASG and grounded verb semantic roles for VSR), which are only available for COCO-Ent and Flickr-Ent. We observe a significant improvement in overall performance for our CIC-BART-SSA model. We notice that the models (SCT and CIC-BART) trained on the original datasets, which describe the entire image, had difficulties generalizing to cases where they had to focus on a specific sub-region of an image. However, our model CIC-BART-SSA was able to generate focused and diverse descriptions of the challenging, highly focused examples present in our SSA data.

We compare the length precision of our model with the baselines utilizing length control in Table 3. LaBERT uses only length-control signals without spatial control, while our model employs both spatial and length-control signals to generate focused captions. This makes the LaBERT task much easier since it only focuses on generating specific length descriptions of an image. On the other hand, our model focuses on generating captions that describe only a specific sub-region of the scene while maintaining a desired description length level. Although our task is more challenging than LaBERT, we achieve competitive length precision performance. Lastly, we want to emphasize that the remarkable improvement in length controllability (L) and length precision (LP) from CIC-BART to CIC-BART-SSA stems from the increased length diversity found in our SSA augmentations, which enriches the original COCO-Ent and Flickr-Ent datasets. In the supplementary, we provide an analysis of the caption length statistics in the original datasets and our SSA-derived captions.

Table 3: Length Precision (LP) for CIC models on COCO-Ent and Flickr-Ent original test sets.

Model	LP↑	LP↑
	COCO-Ent	Flickr-Ent
ComPro [48]	94.7	81.4
LaBERT [19]	<u>99.7</u>	**98.4**
CIC-BART	**99.9**	88.0
CIC-BART-SSA	**99.9**	<u>91.3</u>

6.2 Effect of SSA on Content Controllability

To analyze the impact of our SSA augmentations, we measure the content controllability (IoU) performance of CIC-BART at different levels of focus of the control signals and report it in Fig. 4. We use coverage, defined as the area of the image enclosed by the bounding boxes of the entities of interest in the control

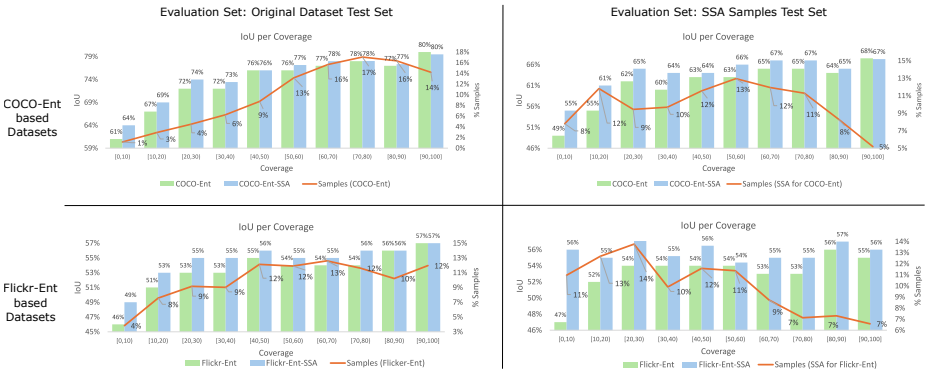

Fig. 4: IoU histograms for CIC-BART (green) and CIC-BART-SSA (blue). The first column depicts the IoU on the original test set and the second on the original test set images' SSA (only) data. The Samples curve (orange) represents the distribution of test captions in each coverage (image area covered by the control signal) interval.

