RoViST: Learning Robust Metrics for Visual Storytelling

Anonymous ACL submission

Abstract

Visual storytelling (VST) is the task of generating a story paragraph that describes a given image sequence. Most existing storytelling approaches have evaluated their models using traditional natural language generation metrics like BLEU or CIDEr. However, such metrics based on n-gram matching tend to have poor correlation with human evaluation scores and do not explicitly consider other criteria necessary for storytelling such as sentence structure or topic coherence. Moreover, a single score is not enough to assess a story as it does not inform us about what specific errors were made by the model. In this paper, we propose 3 evaluation metrics sets that analyses which aspects we would look for in a good story: 1) visual grounding, 2) coherence, and 3) non-redundancy. We measure the reliability of our metric sets by analysing its correlation with human judgement scores on a sample of machine stories obtained from 4 state-of-the-arts models trained on the Visual Storytelling Dataset (VIST). Our metric sets outperforms other metrics on human correlation, and could be served as a learning based evaluation metric set that is complementary to existing rule-based metrics.

1 Introduction

Visual storytelling (VST) is a natural language generation (NLG) task that aims to automatically generate a cohesive story given a sequence of images (Huang et al., 2016). The task is fundamental to the development of intelligent agents capable of understanding complex visual scenarios, and can be further applied to assist the visually impaired in understanding images on the web. Recently, progress has been made on designing network architectures to accomplish the VST task but little work has been done to explore new metrics that automatically evaluate and quantify the errors produced by these systems. As to date, a majority of the past works on VST have used existing popular n-gram based metrics such as BLEU, METEOR, ROUGE, CIDEr, and SPICE to evaluate their models (Wang et al., 2018; Kim et al., 2018; Hsu et al., 2019; Chen et al., 2021). However, it is known that such metrics are unreliable for VST. Figure 1 shows two machine generated stories for a photo sequence and their corresponding n-gram matching based metrics (BLEU, CIDEr, METEOR, ROUGE-L and SPICE). Evidently, the first candidate story is more repetitive and lacks a narrative style but achieves higher scores across a majority of the n-gram based metrics in Figure 1. The second story however, has greater word diversity and is more expressive through its use of phrases like '*completely in disrepair*'. Relevant words like '*trip*', '*countryside*' and '*hills*' are also used but are not rewarded since they are not mentioned in the gold story.

Input Image Sequence:

Gold Story:
We arrived at the house in the mountains. It was very well kept but the patio and fire pit needed some attention. The guest house was pretty but had a strange style. From the deck, we could see hawks and buzzards soaring over the mountains. We didn't have a bird 's eye view, but ours was more than pretty enough for us.

Candidate Story 1 (AREL):
The view from the top of the mountain was beautiful. This is a picture of a tree. The building was very old. We were able to take a picture of a tree in the distance. The view from the top of the mountain was beautiful.

Candidate Story 2 (MCSM+BART):
A family decided to take a trip to the countryside. They took pictures of the different landscape like this picture outside. At the same time , they found a house that was completely in disrepair. At the end of the day, they saw an eagle in the air. We then looked down over the hills.

	B-1	B-2	B-3	B-4	C	M	R	S
Story 1	52.0	27.3	11.6	0.0	0.5	12.7	25.2	11.1
Story 2	36.8	16.3	0.0	0.0	1.1	14.9	24.1	3.2

Figure 1: Example gold story found in the VIST dataset versus machine output from 2 VST models and their n-gram based metrics.

The low level of agreement between human judgement and current automatic metrics may be because such metrics were originally developed to assess machine translation, summarization and image captioning tasks (Sharif et al., 2018), which are significantly different problems to VST. Specifically, VST is a multimodal task that firstly requires: 1) generating text relevant to the image content but unlike image captioning, there is less emphasis on describing relationships between objects and may contain concepts that are inferred from the image. It additionally needs to ensure that: 2) the story must be topically coherent, similar to how a human would tell a story in a social setting. Sentences should not sound disjointed e.g. '*We went to the park. I grew up in Sydney*'. And finally 3) avoids repetition which appears to be a common issue in current VST models. For instance, Candidate Story 1 in Figure 1 exhibits inter-sentence repetition between the first sentence and last sentence. We also find that some output stories may contain repetition *within* sentences (i.e. intra-sentence repetition) e.g. '*we had a good time and had a great time!*'.

Moreover, it is noted that open-ended text generation tasks usually suffer from the one-to-many issue, whereby there are multiple plausible outputs for the same input which are not fully reflected in the reference sentences (Guan and Huang, 2020). This issue is even more prominent in the VST task as different individuals may tell significantly different stories and have diverse interpretations given the same image sequence. All these issues suggest that we require evaluation metrics that do not simply rely on comparison with reference sentences. In addition, given that the VST task requires several aspects, one single metric is not sufficient to evaluate a story and there is a need to design multiple interpretable metrics that each target a specific VST criteria. Hence, in this paper, we propose several unreferenced metrics for the VST task based on the three aforementioned criteria: 1) visual grounding, 2) coherence, and 3) non-redundancy.

