Supplementary for "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

1 Overview

In our supplementary material, we provide more details for our SSA methodology in Sec. 2; in Sec. 3, we provide additional details for our experimental set-up; and finally, in Sec. 4 we provide extended qualitative and quantitative results of our performed experiments and ablations. Specifically, we focus on:

- Dataset statistics before and after SSA augmentation in Sec. 4.1.
- Impact of mixing of original and SSA captions in Sec. 4.2.
- Effects of SSA on content controllability in Sec. 4.3.
- SSA-induced diversity in Sec. 4.4.
- Standard Captioning Performance of CIC models Sec. 4.5
- Qualitative comparisons in Sec. 4.6.
- Comparison of SSA and alternative augmentation strategies in Sec. 4.7, with attention on LLM-based paraphrasing and Scene Graph-based methods.

2 Structured Semantic Augmentation (SSA)

In this section, we will provide additional information on the SSA augmentation strategy we introduced in our main paper. We summarize the steps in constructing the meta-vg Graph (as described in Step 1: Image-level AMR graph generation in our main paper) in Algorithm 1. we include a detailed example of our SSA methodology in Fig. 1. We present the flow diagram that explains the process of determining if two nodes from different vgAMRs refer to the same concept and, therefore, should be merged in Fig. 2.

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

Algorithm 1 meta-vgAMR Graph Construction

1: **Input**: An image I with N human-generated, visually-grounded captions; We denote the visually grounded entities of each caption as $\{G_i^{en}\}_{i=1}^{N}$;
2: **Output**: The meta-vgAMR graph, $\mathcal{A}_{Meta}^{vg}$, of the N captions;
3: **Initialize**: Generate the individual AMR graphs $\{\mathcal{A}_i\}_{i=1}^{N}$ for each image caption using a pre-trained Text-to-AMR semantic parser with (AMR node–caption word) alignment; Construct the vgAMRs, $\mathcal{A}^{vg} = \{\mathcal{A}_i^{vg}\}_{i=1}^{N}$, using the visual grounding annotations and (AMR node–caption word) alignment;
4: Compute $D = 1-$ SmatchScore$(\mathcal{A}^{vg})$; ▷ A symmetric $N \times N$ AMR graph distance matrix between all $\mathcal{A}_i^{vg}, \mathcal{A}_j^{vg}$ pairs.
5: bottomUpHCs = UPGMA(D); ▷ Bottom-up hierarchical clusters, each cluster contains two vgAMR graphs.
6: **for** $(\mathcal{A}_i^{vg}, \mathcal{A}_j^{vg})$ in bottomUpHCs **do** ▷ Following the bottom-up hierarchy, pair-wise merge the vgAMRs of each cluster.
7: $\mathcal{A}_i^{vg} = (\mathcal{N}_i, \mathcal{E}_i)$; $\mathcal{A}_j^{vg} = (\mathcal{N}_j, \mathcal{E}_j)$ ▷ The nodes and edges of each vgAMR graph.
8: Initialize $\mathcal{A}_m^{vg} = (\mathcal{N}_m, \mathcal{E}_m)$ as a null graph;
9: $\mathcal{N}_{\text{common}}$ = getCommonNodes$(\mathcal{A}_i^{vg}, \mathcal{A}_j^{vg})$; ▷ Returns the common nodes between the two vgAMR graphs.
10: **if** $\mathcal{N}_{\text{common}}$ is empty **then** ▷ The two vgAMRs have no overlapping information.
11: $\mathcal{N}_m = \mathcal{N}_i \cup \mathcal{N}_j \cup \mathcal{N}_{\text{multi-sentence}}$; ▷ Introduce a new, AMR-specific "multi-sentence" node, to be the root of the merged graph. This node will connect the two disjoint vgAMR graphs.
12: **else**
13: $\mathcal{N}_i' = \mathcal{N}_i \setminus \mathcal{N}_{\text{common}}$; $\mathcal{N}_j' = \mathcal{N}_j \setminus \mathcal{N}_{\text{common}}$
14: $\mathcal{N}_m = \mathcal{N}_{\text{common}} \cup \mathcal{N}_i' \cup \mathcal{N}_j'$
15: $\mathcal{E}_m$ = getConnectingEdges$(\mathcal{A}_i^{vg}, \mathcal{A}_j^{vg}, \mathcal{N}_m)$
16: $\mathcal{A}^{vg}$.remove$(\mathcal{A}_i^{vg}, \mathcal{A}_j^{vg})$
17: $\mathcal{A}^{vg}$.add$(\mathcal{A}_m^{vg})$
18: $\mathcal{A}_{Meta}^{vg} = \mathcal{A}^{vg}$ ▷ All N vgAMRs are merged into one representation
19: **return** $\mathcal{A}_{Meta}^{vg}$

SSA Algorithm. To construct the hierarchical clusters, we use the UPGMA algorithm, which considers each individual vgAMR as a separate cluster at Level 0. Two clusters are merged at each level based on their distance, starting with the most similar graphs. To measure similarity, we use the SmatchScore between two vgAMR graphs. Since the Smatch score is a metric from 0 to 1, we use 1- Smatch score as the distance metric for the UPGMA algorithm. For this example, the AMR graphs of captions (2) and (3) are the most similar, so they are merged first to create their joint vgAMR graph at Level 1. Every graph from levels 1 to 4 results from the 6-17 step of our Algorithm 1 where we merge two graphs from lower layers according to the hierarchical clusters computed by UPGMA. The final layer (4) graph is our meta-vgAMR graph, which contains all information from the original vgAMRs and, thus, from the available original captions. By applying our event-focused sampling approach, we can generate novel, focused, visually grounded descriptions from this new structure. Some examples can be

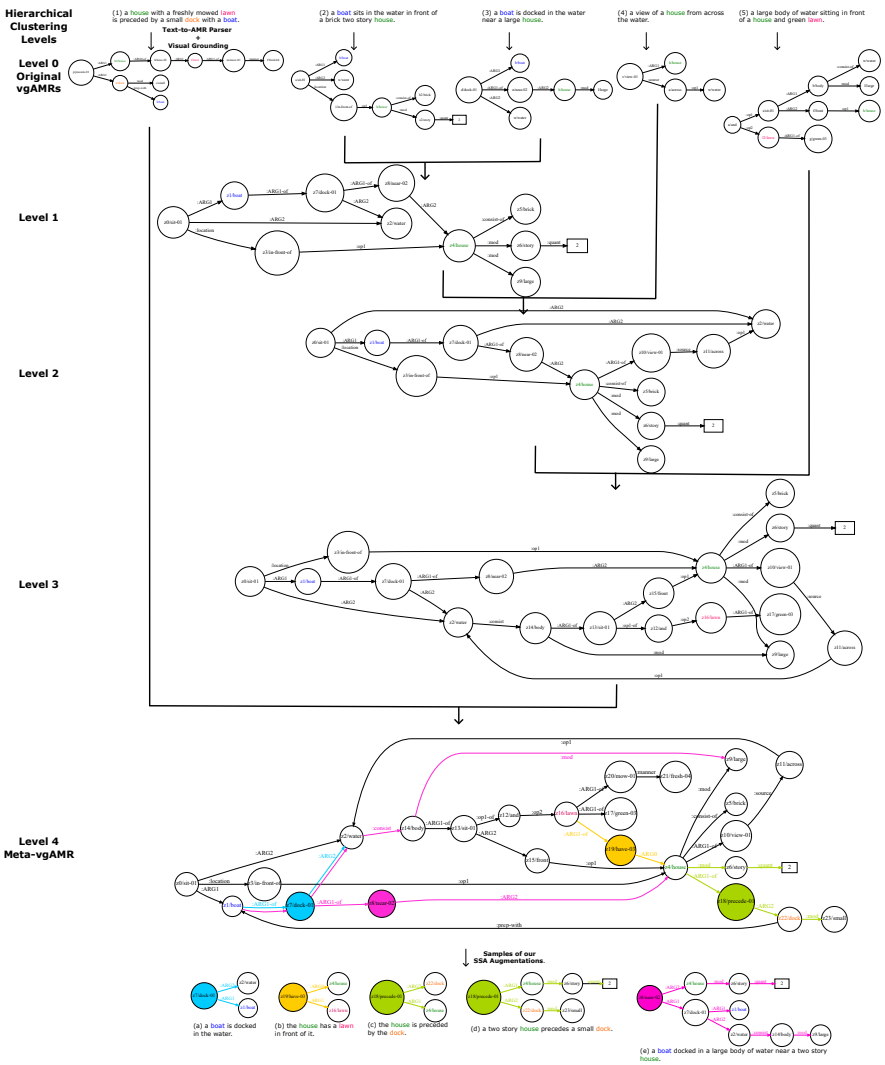

Fig. 1: Our Structured Semantic Augmentations process generates new focused captions from visually grounded captions. We derive individual vgAMRs from the original captions and merge them using hierarchical clusters based on the graph similarity of the original vgAMRs (Level 0). At the final layer, we obtain the meta-vgAMR, which combines all available information into one structure. Then, we sample sub-graphs from our meta-vgAMR to generate new captions. Examples (a)-(e) show some of the vgAMRs we sampled with their generated captions.

seen at the bottom of Fig. 1, along with the resulting captions generated by pre-trained AMR-To-Text parsers.