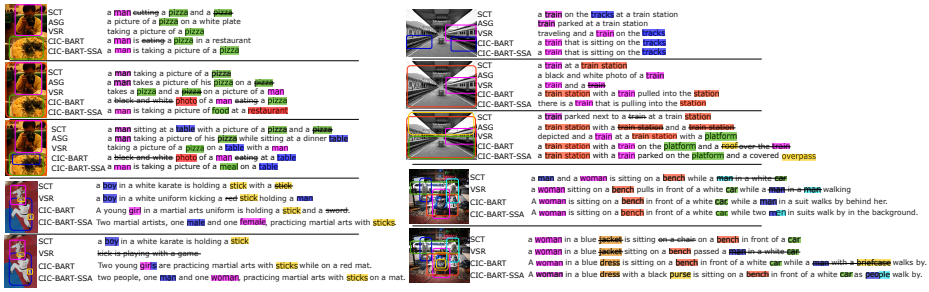

Fig. 5: Qualitative examples for the original test sets. Strikethrough marks hallucinations and redundancies. a, b licensed under CC BY-SA 2.0; c, d under CC BY 2.0.

signal, to quantify that focus. For example, highly focused control signals cover a small area, yielding low coverage, while broader signals cover a larger area and have high coverage. We 'break down' the IoU performance into 10 coverage bands and report the average IoU over control signals in those bands. In addition, the 'Samples' curve shows the distribution of test captions over the same bands. The results in Fig. 4 indicate that by training with SSA (blue bars), the spatial controllability improves significantly in the low-coverage regime, where the control signals are highly focused. Interestingly, these are also the most underrepresented (data deprived) parts of the original dataset Flickr-Ent. Therefore, SSA which enriches the original datasets with highly focused examples (please refer to our supplementary for % Samples per coverage bands for the training sets), is effective in improving generalization performance in CIC. We include the performance of hallucinating nouns metric in the supplementary material.

6.3 Qualitative Analysis

In Fig. 5, we present qualitative examples from the original test sets, and in Fig. 6 examples from our SSA (only) test set control signals. In the two figures, each highlighted word found in the generated controlled captions corresponds to the control entity of the same color. This shows the match between the captions produced and the control signal. We also strike through the parts where the model hallucinates or generates redundant references to the entities of interest. Our models have been observed to outperform the previous state-of-the-art models by substantially enhancing the quality of the generated controlled captions. This behavior was expected from our quantitative analysis, which showed that our models have significantly higher text quality (G). More importantly, our CIC-BART-SSA model is capable of generating captions that are faithful to the control signal and better understand the relationships that connect the entities of interest. We include additional qualitative samples in the Supplementary Material.

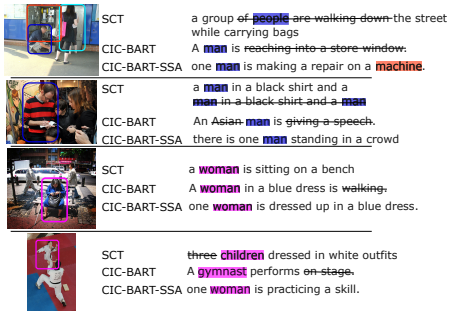

Fig. 6: Examples for SSA-only test set. Strikethrough marks hallucinations. Up to Down: a licensed under CC BY 2.0; b under CC PDM 1.0, for last two see Fig. 5.

7 Conclusions

We address two main challenges faced by the controllable image captioning (CIC) models. First, standard image–caption datasets lack the controllability and diversity needed for proper training and evaluation of CIC. Second, most recent SOTA models require complex and overly descriptive control signals as input (including, e.g., the main action/verb to appear in the generated caption). To address the first challenge, we propose a novel technique that draws on a structured semantic augmentation (SSA) formalism to generate focused captions and the corresponding control signals for images. For the second challenge, we propose a transformer-based vision-language model attuned to the CIC task. We show that this model performs competitively with SOTA models without requiring complex and explicit control signals. Importantly, when combined with our SSA approach, our model generates highly diverse captions and significantly reduces the content controllability performance gap between the different levels of focus of the generated controlled captions. Finally, when provided with the commonly used verb guidance of other SOTA approaches, our model shows a substantial improvement in performance.