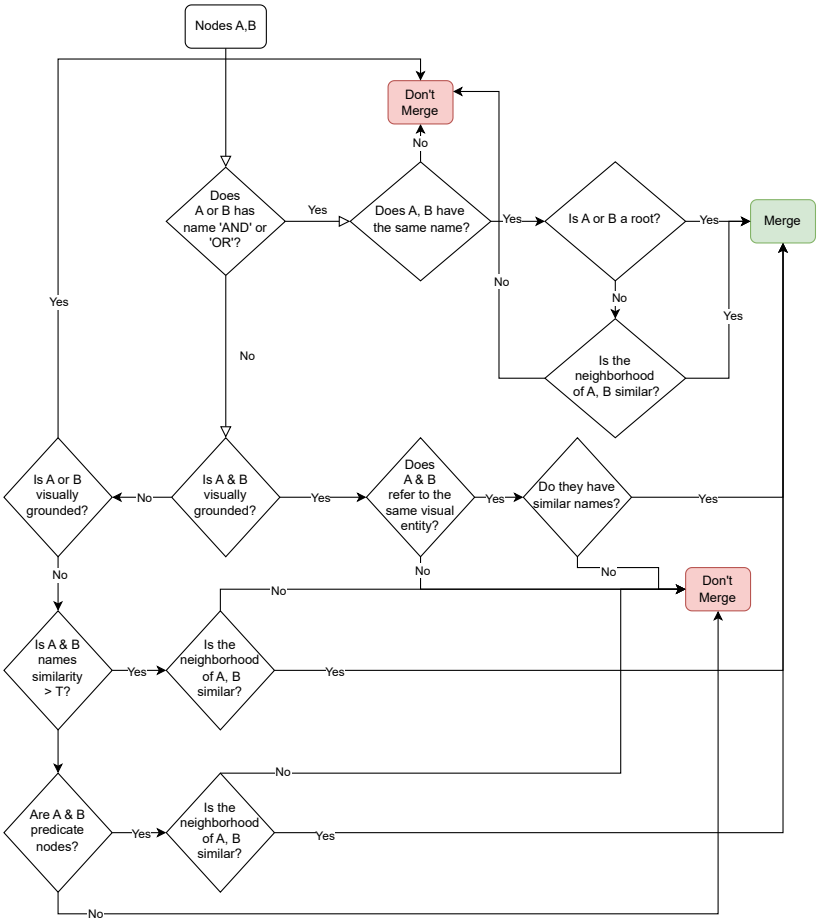

Fig. 2: The flow diagram shows the process of merging nodes A and B from two vgAMRs. The merging process starts with handling the AMR-specific nodes (`AND`), followed by the visually grounded vgAMR nodes. For the remaining non-grounded nodes (such as predicates, adjectives, and nouns), we check if their names are synonyms (semantically similar) to decide if they refer to the same concepts and can be merged. If the nodes do not fit the above categories, we check if they are predicate nodes. If they are, we examine their neighborhood to determine if they can be merged.

Finding same-concept nodes in two vgAMR graphs. In the flow diagram labeled as Fig. 2, we can observe that when we merge two vgAMR graphs, vgAMR-A and vgAMR-B, we need to identify the common concept nodes between the two representations and combine them. This merging process serves two purposes: a) it allows for a more efficient and compressed representation by reducing re-

dundancies and eliminating multiple nodes for the same concept, and b) it consolidates all available information about a particular concept found in different captions. For example, in Fig. 8 c) for the top player, one caption may describe her clothing, another her physical characteristics, and yet another her actions (for instance, practicing martial arts). Despite the differences, all these captions have a common concept: the person in the picture. Instead of having three separate nodes with partial information, we aim to create a single node (person) that consolidates all available information about the person in the image, making it easier to explore all connected nodes and access the complete information.

The process for identifying common nodes involves the steps outlined in Fig. 2. We start by checking if the two nodes are AMR-specific nodes of type `AND`. If these nodes are found at the root of the graph, it indicates that the corresponding sentence follows the format 'FACT-1 `AND` FACT-2'. In this case, we can merge them as they represent aggregated facts about the image. If they are not root nodes, we need to be more cautious and ensure that they originate from the same concept. For this reason, we assess the similarity of their neighboring nodes. If they link to the same nodes, we merge them and combine the provided facts.

As shown in Fig. 2, if both nodes are visually grounded and refer to the same visual entity (i.e. if they have the same bounding boxes), we are hesitant to merge the nodes without first verifying that their names are synonyms [4]. This additional condition is helpful in cases where a) the original dataset visually grounds phrases instead of nouns, and b) there is noise from the Text-to-AMR parser. This check ensures that entity attributes (such as 'young' and 'tall') which may be visually grounded, are not mistakenly merged with noun nodes. Finally, if the two nodes are semantically distant or do not refer to the same visual entity, or if one of them is grounded and the other is not, we conclude that the two nodes cannot be merged.

When we don't have visual cues to help us identify similar concept nodes, we rely on the names of the nodes and their surrounding information to make decisions. If two nodes are synonyms, we look at how similar their neighbors are (e.g., if the two nodes are nouns or adjectives, do they share the same parent?). If they do, we merge them as similar concept nodes.

In our final step, we have an additional procedure for predicate nodes. In our experiments, we observed that the GloVe embeddings of predicate/verb words tend to be more distant. Therefore, in the last step, if the two nodes are predicates, and their child nodes (ARG0, ARG1, and so on) are the same, and the similarity of their names is above a certain threshold (which is smaller than the thresholds used in the previous steps), then we merge the two predicate nodes. This concludes our node merge process.

[4] We determine if two nodes are synonyms by comparing the cosine similarity of their GloVe embeddings. If the cosine similarity is above a certain threshold, we consider them synonyms.

3 Experimental Setup

3.1 Evaluation Metrics

Content Controllability: IoU. We measure content controllability using the IoU (the overlap between two sets) of the set of nouns in the control signal and the set of nouns in the generated sentence.

- For the set of control signal nouns $\mathcal{E}$ in COCO-Ent test set, we use the existing annotations. For each caption, the head noun of a noun chunk is provided. We use the set of head nouns as our $\mathcal{E}$.
- For Flickr-Ent this information is not available. We use the object labels from Faster R-CNN to get the control signal nouns.

Our IoU metric is based on the corresponding score in [8]. We modify the following parts:

- Ground truth nouns: Instead of using the ground truth captions as a proxy, we directly extract control signal nouns from the control signal itself, as described in the previous bullet points.
- Generated sentence nouns: For the generated controlled sentence, instead of looking if each word is in a dictionary of nouns prepared by [8], we use part-of-speech tagging [14, 15] to extract the nouns of the sentence. We use this approach because the provided dictionary, although it contained many nouns, was not a complete list, so in many cases, during evaluation, nouns were discarded because there was no entry for them in the dictionary, which added noise to the original metric.

Our next steps are as described in [8], that is, the Hungarian matching of the two sets of nouns using the cosine similarity of the corresponding GLoVE embeddings for each noun word. The final IoU is the sum of cosine similarities for the aligned nouns.

The advantage of our IoU score from the one proposed in [8] is that it directly compares the control signal with the generated sentence; instead of the dataset ground truth sentences, which are just a proxy of entities in the control signal. This helps reduce the metric noise, since our IoU are not affected from annotation errors (for example, ground truth captions where not all entities are annotated/grounded to a bounding box, which will lead to a noisy proxy of the control signal) or from missing entries in the noun dictionary used in [8].

Content Controllability: Hallucinations. We propose the Hallucinating Nouns (Hal) content controllability metric to help us to determine the number of hallucinations present in the generated captions. These hallucinations refer to nouns or visual entities that are not part of the control signal. They could be visual objects present in the image but not in focus of the control signal, or visual objects that are not present in the image at all. To measure this, we

propose the "Hal-lucinating Nouns" metric, which can be computed using the following equation:

$$\text{Hal} = \frac{1}{|\mathcal{N}|} \left(|\mathcal{N}| - \text{IoU}(\mathcal{N}, \mathcal{E}) \right). \tag{1}$$

where $\mathcal{N}$ is the set of nouns extracted from the generated controlled caption and $\mathcal{E}$ is the set of nouns (visual entities) in the control signal.

Diversity. To measure *diversity*, we compute n-gram diversity, D-n for n = 1, 2 [3], as well as self-CIDEr-based diversity (sC) [18]. D-n measures the ratio of distinct n-grams to the total number of words generated per set of diverse captions. sC computes the diversity of a set of captions by using their CIDEr score [17], a metric that measures sentence similarity by giving more weight to the matching of novel words. 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. Note that not all images in COCO-Ent and Flickr-Ent have five caption–control signal pairs, especially for COCO-Ent that is automatically annotated. We only considered the ones with five available pairs for diversity evaluation, including 985 images for Flickr-Ent and 112 images for COCO-Ent.

Best-5 Diversity. For completeness, we compute the best-5 diversity, proposed in [6]. Specifically, we generate $M = 10$ randomly generated control signals for a given image. From the M captions, we form all possible sets of 5 captions (M choose 5) and measure the ratio of n-grams to the total number of words for each set. We report the average of the best Div-n scores for all images in the test set.

Length Controllability. 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) [9] by determining the percentage of generated captions that match the desired coarse length level.

Text Quality. We assess *text quality* of generated captions using GRUEN (G) [21], a reference-free metric based on BERT contextual embeddings that measure the syntactic and semantic well-formedness of a text segment.

Overall Performance using Harmonic Means. Finally, we measure 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.

Standard Captioning Metrics. Following prior work, we also report performance with respect to standard captioning metrics, namely Bleu-4 (B4) [13], Meteor (M) [4], Rouge (R) [11], CIDEr (C) [17] and Spice (S) [2]. Specifically, Bleu measures the n-gram similarity of the two sentences, but without examining their synonymity, something that is addressed by Meteor. Rouge estimates the recall of their largest common subphrase. CIDEr score gives more weight to the matching of novel words and finally, Spice computes the semantic similarity by comparing the scene graph representations of the two sentences. For all these measures, higher is better. Note that these metrics are not sufficient for evaluating CIC, as they compare a generated (controlled) caption with a reference ground-truth caption, ignoring the desired effect of the control signal. We report them for completeness, but we believe that our well-formedness metric (GRUEN) is more suited for evaluating the quality of the controlled captions.

3.2 CIC-BART-SSA and Baselines Setup

SSA Parameters. For our AMR-to-Text generated sentence filtering, we set the GRUEN (G) threshold to 0.7.

Model Parameters. We initialize CIC-BART encoder and decoder from the pre-trained weights of VL-BART [7] to benefit from transfer learning. We further train our model on data that contains image–control–caption triplets, where control consists of the above-mentioned signals. We incorporate five different length levels to control the length of our output. Each level has a specific range, with level A ranging from one to nine words, level B spanning ten to nineteen words, level C covering twenty to twenty-nine words, level D consisting of thirty to thirty-nine words, and level E including sentences with forty or more words. In our CIC-BART vocabulary, we have added five tokens to represent these five caption length levels. For optimizing the cross-entropy loss, we utilize the RAdam optimizer [12] with a learning rate of $5 \cdot 10^{-5}$ and a batch size of 80. We train our models for 20 epochs and select our trained model based on the best content controllability IoU and CIDEr scores.

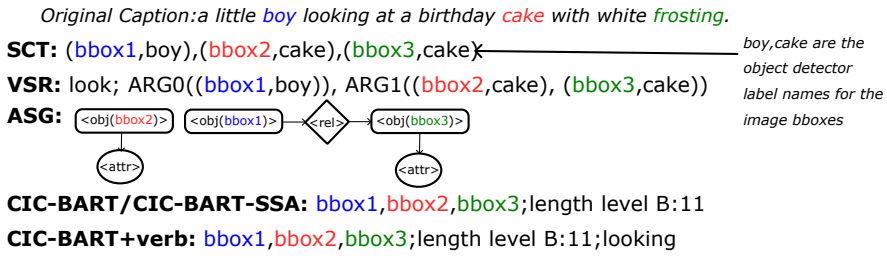

Fig. 3: An illustrative example of the used control signals from different CIC models.

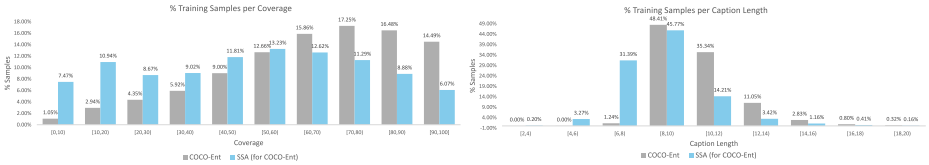

Fig. 4: Training set coverage and caption length histograms of COCO-Ent and SSA augmentations for this dataset. Coverage represents the % of the image covered by the control signal, so the left and right parts of the graph represent more focused and broader control signals, respectively. We note that the COCO-Ent-SSA dataset will contain both the original data (gray bars) and our SSA (blue bars).

Baseline Models. We conducted evaluations for metrics such as content controllability (IoU), text quality (G), and diversity (D-1, D-2 and sC) for SCT and VSR, using the code and pre-trained checkpoints available on their official project GitHub pages. However, for ASG, we re-trained and evaluated the ASG2Caption model for COCO-Ent using the official GitHub codebase since the pre-trained checkpoints were not available. Unfortunately, we could not train the ASG2Caption model on the Flickr-Ent dataset as the ASG dataset for Flickr-Ent has not been released. For the standard captioning metrics, best-5 diversity, and length precision of the testing sets of the datasets COCO-Ent and Flickr-Ent, we used the values presented in the corresponding papers.

We mention, that in our main paper, we used the strongest model performance from ComPro, which employs GPT-2 Large. This model has a total of 881M parameters, 107M of which are used for the mapping network and 774M are from GPT-2. It's worth noting that our models, namely, CIC-BART, CIC-BART-SSA, and CIC-BART +verb, use only 140M parameters, making them more than six times smaller than ComPro with GPT-2 Large.

In Fig. 3, we present an example with the control signals used by the baselines and our models for a specific instance where we need a focused caption on the boy (bbox 1) and the cake (bbox2, bbox3). The SCT model [8] uses bounding boxes of entities of interest and GLoVE embeddings of their Faster R-CNN labels as control signals. The VSR model [5] adds ground truth caption verbs and their PropBank grounded verb semantic roles to the SCT control signal. The ASG model [6] employs abstract scene graphs as control signals that provide information about how visual entities are related or connected and how many attributes they have. Our models (CIC-BART and CIC-BART-SSA) use only the bounding boxes of interest and the desired caption length level as control signals. We have also explored the use of ground truth verbs in the control signal, like in VSR, in our CIC-BART +verb model. However, unlike VSR, we only use the ground truth verb name and not their PropBank grounded semantic roles.

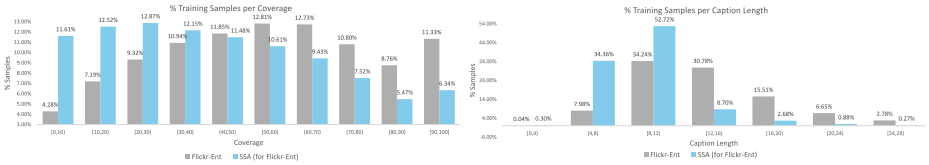

Fig. 5: Training set coverage and caption length histograms of Flickr-Ent and SSA augmentations for this dataset. Coverage represents the % of the image covered by the control signal, so the left and right parts of the graph represent more focused and broader control signals, respectively. We mention that Flickr-Ent-SSA dataset contains both the original (gray bars) and SSA (blue bars) samples.

4 Results

4.1 Original and SSA Augmented Datasets Analysis

In Figs. 4 and 5, we present the coverage and caption length statistics of the COCO-Ent and Flickr-Ent training set and their derived SSA augmentations, respectively. When analyzing the scene coverage based on the control signal, it becomes apparent that the original datasets predominantly feature samples that describe the entire image (high coverage), with very few focusing on a small portion of the scene (highly focused control signals, low coverage). This is particularly evident in the COCO-Ent dataset, where examples with focused control signals are minuscule. For the caption length statistics, we notice that COCO-Ent dataset is far from diverse with approximately 84% of the captions having 8-12 words. Similarly, in the Flickr-Ent dataset, approximately 65% of its descriptions have 8-16 words.

With our SSA data (blue bars), we augment the original datasets (gray bars) with highly focused control–caption pairs and diverse caption length, to construct a new dataset of spatially and linguistically diverse data for controllable image captioning, namely the COCO-Ent-SSA and Flickr-Ent-SSA datasets.

4.2 Original and SSA captions Mixtures

In this section, we delve deeper into the impact of our SSA augmentation on CIC models. To conduct our experiments, we utilize our mixing methodology described in our main paper (Section 4.1), to create various versions of COCO-Ent-SSA and Flickr-Ent-SSA datasets. We aim to examine the effect of our SSA examples, so we include all original samples in the mixed dataset $\mathcal{D}_{SSA}$. Formally, we state that $\text{sam}_{\mathcal{D}}(\tau_{\mathcal{D}}, p_{\mathcal{D}}) = \mathcal{D}$.

We conducted six experiments for our augmentation function sam_{SSA}. In the first scenario, we randomly sampled x% of the SSA samples. This is equivalent to the 'Random Sampling Strategy' (R). To test the impact of parameter p_{SSA}, we experimented with the following percentages: 0% (no SSA samples); 25% (all original and a random 25% of the generated SSA samples);

Table 1: Content (IoU, Hal) and length (L) controllability, text quality (G), diversity (D-1, D-2, sC), and harmonic mean (H) of (IoU, G, and sC) for our CIC-BART-SSA models evaluated only on the original Flickr-Ent test set. Each of our CIC-BART-SSA models is trained on a different Flickr-Ent-SSA mixture, described by the augmentation strategy type τ_{SSA} and parameters p_{SSA}. Each blended data version has all the original data but different percentages of our SSA augmentations. The row order of experiments corresponds to the included SSA percentage in Flickr-Ent ranging from 0 to 100%.

Model: CIC-BART-SSA | Evaluation: Flickr-Ent Test Set									
τ_{SSA}	p_{SSA}	$H\uparrow$	IoU↑	G↑	sC↑	L↓	Hal↓	D-1↑	D-2↑
R	0%	69.8	54.0	85.0	78.6	1.24	36.5	43.6	58.2
R	25%	69.9	53.7	85.1	79.8	1.29	36.5	44.7	59.5
R	50%	70.3	53.9	85.6	80.5	1.23	36.2	45.3	60.6
U	10	70.6	54.3	85.6	80.5	1.07	35.6	45.9	61.0
R	75%	70.5	53.9	85.5	81.1	**1.05**	35.9	46.2	61.7
R	100%	**71.3**	**55.0**	**86.0**	**81.7**	**1.05**	**34.1**	**47.0**	**62.6**

Table 2: Content (IoU, Hal) controllability, text quality (G), diversity (D-1, D-2, sC), and harmonic mean (H) of (IoU, G, and sC) for our CIC-BART-SSA models evaluated only on the SSA data generated using Flickr-Ent test set. Each of our CIC-BART-SSA models is trained on a different Flickr-Ent-SSA mixture, described by the augmentation strategy type τ_{SSA} and parameters p_{SSA}. Each blended data version has all the original data but different percentages of our SSA augmentations. The row order of experiments corresponds to the included SSA percentage in Flickr-Ent-SSA ranging from 0 to 100%.