References

1. Abdelsalam, M.A., Shi, Z., Fancellu, F., Basioti, K., Bhatt, D., Pavlovic, V., Fazly, A.: Visual semantic parsing: From images to abstract meaning representation. In: Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL). pp. 282–300 (2022)
2. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. pp. 382–398. Springer (2016)
3. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR (2018)
4. Aneja, J., Agrawal, H., Batra, D., Schwing, A.: Sequential latent spaces for modeling the intention during diverse image captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4261–4270 (2019)
5. Astudillo, R.F., Ballesteros, M., Naseem, T., Blodgett, A., Florian, R.: Transition-based parsing with stack-transformers. arXiv preprint arXiv:2010.10669 (2020)
6. Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., Schneider, N.: Abstract meaning representation (amr) 1.0 specification. In: Parsing on Freebase from Question-Answer Pairs. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle: ACL. pp. 1533–1544 (2012)
7. Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., Schneider, N.: Abstract meaning representation for sembanking. In: Proceedings of the 7th linguistic annotation workshop and interoperability with discourse. pp. 178–186 (2013)
8. Bevilacqua, M., Blloshmi, R., Navigli, R.: One spring to rule them both: Symmetric amr semantic parsing and generation without a complex pipeline. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 12564–12573 (2021)
9. Blloshmi, R., Bevilacqua, M., Fabiano, E., Caruso, V., Navigli, R.: Spring goes online: end-to-end AMR parsing and generation. In: Proceedings of the 2021 conference on empirical methods in natural language processing: system demonstrations. pp. 134–142 (2021)
10. Cai, S., Knight, K.: Smatch: an evaluation metric for semantic feature structures. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 748–752 (2013)
11. Chang, X., Ren, P., Xu, P., Li, Z., Chen, X., Hauptmann, A.: A comprehensive survey of scene graphs: Generation and application. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(1), 1–26 (2021)
12. Chen, L., Jiang, Z., Xiao, J., Liu, W.: Human-like controllable image captioning with verb-specific semantic roles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16846–16856 (2021)
13. Chen, S., Jin, Q., Wang, P., Wu, Q.: 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. pp. 9962–9971 (2020)
14. Cho, J., Lei, J., Tan, H., Bansal, M.: Unifying vision-and-language tasks via text generation. In: International Conference on Machine Learning. pp. 1931–1942. PMLR (2021)

15. Choi, W.S., Heo, Y.J., Punithan, D., Zhang, B.T.: Scene graph parsing via abstract meaning representation in pre-trained language models. In: NAACL 2022 Workshop on Deep Learning on Graphs for Natural Language Processing (2022)
16. Choi, W.S., Heo, Y.J., Zhang, B.T.: Sgram: Improving scene graph parsing via abstract meaning representation. arXiv preprint arXiv:2210.08675 (2022)
17. Chunseong Park, C., Kim, B., Kim, G.: Attend to you: Personalized image captioning with context sequence memory networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 895–903 (2017)
18. Cornia, M., Baraldi, L., Cucchiara, R.: 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. pp. 8307–8316 (2019)
19. Deng, C., Ding, N., Tan, M., Wu, Q.: Length-controllable image captioning. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16. pp. 712–729. Springer (2020)
20. Gan, C., Gan, Z., He, X., Gao, J., Deng, L.: Stylenet: Generating attractive visual captions with styles. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3137–3146 (2017)
21. Guo, L., Liu, J., Yao, P., Li, J., Lu, H.: Mscap: Multi-style image captioning with unpaired stylized text. 2019 ieee. In: CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4199–4208 (2019)
22. Hirsch, E., Tal, A.: Clid: Controlled-length image descriptions with limited data. arXiv preprint arXiv:2211.14835 (2022)
23. Johnson, J., Karpathy, A., Fei-Fei, L.: Densecap: Fully convolutional localization networks for dense captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4565–4574 (2016)
24. Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D., Bernstein, M., Fei-Fei, L.: Image retrieval using scene graphs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3668–3678 (2015)
25. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3128–3137 (2015)
26. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16. pp. 121–137. Springer (2020)
27. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
28. Lindh, A., Ross, R.J., Kelleher, J.D.: Language-driven region pointer advancement for controllable image captioning. arXiv preprint arXiv:2011.14901 (2020)
29. Lu, J., Yang, J., Batra, D., Parikh, D.: Neural baby talk. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7219–7228 (2018)
30. Lukasová, A.: Hierarchical agglomerative clustering procedure. Pattern Recognition **11**(5-6), 365–381 (1979)
31. Luo, J., Li, Y., Pan, Y., Yao, T., Feng, J., Chao, H., Mei, T.: Semantic-conditional diffusion networks for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23359–23368 (2023)
32. Mathews, A., Xie, L., He, X.: Senticap: Generating image descriptions with sentiments. In: Proceedings of the AAAI conference on artificial intelligence. vol. 30 (2016)