Model: CIC-BART-SSA | Evaluation: SSA Test Set								
τ_{SSA}	p_{SSA}	$H\uparrow$	IoU↑	G↑	sC↑	Hal↓	D-1↑	D-2↑
R	0%	68.5	53.0	80.5	79.8	37.3	52.9	62.9
R	25%	70.5	54.4	81.4	84.0	35.3	55.5	67.2
U	10	71.3	55.4	82.6	83.7	33.9	56.1	67.4
R	50%	71.5	55.2	82.7	85.1	33.9	56.4	68.3
R	75%	71.6	55.4	82.8	84.8	33.3	56.4	68.7
R	100%	**72.0**	**55.6**	**82.9**	**86.1**	**33.0**	**56.5**	**69.3**

50%; 75%; and 100% (all original and all generated SSA samples). Additionally, we conducted an experiment using the 'Uniform Coverage Sampling Strategy' (U) for τ_{SSA}, with p_{SSA} set to ten (10) uniform coverage bins ($p_{SSA} = \{[0\%, 10\%), [10\%, 20\%), \ldots, [90\%, 100\%]\}$). We note that the 'Uniform Coverage Sampling Strategy' contains approximately the same number of SSA samples as the 50% random sampling strategy.

We repeat the procedure for both the COCO-Ent-SSA and Flickr-Ent-SSA datasets. We evaluate all models on content (IoU, Hal) and length (L) controllability, text quality (G), diversity (sC, D-1, D-2), and the harmonic mean (H) of IoU, G, and sC. We present the evaluation results for the COCO-Ent and Flickr-Ent test sets in Tabs. 1 and 3. We observe a similar trend in both datasets where adding our SSA samples improves context and length controllability, text quality, and diversity. Our significant improvement in diversity and length controllability is due to the linguistic diversity offered by our SSA augmentations. For example, in Fig. 4 caption length histogram, we can see how narrow it is for

Table 3: Content (IoU, Hal) and length (L) controllability, text quality (G), diversity (D-1, D-2, sC) and harmonic mean (H) of (IoU, G, and sC) for our CIC-BART-SSA models evaluated only on the original COCO-Ent test set. Each of our CIC-BART-SSA models is trained on a different COCO-Ent-SSA mixture, described by the augmentation strategy type τ_{SSA} and parameters p_{SSA}. Each blended data version has all the original data but different percentages of our SSA augmentations. The row order of experiments corresponds to the included SSA percentage in COCO-Ent-SSA ranging from 0 to 100%.

Model: CIC-BART-SSA | Evaluation: COCO-Ent Test Set									
τ_{SSA}	p_{SSA}	$H\uparrow$	IoU↑	G↑	sC↑	L↓	Hal↓	D-1↑	D-2↑
R	0%	75.9	76.2	73.0	78.7	.490	19.0	38.0	56.2
R	25%	76.9	76.5	74.6	80.1	.148	18.5	42.6	59.7
R	50%	76.8	76.7	74.9	79.0	.163	**17.8**	42.9	59.0
U	10	76.7	77.0	74.0	79.4	.150	**17.8**	42.0	58.6
R	75%	77.8	**77.2**	74.0	**82.6**	.116	**17.8**	42.9	61.6
R	100%	**78.3**	**77.2**	**74.8**	82.5	**.106**	**17.8**	**44.6**	**63.2**

Table 4: Content (IoU, Hal) controllability, text quality (G), diversity (D-1, D-2, sC), and harmonic mean (H) of (IoU, G, and sC) for our CIC-BART-SSA models evaluated only on the SSA data generated using COCO-Ent test set. Each of our CIC-BART-SSA models is trained on a different COCO-Ent-SSA mixture, described by the augmentation strategy type τ_{SSA} and parameters p_{SSA}. Each blended data version has all the original data but different percentages of our SSA augmentations. The row order of experiments corresponds to the included SSA percentage in COCO-Ent-SSA ranging from 0 to 100%.

Model: CIC-BART-SSA | Evaluation: SSA Test Set								
τ_{SSA}	p_{SSA}	$H\uparrow$	IoU↑	G↑	sC↑	Hal↓	D-1↑	D-2↑
R	0%	69.2	61.4	73.9	74.0	28.4	44.2	57.0
R	25%	74.4	65.1	80.1	80.0	**23.0**	50.1	63.1
U	10	74.6	65.0	80.3	80.8	23.2	51.2	64.4
R	50%	74.9	64.9	**80.7**	81.7	23.1	51.7	64.3
R	75%	74.9	64.9	**80.7**	81.6	23.1	51.7	65.1
R	100%	**75.6**	**65.2**	**80.7**	**83.7**	23.2	**53.8**	**67.8**

COCO-Ent dataset which mainly contains captions with 11, 12, or 13 words, so it is difficult for models trained on just COCO-Ent to generalize and generate captions of other lengths. On the contrary, our models trained jointly with our SSA augmentations can generate captions faithful to the length control signal.

In Tabs. 2 and 4, we have evaluated the performance of each model on the SSA augmentations from the COCO-Ent and Flickr-Ent testing images, respectively. We have excluded the length (L) controllability as it is the same as in Tabs. 1 and 3. This is because it is computed on random control signals of each dataset testing images.

We observe an even more evident improvement across all metrics as we progressively include more of our focused SSA examples for CIC model training. This results from our SSA augmentations, which provide focused examples for training controllable image captioning models. This is exemplified in Figs. 4

and 5 coverage histograms, where the low coverage (high focus) regime is highly under-represented in the original COCO-Ent and Flickr-Ent datasets. Finally, our quantitative analysis demonstrates that training with our SSA augmentations improves controllability, text quality, and diversity performance. Particularly, the improvement is remarkable in cases where the CIC models need to focus and describe a specific, small region of a complex and large scene.

4.3 Effect of SSA on Content Controllability

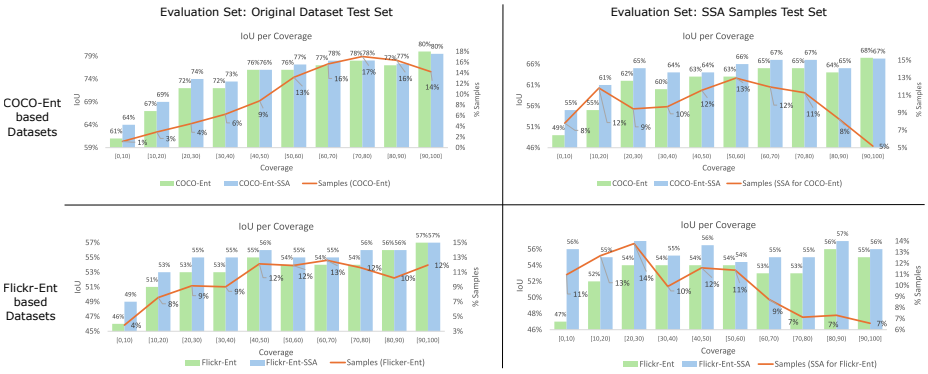

Fig. 6: Content controllability (IoU) performance of CIC-BART when trained with and without SSA. The abscissa is the % of the image covered by the control signal, so the left and right parts of the graph represent more focused and broader control signals, respectively. The %Samples curve represents the distribution of Flickr-Ent test images in each coverage interval. The results show that SSA plays a crucial role in boosting CIC-BART performance in data-deprived, focused CIC settings.

In this section, we present the performance of our models, namely CIC-BART and CIC-BART-SSA, with regards to IoU (Intersection over Union) analysis. The former is trained on the original datasets, COCO-Ent and Flickr-Ent, while the latter is trained on our proposed datasets, COCO-Ent-SSA and Flickr-Ent-SSA. We break down the content controllability (IoU, Hal) performance on coverage bands, where coverage refers to the percentage of the image covered by the control signal.

In Fig. 6, we show the results of our models on different evaluation test sets. The first row of the figure represents the performance of models trained on either COCO-Ent or COCO-Ent-SSA datasets, while the second row represents models trained on Flickr-Ent or Flickr-Ent-SSA. The two columns describe the evaluation test sets. In the first column, we evaluate the models on the original datasets (COCO-Ent, Flickr-Ent) test sets. In contrast, in the second column, we evaluate our SSA augmentations derived from the test sets of the original datasets.

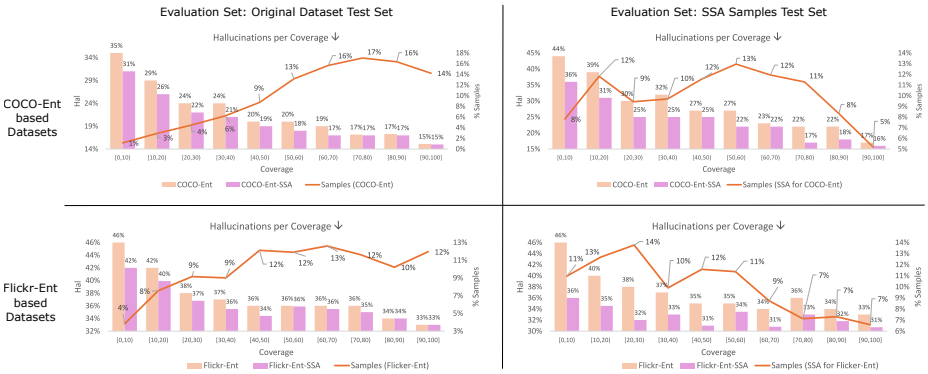

Fig. 7: Hallucinating Nouns (Hal) performance of CIC-BART when trained with and without SSA. The abscissa is the % of the image covered by the control signal, so the left and right parts of the graph represent more focused and broader control signals, respectively. The %Samples curve represents the distribution of Flickr-Ent test images in each coverage interval. The results show that SSA plays a crucial role in boosting CIC-BART performance in data-deprived, focused CIC settings.

Table 5: Best-5 Diversity for randomly generated control signals of the COCO-Ent and Flickr-Ent testing images. Our model CIC-BART was trained on the original COCO-Ent and Flickr-Ent datasets while CIC-BART-SSA was trained with our COCO-Ent-SSA and Flickr-Ent-SSA augmented datasets. **ASG-type dataset was not released for Flickr-Ent, precluding us from evaluating its best-5 diversity scores.*

Method	D-1↑	D-2↑	D-1↑	D-2↑
	COCO-Ent		Flickr-Ent	
ASG [6]	43	56	-	-
CIC-BART	58	86	67	90
CIC-BART-SSA	**67**	**92**	**68**	**93**

The orange line in all plots represents the percentage of examples in each coverage band. We observe that the test sets of the original datasets, COCO-Ent and Flickr-Ent, have more examples with control signals covering a broad aspect of the image. In contrast, the SSA test sets have more samples for focused control signals covering a small percentage of the image. We also notice that the test sets of the original and SSA datasets are consistent with their respective training set statistics presented in Figs. 4 and 5.

Furthermore, in Fig. 7, we present the corresponding coverage histograms for our Hallucinations (Hal) metric. We observe a similar trend in all cases, wherein our model CIC-BART-SSA, trained on our SSA augmentations, shows an improvement in content controllability performance. This means higher IoU and reduced Hal. We note that breaking down the content controllability metrics in coverage bands reveals the major improvement in the low coverage regions, where the original datasets, COCO-Ent and Flickr-Ent, have very few data points.

Table 6: Captioning metrics on COCO-Ent and Flickr-Ent original test sets.

Model	B4↑	M↑	R↑	C↑	S↑	B4↑	M↑	R↑	C↑	S↑
	COCO-Ent					Flickr-Ent				
LaBERT [9]	13.5	20.6	42.3	136.6	32.4	8.1	14.6	32.7	70.8	19.6
SCT [8]	22.3	25.6	55.3	209.7	48.5	12.5	16.8	38.9	84.0	23.5
ComPro [19]	24.0	27.3	56.1	232.2	50.4	11.9	17.3	37.8	89.4	23.9
ASG [6]	23.0	24.5	50.1	204.2	42.1	-	-	-	-	-
VSR [5]	25.4	28.8	57.8	265.0	49.8	12.3	19.8	40.9	131.4	22.4
CIC-BART	21.0	26.2	50.2	225.0	46.3	14.2	19.4	39.7	136.4	27.2
CIC-BART-SSA	20.0	25.5	48.9	216.2	46.1	13.0	18.9	37.8	123.3	27.0
CIC-BART +verb	**36.2**	**33.7**	**62.9**	**366.8**	**53.7**	**26.6**	**27.2**	**53.9**	**275.1**	**32.4**

4.4 Best-5 Diversity

In Tab. 5, we present the best-5 D-1, D-2 diversity for our models CIC-BART and CIC-BART-SSA, which was proposed in ASG [6]. We notice an important diversity improvement, especially for the COCO-Ent dataset, when we train our models using our SSA augmentations (CIC-BART-SSA).

4.5 Measuring Performance via Standard Captioning Metrics

Tab. 6 reports the results of all models in standard captioning metrics (i.e., B4, M, R, C and S). As we can see, both CIC-BART and CIC-BART-SSA perform comparably to the three baselines with respect to these metrics. Nonetheless, as we noted earlier, these scores reflect the match between a generated controlled caption and a ground-truth image-level caption. Given a focused control signal (e.g., one focusing on a subset of entities in an image), we expect a partial match between the generated controlled caption and the ground-truth caption. VSR has the best scores for most of these metrics, but this is partially due to this model using the exact verb as the control signal and is not necessarily an indicator of this model's caption quality (as we saw earlier with the low G score). To understand the role of such descriptive control signals, we present results for a variation of our CIC-BART where we also input the verb as an additional control signal; see the last row of the table (CIC-BART +verb). Note that even with this additional information, our control signal is still simpler than that of the VSR, as we do not provide the verb-specific semantic roles. Nevertheless, by adding a verb as the control signal, we can see a substantial increase in all standard captioning metrics.

4.6 Qualitative Results

In Fig. 8, we present qualitative examples from COCO-Ent and Flickr-Ent test sets. In these examples, the control signals are extracted from the ground-truth captions. Each colored oval under 'Cntl' corresponds to a bounding box of the same color in the image. The collection of ovals identifies the entities of interest, that is, the control signal. (Note that we do not show the colored ovals for image (c), since both sets of control signals include all bounding boxes.) For example, in

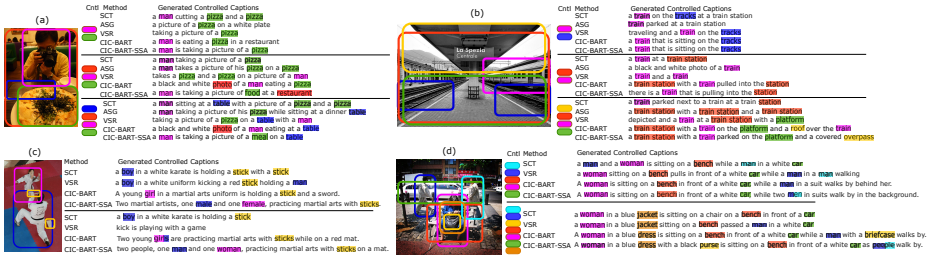

Fig. 8: Qualitative examples of generated controllable captions from COCO-Ent (images (a) and (b)) and Flickr-Ent (images (c) and (d)) test set. Images (a) `https://farm5.staticflickr.com/4102/4888234256_538b8dee56_z.jpg` and (b) `https://farm9.staticflickr.com/8501/8308004994_44eb2d562d_z.jpg` are licensed under a Creative Commons CC BY-SA 2.0 `https://creativecommons.org/licenses/by-sa/2.0/`; (c) `flickr.com/photo.gne?id=101362133` and (d) `flickr.com/photo.gne?id=151970521` are licensed under a Creative Commons CC BY 2.0 `https://creativecommons.org/licenses/by/2.0/`.

(a), the first control signal (at the top) focuses on the regions *restaurant*, *man*, and *food*, while the second control signal also includes the entity *table*. Each highlighted word in the generated controlled captions corresponds to the control entity of the same color, showing the match between the generated captions and the control signal.

We notice that our models outperform previous SOTA by substantially improving the quality of the generated controlled captions. This behavior was expected from our quantitative analysis, showing that our models have significantly higher text quality (G) performance. In addition, more evidently in figures (b) and (d), our models have better content controllability performance by correctly referring to all entities of interest in the control signal. Especially in the highly challenging, complex scene (d) in which many objects are present, it successfully describes all entities of interest in the generated controllable caption.

Our CIC-BART-SSA model generates captions that are faithful to the control signal and better understand the relationships connecting the entities of interest. For example, in (a), it correctly identifies that the person is photographing his food rather than eating it and that the image is not black and white, or in (d) that the woman holds the purse and not the man in the background.

In Fig. 9 we present additional qualitative examples from COCO-Ent test set. We include the generated controllable captions from the baseline models (SCT, ASG, and VSR) and the proposed models CIC-BART and CIC-BART-SSA. Our qualitative examples also show that our models generate diverse, high-text-quality captions with improved content controllability when compared to the baseline models.

Next in Fig. 10 we present examples using the SSA augmentations control signals which are derived from Flickr-Ent test set. We notice that our CIC-BART-SSA better conveys the image concept without hallucinating. For example, in