33. Mathews, A., Xie, L., He, X.: Semstyle: Learning to generate stylised image captions using unaligned text. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8591–8600 (2018)
34. Müllner, D.: Modern hierarchical, agglomerative clustering algorithms. arXiv preprint arXiv:1109.2378 (2011)
35. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
36. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision. pp. 2641–2649 (2015)
37. Ramos, R., Martins, B., Elliott, D., Kementchedjhieva, Y.: Smallcap: lightweight image captioning prompted with retrieval augmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2840–2849 (2023)
38. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
39. Ren, Y., Mao, Z., Fang, S., Lu, Y., He, T., Du, H., Zhang, Y., Ouyang, W.: Crossing the gap: Domain generalization for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2871–2880 (2023)
40. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7008–7024 (2017)
41. Shuster, K., Humeau, S., Hu, H., Bordes, A., Weston, J.: Engaging image captioning via personality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12516–12526 (2019)
42. Toutanova, K., Klein, D., Manning, C.D., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 human language technology conference of the north american chapter of the association for computational linguistics. pp. 252–259 (2003)
43. Toutanvoa, K., Manning, C.D.: Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora. pp. 63–70 (2000)
44. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
45. Wang, N., Xie, J., Wu, J., Jia, M., Li, L.: Controllable image captioning via prompting. In: AAAI (2023)
46. Wang, Q., Chan, A.B.: Describing like humans: on diversity in image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4195–4203 (2019)
47. Wang, T., Zhang, J., Fei, J., Ge, Y., Zheng, H., Tang, Y., Li, Z., Gao, M., Zhao, S., Shan, Y., et al.: Caption anything: Interactive image description with diverse multimodal controls. arXiv preprint arXiv:2305.02677 (2023)
48. Wang, Z., Xiao, J., Chen, L., Gao, F., Shao, J., Chen, L.: Learning combinatorial prompts for universal controllable image captioning. arXiv preprint arXiv:2303.06338 (2023)

49. Xia, Q., Huang, H., Duan, N., Zhang, D., Ji, L., Sui, Z., Cui, E., Bharti, T., Zhou, M.: Xgpt: Cross-modal generative pre-training for image captioning. In: Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part I 10. pp. 786–797. Springer (2021)
50. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. pp. 2048–2057. PMLR (2015)
51. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics **2**, 67–78 (2014)
52. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: Scene graph parsing with global context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5831–5840 (2018)
53. Zeng, Z., Zhang, H., Lu, R., Wang, D., Chen, B., Wang, Z.: Conzic: Controllable zero-shot image captioning by sampling-based polishing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23465–23476 (2023)
54. Zhao, W., Wu, X., Zhang, X.: Memcap: Memorizing style knowledge for image captioning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12984–12992 (2020)
55. Zhao, Y., Wei, J., Lin, Z., Sun, Y., Zhang, M., Zhang, M.: Visual spatial description: Controlled spatial-oriented image-to-text generation. arXiv preprint arXiv:2210.11109 (2022)
56. Zheng, Y., Li, Y., Wang, S.: Intention oriented image captions with guiding objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8395–8404 (2019)
57. Zhong, Y., Wang, L., Chen, J., Yu, D., Li, Y.: Comprehensive image captioning via scene graph decomposition. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16. pp. 211–229. Springer (2020)
58. Zhu, W., Bhat, S.: Gruen for evaluating linguistic quality of generated text. arXiv preprint arXiv:2010.02498 (2020)