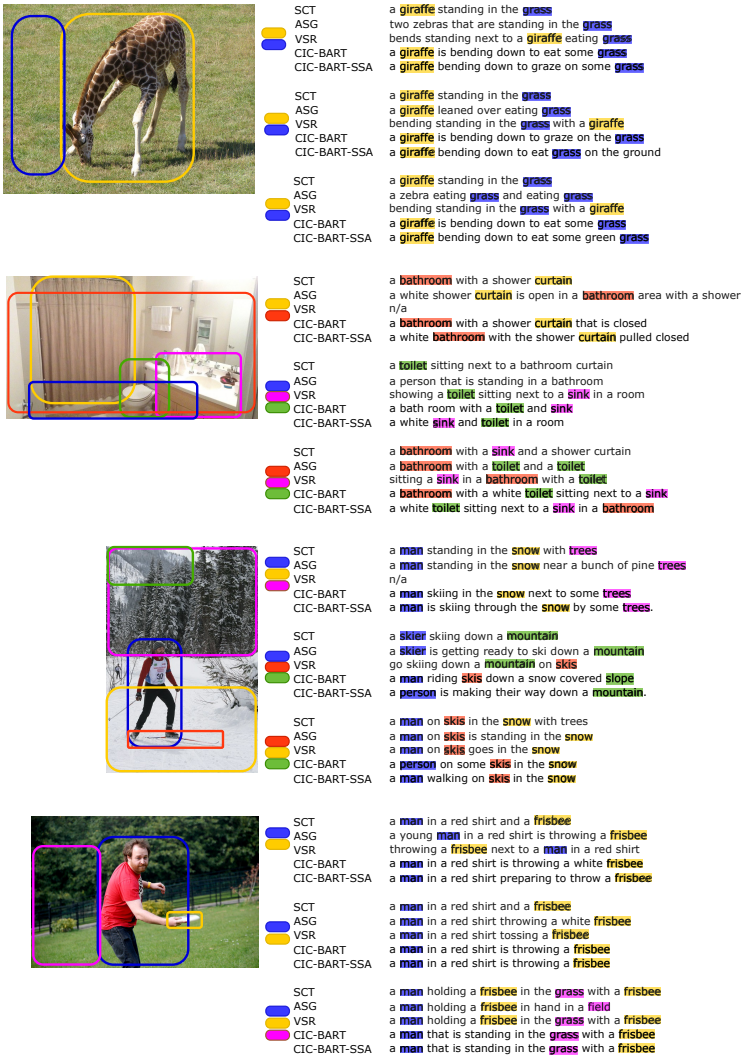

Fig. 9: Qualitative examples from COCO-Ent test set. We include the generated controlled captions of SCT, ASG, and VSR methods and our proposed CIC-BART and CIC-BART-SSA models. CIC-BART was trained using COCO-Ent training set, whereas CIC-BART-SSA used COCO-Ent-SSA training set. Images `http://images.cocodataset.org/train2017/000000001448.jpg`, `http://images.cocodataset.org/train2017/000000281019.jpg` and `http://images.cocodataset.org/val2017/000000325991.jpg` are licensed under a Creative Commons CC BY-SA 2.0 `https://creativecommons.org/licenses/by-sa/2.0/` and image `http://images.cocodataset.org/val2017/000000038210.jpg` licensed under a Creative Commons CC BY 2.0 `https://creativecommons.org/licenses/by/2.0/`.

Fig. 10: Qualitative examples from our SSA test set constructed from Flickr-Ent. CIC-BART was trained using Flickr-Ent training set, whereas CIC-BART-SSA used Flickr-Ent-SSA training set. Images `https://www.flickr.com/photos/thunderchild5/183647966/`, `flickr.com/photo.gne?id=151970521` and `flickr.com/photo.gne?id=101362133` are licensed under a Creative Commons CC BY 2.0 `https://creativecommons.org/licenses/by/2.0/`, and image `flickr.com/photo.gne?id=7249763658` under a Creative Commons PDM 1.0 `https://creativecommons.org/publicdomain/mark/1.0/`.

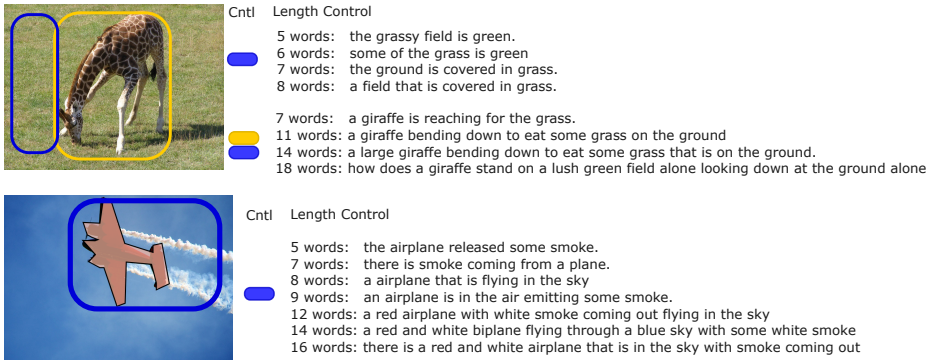

Fig. 11: Qualitative examples for various length control values given a specific image sub-region. The generated controllable captions are from our trained CIC-BART-SSA model. Image `http://images.cocodataset.org/train2017/000000001448.jpg` is licensed under Creative Commons CC BY-SA 2.0 `https://creativecommons.org/licenses/by-sa/2.0/` and `http://images.cocodataset.org/train2017/000000115178.jpg` under a Creative Commons CC BY 2.0 `https://creativecommons.org/licenses/by/2.0/`.

the second image, it correctly describes that the man is fixing the ticket booth or that the woman carries a bag and is not walking.

Further, we conducted an experiment to evaluate our length control performance qualitatively. In the experiment, we generated controllable captions for a fixed image region and various caption length controls. We present some of our qualitative results in Fig. 11 showing that the generated captions were faithful to the length control signal, indicating that our model is effective in controlling the length of captions. Furthermore, we observe that our model generates a diverse set of captions for a specific image region.

4.7 SSA vs Other Augmentation Strategies

Augmentations via LLM Paraphrasing To understand the impact of our SSA enhancements, we perform an experiment in which we augment the original training data with paraphrases generated using an LLM, Llama-2 [16]. We generate one paraphrase per original caption, effectively doubling the size of the training data[5]. Specifically, we instructed the Llama-2 model to rephrase the initial captions using few shot prompting like

> *If the phrase 'Children wearing team uniforms playing soccer in a grassy field' can be paraphrased as 'Kids in a grassy field playing soccer in uniforms', and the phrase 'A little girl sitting in the middle of a restaurant and smiling for picture' can be paraphrased as 'A smiling little girl taking a picture while sitting in a restaurant', then the phrase '{caption}' can be paraphrased as ...*

[5] For 20% of the captions, Llama generates paraphrases identical to original sentences.

We replace {caption} with original dataset captions, relying on Llama-2 to paraphrase them. Since we only paraphrased the original sentence, we can assume that it pertains to the same set of bounding boxes since it refers to the same visual entities of interest, which is the only information required for our CIC-BART model.

The results in Tab. 7 (bottom panel) demonstrate that CIC-BART-SSA outperforms CIC-BART-par on all controllable captioning metrics. We conclude that the improved performance of CIC-BART-SSA is not just due to the increase in training data size; the model benefits from the intricate structured and visually grounded guidance of our SSA.

Table 7: Performance of our SSA augmentations (CIC-BART-SSA) vs LLM paraphrases (CIC-BART-par). All models are evaluated on the original COCO-Ent test sets.

Model	$H\uparrow$	IoU↑	Hal↓	G↑	sC↑	D-1↑	D-2↑	L↓
CIC-BART-par	74.3	76.2	18.8	72.0	74.9	36.1	52.7	.19
CIC-BART-SSA	**78.3**	**77.2**	**17.8**	**74.8**	**82.5**	**44.6**	**63.2**	**.11**

In Figs. 12 to 14, we present some examples of the Llama-2 paraphrases of the original COCO captions. For example, the paraphrase of 1-O) is 1-Llama-2), the paraphrase of 2-O) is 2-Llama-2), and so on. We have included the SSA-generated focused captions for each image, along with the corresponding synthetic caption and its GRUEN score. SSA uses this metric to filter out poor-quality sentences (in our experiments, we set the GRUEN threshold to 0.7).

Scene Graphs with LLM Paraphrasing Augmentations The scope of this section is to illustrate the benefits of AMRs when compared to scene graphs for CIC augmentation. Figs. 14, 16 and 18 depict examples from COCO-Ent and Flickr-Ent that contrast Original Captions, LLM-paraphrased captions, SSA-augmented captions, and CLID-augmented captions.

Nature of captured relations and entities in AMR vs. Scene Graph representations. As stated in our main paper, prior analysis [1,20] has shown that existing scene graph annotations focus mainly on geometric and possessive relations. For example, in Fig. 15 examples of geometric relations are 'a man in front of a door', 'one man next to another man,' etc., and possessive 'a man has hair,' 'the man has a head' etc. Regarding entities, scene graphs focus mainly on object/body parts (hair, head, arm, etc.) and clothing (dress, shirt, etc.). On the contrary, the AMRs derived from the image captions contain a wide range of semantic relations that are inherited from the natural language image descriptions drawn from the image captioning datasets. For example, in Fig. 15, in the original captions, we can find the semantic relations, 1-O) 'the men are *hanging out* in the yard', 5-O) 'the friends *enjoy time spent* together' which will be inherited

Original Captions (COCO Entities, ID 101155)

1-O) A sheep and its lamb in a fenced grass enclosure.

2-O) a couple of sheep stand in front of a rock

3-O) Young lambs with adult in fenced grassy area.

4-O) Sheep are grazing in the fenced in area.

5-O) A lamb is standing next to a rock.

Llama-2 paraphrases.

1-Lllama-2) A sheep and its lamb in a fenced enclosure

2-Lllama-2) A couple of sheep standing in front of a rock

3-Lllama-2) Young lambs with an adult in a fenced grassy area.

4-Lllama-2) A fenced in area is grazing sheep.

5-Lllama-2) A rock is standing next to a lamb.

SSA Captions with GRUEN score.

1-SSA) sheep are roaming in a grassy area. GRUEN 0.90

2-SSA) a couple of sheep are standing in front GRUEN 0.89

3-SSA) a young lamb is standing next to a rock. GRUEN 0.82

4-SSA) the grassy area is free of trees. GRUEN 0.77

5-SSA) a lamb is standing next to a sheep. GRUEN 0.74

6-SSA) a couple of sheep standing in front of rocks. GRUEN 0.55

CLID Captions

1-CLID) There is a hill, a rock, a fence, an animal and a sheep.

2-CLID) There is a fence, a sheep, a hill, a dog, a cow, and a head of a sheep.

Fig. 12: Qualitative examples from different augmentation strategies. We present the original five dataset captions, their Llama-2 paraphrases, our SSA (AMR-based) generated descriptions, and finally, CLID (scene-graph-based) augmentations. For the SSA captions, their GRUEN score is included, which is used to filter out poorly generated sentences. *The captions are from the image with ID 101155 from the COCO-Ent dataset. The image is not depicted here due to license limitations.*

Original Captions (COCO Entities, ID 101194)

1-O) A young woman is pulling a casserole out of the oven.

2-O) A cook removes a hot dish from the oven.

3-O) A woman pulling her food out of the oven.

4-O) A person wearing mitts reaching into an over.

5-O) A woman bending over and reaching into an oven.

Llama-2 paraphrases.

1-Llama-2) A young woman is removing a casserole from the oven.

2-Llama-2) A cook removing a hot dish from the oven.

3-Llama-2) A woman taking food out of the oven.

4-Llama-2) A person reaching into an over with mitts on.

5-Llama-2) A woman bending over and reaching into an oven.

SSA Captions with GRUEN score.

1-SSA) a person is removing a dish from an oven. GRUEN 0.89

2-SSA) a person is removing a casserole from the oven. GRUEN 0.87

3-SSA) a woman is reaching over to an oven GRUEN 0.85

4-SSA) a young woman is bent over GRUEN 0.79

5-SSA) a person is reaching over into an oven. GRUEN 0.72

CLID Captions

1-CLID) A woman is removing a good food from an oven and a blue on an oven mitt with flowers on a floral white oven mitt and a top of a stove and a flower on the floral oven mitt and a full dish in the open oven.

2-CLID) A woman wearing an oven mitt with flowers on a hot full dish in an open oven and an oven mitt with a blue on it and an open door on the hot oven has a hot food being removed from the black oven.

3-CLID) A woman is reaching in an oven removing a good hot food and a colorful oven mitt with flowers and a filled full dish in the oven and a door on the oven and a left oven mitt on her and a dark blue patch on an oven mitt.

4-CLID) A woman in an oven mitt with flowers and an open door is removing a hot food item from the oven while wearing a colorful oven mitt.

5-CLID) A woman wearing a colorful white oven mitt and a right oven mitt is removing a good hot food from a white oven and flowers from a dish in the open hot oven.

6-CLID) A woman wearing a white colorful oven mitt with flowers on an open black oven and an open door of the oven and a hot food being removed from the oven and a full dish in the open hot black oven and a right oven mitt on her is standing in front of the oven.

7-CLID) An open door on an oven, an oven mitt with flowers, a woman wearing a colorful oven mitt, and a good hot food in the oven, a hot filled dish in the oven, and a left oven mitt on the woman were all present.

8-CLID) A woman wearing an oven mitt removes a hot dish in a black white oven and a good food being removed from the hot open black oven and a flower on a floral oven mitt with flowers and an open door on the black white oven and an oven mitt with the flowers and a tray in the

9-CLID) A woman is reaching in a white oven removing a hot food and a blue on an oven mitt with flowers on a colorful white floral oven mitt.

10-CLID) A woman in an oven mitt with flowers and a black hot oven mitt with a silver tray and an open door is removing a hot food item from the oven.

Fig. 13: Qualitative examples from different augmentation strategies. We present the original five dataset captions, their Llama-2 paraphrases, our SSA (AMR-based) generated descriptions, and finally, CLID (scene-graph-based) augmentations. For the SSA captions, their GRUEN score is included, which is used to filter out poorly generated sentences. *The captions are from the image with ID 101194 from the COCO-Ent dataset. The image is not depicted here due to license limitations.*

Original Captions (COCO Entities, ID 101310)

1-O) there is a airplane being boarded at a air port

2-O) An aircraft that is Parked at the terminal ready to be boarded.

3-O) A plane at the airport with the air bridge pulled up to it.

4-O) A airplane that is sitting on the runway outside of a terminal.

5-O) An airplane is parked at a jet way at an airport.

Llama-2 paraphrases.

1-Llama-2) There is a plane being boarded at an airport

2-Llama-2) An aircraft that is waiting to be boarded.

3-Llama-2) A plane with the air bridge pulled up to it.

4-Llama-2) A plane sitting on a runway outside a terminal.

5-Llama-2) An airplane parked at a jet way at an airport

SSA Captions with GRUEN score.

1-SSA) an airplane is being pulled into the airport. GRUEN 0.91

2-SSA) a plane is pulled up to the airport. GRUEN 0.90

3-SSA) an aircraft is parked at a terminal. GRUEN 0.88

4-SSA) the aircraft is ready for the boarding. GRUEN 0.82

5-SSA) an aircraft is boarding at a jet port. GRUEN 0.82

6-SSA) a plane is ready to board at an air port. GRUEN 0.81

7-SSA) a plane is sitting on the runway GRUEN 0.75

8-SSA) an airplane parked in a jet way at an airport. GRUEN 0.73

9-SSA) an airplane sitting on a runway outside of a terminal. GRUEN 0.69

10-SSA) an airport has a jet bridge by it. GRUEN 0.57

11-SSA) an airport has a bridge by it. GRUEN 0.55

CLID Captions

1-CLID) An airplane parked at a gate on a tarmac and a black backpack by an airport chair are in a distance from the tarmac.

2-CLID) There is a white airplane parked at a gate at an airport and buildings across from a tarmac in a distance.

3-CLID) There is a white airplane parked at a gate ready for a boarding and a black backpack next to a chair by an airport and a van on a tarmac and silver steps across from the tarmac.

4-CLID) A white airplane is parked at a gate at an airport and buildings are a distance from a tarmac.

5-CLID) A white airplane ready for a boarding at an airport, a brown suitcase in the airport, a black backpack next to a chair by the airport, a silver step ladder on the tarmac, and a white van on the tarmac.

6-CLID) A black backpack next to a chair by an airport and an airplane parked at a gate at the airport and buildings in a distance across from a tarmac and a conveyor belt leading to the airplane and a van on the tarmac.

7-CLID) An airplane on a tarmac at an airport and buildings across from the tarmac, a black backpack by a chair by the airport, and a van on the tarmac.

8-CLID) A white airplane at an airport on a tarmac and an airplane at the airport and a black backpack by a chair by the airport and buildings in a distance across from the tarmac and a conveyor belt leads to the white airplane.

9-CLID) There is a black backpack next to a chair by an airport, a white airplane parked at a gate, a brown suitcase in the airport, and a white van on the tarmac.

10-CLID) There are buildings across from a tarmac and a brown suitcase in an airport and a white airplane on the tarmac.

Fig. 14: Qualitative examples from different augmentation strategies. We present the original five dataset captions, their Llama-2 paraphrases, our SSA (AMR-based) generated descriptions, and finally, CLID (scene-graph-based) augmentations. The GRUEN score is included for the SSA captions, which is used to filter out poorly generated sentences. Image `http://images.cocodataset.org/train2017/000000101310.jpg` is licensed under a Commons Creative CC BY-SA 2.0 `https://creativecommons.org/licenses/by-sa/2.0/`.

Original Captions (Flickr-30k Entities, ID 1000092795)

1-O) Two young guys with shaggy hair look at their hands while hanging out in the yard.

2-O) Two young, White males are outside near many bushes.

3-O) Two men in green shirts are standing in a yard.

4-O) A man in a blue shirt standing in a garden.

5-O) Two friends enjoy time spent together.

SSA Captions with GRUEN score.

1-SSA) a man in a blue shirt with shaggy hair wears a jersey. GRUEN 0.91

2-SSA) two men who are spending time together. GRUEN 0.84

3-SSA) the two men are enjoying the time. GRUEN 0.84

4-SSA) two shaggy haired men, one in a blue shirt, spending time together. GRUEN 0.83

5-SSA) two shaggy haired men, one in a blue shirt, enjoying time. GRUEN 0.77

6-SSA) two men who are near each other outside. GRUEN 0.75

7-SSA) two shaggy haired men, one in a blue shirt, looking at hands while hanging out in a yard. GRUEN 0.69

8-SSA) two men with shaggy hair, one in a blue shirt, near many bushes. GRUEN 0.64

9-SSA) two men in a field wearing jersies. GRUEN 0.61

10-SSA) one man in a blue shirt standing in a yard. GRUEN 0.58

CLID Captions

1-CLID) A man wearing a shirt and a pant is standing in front of a building, behind a woman and a door, and with a hair on his head.

2-CLID) A person and a man are standing in front of a building with a tree behind them and a girl with an arm and a hair on a head.

3-CLID) A man in a shirt and pants is standing in front of a building with a man and a woman in front of a tree behind him.

Fig. 15: Qualitative examples from different augmentation strategies. We present the original five dataset captions, our SSA (AMR-based) generated descriptions, and finally, CLID (scene-graph-based) augmentations. For the SSA captions, their GRUEN score is included, which is used to filter out poorly generated sentences. *The captions are from the image with ID 1000092795 from the Flickr-Ent dataset. The image is not depicted here due to license limitations.*

Original Captions (Flickr-30k Entities, ID 1001573224)

1-O) Five ballet dancers caught mid jump in a dancing studio with sunlight coming through a window.

2-O) Ballet dancers in a studio practice jumping with wonderful form.

3-O) Five girls are leaping simultaneously in a dance practice room.

4-O) Five girls dancing and bending feet in ballet class.

5-O) A ballet class of five girls jumping in sequence.

SSA Captions with GRUEN score.

1-SSA) the people in the studio are dancing a ball. GRUEN 0.82

2-SSA) there is a person that is dancing the ballet. GRUEN 0.80

3-SSA) there are people that are dancing a ball. GRUEN 0.78

4-SSA) a person in a studio is dancing ballet. GRUEN 0.71

5-SSA) a person in a studio bending over with a foot. GRUEN 0.61

6-SSA) people in a studio practice form catching in mid jump in a ball studio simultaneously. GRUEN 0.52

7-SSA) sunlight coming through the window in a ball studio. GRUEN 0.50

8-SSA) people in a studio jumping simultaneously in a ball studio. GRUEN 0.47

9-SSA) people in a studio in mid jump in a ball studio simultaneously. GRUEN 0.39

CLID Captions

1-CLID) There are people with a leg and a building, a girl with an arm and a leg, and a woman with a head and a shirt.

2-CLID) There are people with a head and a girl and a woman with a head and a girl and a building.

3-CLID) A building and a door in a room with people and a girl and a girl and a shirt him.

Fig. 16: Qualitative examples from different augmentation strategies. We present the original five dataset captions, our SSA (AMR-based) generated descriptions, and finally, CLID (scene-graph-based) augmentations. The GRUEN score is included for the SSA captions, which is used to filter out poorly generated sentences. Image `https://www.flickr.com/photos/bombarosa/1001573224/` is under Creative Commons CC BY-ND 2.0 `https://creativecommons.org/licenses/by-nd/2.0/` license.

in their AMR representations and therefore in the SSA samples (i.e. 2-SSA), 3-SSA), 5-SSA), 7-SSA)).

Limitations of Scene Graph representations. Scene graphs are useful in capturing the visual elements of a scene and their relationships, such as geometric and possessive relationships. However, they lack the ability to represent abstract concepts like time-related information, such as 'a sunny morning' or 'a quiet afternoon'. For instance, in Fig. 18 caption 2-O), the phrase 'during a sunny afternoon' is not directly related to low-level semantics like individual visual objects and their relationships. Rather, it relates to higher-level reasoning, such as observing the shadows of the trees and the annotators' experiences that helped them conclude that the picture was taken during a sunny afternoon.

Another important difference between AMRs and scene graphs is that the edges of an AMR carry linguistic information, which is not the case with scene graphs. In Fig. 1, we can see only a fraction of the available edge roles, which are crucial when converting vgAMRs to natural language sentences as they help generate accurate descriptions. On the other hand, scene graph edges lack the ability to convey additional linguistic information, except for the edge direction, which indicates the object and subject of a relation. This lack of additional information makes it challenging to augment captions using sampled scene graphs and forces reliance on LLM paraphrasing, which, as previously discussed in Sec. 4.7, can introduce errors and inaccuracies (increased hallucinations (Hal) and poorer text quality (G) as seen in Tab. 7). For instance, in the scene graph-based aug-

Original Captions (Flickr-30k Entities, ID 1003420127)

1-O) A group of adults, inside a home, sitting on chairs arranged in a circle, playing a type of musical instruments.

2-O) Five musicians, a man and four women, practicing sheet music (using flutes) in a living room.

3-O) People gathered in a circle, some holding musical instruments.

4-O) People gathered in a room to talk about their favorite tunes.

5-O) Five people are sitting in a circle with instruments.

SSA Captions with GRUEN score.

1-SSA) the chairs are arranged in a circle. GRUEN 0.89

2-SSA) there are five musicians sitting in a circle. GRUEN 0.88

3-SSA) there are five musicians practicing music. GRUEN 0.88

4-SSA) the five musicians are talking about the tune. GRUEN 0.87

5-SSA) five musicians, four women and one man, are talking about a tune. GRUEN 0.86

6-SSA) there are five musicians that are playing instruments GRUEN 0.86

7-SSA) there is a man sitting inside on a chair GRUEN 0.85

8-SSA) five musicians, four women and one man, practicing sheet music in a living room. GRUEN 0.85

9-SSA) there are some people that are holding instruments GRUEN 0.84

10-SSA) some people are holding a sheet music instrument GRUEN 0.84

11-SSA) five musicians, one man and four women, are all in favor of the tune. GRUEN 0.83

12-SSA) the instruments are being used for practice. GRUEN 0.83

13-SSA) five musicians are in favor of a tune. GRUEN 0.82

14-SSA) five musicians, four women and one man, sitting in a circle. GRUEN 0.80

15-SSA) five musicians, one man and four women in a living room practicing sheet music instruments. GRUEN 0.75

16-SSA) five musicians, one man and four women playing musical instruments. GRUEN 0.69

17-SSA) a gathering of five tune-talkers, one man, four women, in a room. GRUEN 0.63

18-SSA) a chair sitting inside of a home. GRUEN 0.52

CLID Captions

1-CLID) A man wearing a shirt on a shirt and a chair with a pillow in a room with people and a woman sitting on a chair with a shirt and a lamp behind them.

2-CLID) A man wearing a shirt and a woman sitting on a chair in a room with a paper and a hair and a chair with a pillow on the floor.

3-CLID) A man wearing a shirt on a chair in a room with a man and a woman sitting on a chair with a pillow on the room and a woman sitting on a chair with a lamp and a head in the window.

Fig. 17: Qualitative examples from different augmentation strategies. We present the original five dataset captions, our SSA (AMR-based) generated descriptions, and finally, CLID (scene-graph-based) augmentations. For the SSA captions, their GRUEN score is included, which is used to filter out poorly generated sentences. *The captions are from the image with ID 1003420127 from the Flickr-Ent dataset. The image is not depicted here due to license limitations.*

Original Captions (Flickr-30k Entities, ID 151970521)

1-O) Woman dressed in blue sitting on a bench with a cane next to her looking through her purse with a stretch limo behind her with a man and a woman walking in opposite directions from one another .

2-O) A woman in blue looks in a black leather bag while sitting on a bench during a sunny afternoon while people and limousines passed behind her .

3-O) An old woman is sitting on a bench, while behind her a limo pulls up and two people in white are walking by .

4-O) Decorated limo passing by elderly woman waiting on park bench in Chinese street .

5-O) A woman in a blue shirt is looking in her change purse .

SSA Captions with GRUEN score.

1-SSA) a woman is dressed up in a blue shirt. GRUEN: 0.83

2-SSA) one woman is waiting on a bench.GRUEN0.83

3-SSA) one woman is sitting on a bench.GRUEN0.79

4-SSA) one woman is looking at a bag GRUEN: 0.72

5-SSA) a woman is dressed up in a shirt GRUEN: 0.71

6-SSA) a woman is looking in a purse GRUEN: 0.69

7-SSA) one woman in a blue shirt waiting on a bench on a chinese street. GRUEN: 0.68

8-SSA) one woman in a blue shirt sitting on a bench on a chinese street on a sunny afternoon. GRUEN: 0.65

9-SSA) one woman in a blue shirt looking through a purse. GRUEN: 0.54

10-SSA) one woman in a blue shirt looks like a leatopposite direction bag as two men and a limo pass behind her in the direction. GRUEN: 0.59

11-SSA) a woman in a blue shirt walks by with a limo behind in the leatopposite direction and one woman in a blue shirt looking through a purse and sitting on a bench on a chinese street in the sunny afternoon. GRUEN: 0.52

12-SSA) two men and a limo passing behind each other in the leatopp opposite direction. GRUEN: 0.43

13-SSA) a woman sits on a bench on a sunny afternoon while a limo is behind in the opposite direction and pulling up. GRUEN: 0.39

14-SSA) a purse has a limo behind it in the leatopposite direction. GRUEN: 0.37

CLID Captions

None

Fig. 18: Qualitative examples from different augmentation strategies. We present the original five dataset captions and SSA (AMR-based) generated descriptions. The GRUEN score is included for the SSA captions, which is used to filter out poorly generated sentences. The CLID dataset does not provide additional annotations for this image. Image `flickr.com/photo.gne?id=151970521` is under a Creative Commons CC BY 2.0 `https://creativecommons.org/licenses/by/2.0/` license.

mentation of CLID [10], they first place the sampled scene graph node names in a sequence and then ask an LLM paraphraser to generate a sentence.

In Figs. 12 to 18, we can see instances from our SSA (AMR-based) and CLID (scene graph-based) augmentations for a particular image. Augmentations in the CLID dataset are not visually grounded and, hence, cannot be utilized for spatial CIC. However, our SSA approach includes visual grounding information, which enables it to effectively generate visually grounded augmentations, making them suitable for spatial CIC tasks as well.

CLID augmentations mainly describe the geometric and possessive relationships of the visual entities, and not their semantic relations. The main focus of these augmentations is on body parts (i.e. in Fig. 16, caption 1-CLID: 'people with a leg', 'a girl with an arm and a leg', and 'a woman with a head') as well as clothing. This is expected, since this type of information is typically captured in existing scene graphs. It is also noticed that most of the generated sentences are difficult to read, containing redundancies and hallucinations. This may be the result of scene graph annotation or generation errors or due to LLM-induced hallucinations and poor quality text generation. On the other hand, the SSA augmentations are based on AMR representations and capture various types of relations, including semantic, geometric, possessive, etc. Unlike other methods, they provide a more natural and human-like description of visual entities, as they are derived from the original dataset of human-annotated captions.

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. 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)
4. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
5. 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)
6. 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)
7. 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)

8. 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)
9. 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)
10. Hirsch, E., Tal, A.: Clid: Controlled-length image descriptions with limited data. arXiv preprint arXiv:2211.14835 (2022)
11. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
12. Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Han, J.: On the variance of the adaptive learning rate and beyond. arXiv preprint arXiv:1908.03265 (2019)
13. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
14. 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)
15. 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)
16. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
17. 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)
18. 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)
19. 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)
20. 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)
21. Zhu, W., Bhat, S.: Gruen for evaluating linguistic quality of generated text. arXiv preprint arXiv:2010.02498 (2020)