To address criteria 1), we propose a learned metric to calculate relevance scores between nouns in the VST sentences with the bounding box regions in the images. We decide to focus on nouns as they provide the most visual information. Other words like adjectives and adverbs are difficult to ground and such words may differ significantly depending on the person writing the story. The second criteria which is story coherence requires that consecutive sentences flow and that each sentence is not just an isolated description of the image. Existing methods for measuring coherence have used next sentence prediction (NSP) to find the probability that a sentence comes after a preceding sentence (Hu et al., 2020). Inspired by this method, we fine-tune the ALBERT (Lan et al., 2019) model on story sentences and build a sentence-order prediction (SOP) model. Finally, to address criteria 3), we propose an additional metric to explicitly measure inter-sentence and intra-sentence repetition.

The contributions are summarized as follows: 1) We propose an interpretable and reference-free metric that addresses 3 criteria required for VST - visual grounding, coherence and non-redundancy. 2) We conduct human evaluation studies to assess a sample of machine generated stories obtained from 4 state-of-the arts VST models. 3) We test the effectiveness of our proposed metrics by analyzing its correlation with human scores and show that our metrics outperform other existing metrics that are commonly used for VST and NLG tasks.

2 Related Works

Natural Language Generation Metrics The most popular NLG evaluation metrics are BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016). All these metrics are widely used in evaluating image captioning tasks (Anderson et al., 2018; Zhou et al., 2020) and have also been predominantly used in VST tasks (Wang et al., 2018; Hsu et al., 2019; Chen et al., 2021) due to the lack of metrics designed for VST. While these metrics are computationally efficient, they have limited ability in accounting for synonym matches or phrase reordering. This poses a problem for many open-ended text generation tasks like VST where different annotators may have slightly different (but still plausible) ways of describing the same image. To address this, some metrics focus on comparing distance and similarity between word embeddings such as Word Mover's Distance (Kusner et al., 2015), MoverScore (Zhao et al., 2019) and BERTScore (Zhang et al., 2019). However, these metrics mentioned so far still heavily rely on similarity with references, potentially leading to bias for VST tasks as the references may not fully cover the possible ways to write a story for an image sequence.

Visual Grounding Metrics Past studies have proposed examining the images in addition to human written references. Cui et al. (2018) trained a binary classifier to discriminate between human and machine captions using image and text representations obtained from a CNN and RNN. TIGEr (Jiang et al., 2019) employs the pretrained SCAN model (Lee et al., 2018) to calculate the text-to-image grounding scores and compares the relevance ranking and grounding weights distribution among image regions between the references and the candidate. Lee et al. (2020) later introduced ViLBERTScore which uses the same approach as BERTScore but utilizes the ViLBERT model (Lu et al., 2019) to retrieve image-conditioned token embeddings. However, we note that these methods are initially designed for evaluating image captioning systems. Hence, while they do consider the text-to-image similarity aspect, they do not explicitly address the extra criteria required for VST such as story coherence. Moreover, such metrics still rely on reference sentences to some extent.

Story Generation Metrics Language models like BERT (Devlin et al., 2018) trained with NSP and masked language modelling tasks can identify appropriate use of words and sentences and hence, may show promising results when applied to evaluating open-ended text generation. Guan and Huang (2020) proposed UNION, an unreferenced metric for scoring machine generated stories. They leverage a BERT model trained with negative samples created by perturbing ground truth stories and predicts a score representing how human-like a story is. They showed the effectiveness of BERT in identifying stories with conflicting logic, repeated plots and incoherence. However, UNION purely evaluates the output text and cannot be applied to analyse the text-to-image relatedness required for the VST multimodal task. Additionally, a single score is outputted which is not informative enough to gauge what specific errors were made by the model. Moving to VST, Hu et al. (2020) designed reward functions to capture story quality for VST models that use a reinforcement learning framework based on 3 criteria: image relevance, coherence and expressiveness. Image relevance is measured by n-gram precision of entities between candidate and reference sentences, coherence through BERT's NSP task, and word diversity by computing BLEU scores between generated sentences.

Inspired by this, we also analyze story quality from 3 similar perspectives 1) visual grounding, 2) coherence, and 3) non-redundancy. We attempt to extend the methods of Hsu et al. (2019), provide a reference-free approach and conduct a comprehensive analysis with human evaluation.

3 Method

We describe our proposed metric in detail. Given a machine story, we aim to output 3 scores that explicitly evaluates the story based on 1) visual grounding, 2) coherence, and 3) non-redundancy.

3.1 RoViST-VG: Visual Grounding Scorer

To detect the visual relationship between image and text, we build a model that computes the similarity between the nouns in the story sentences with the bounding box regions in the images. We focus specifically on nouns because despite the diverse range of words one can use when storytelling, we notice that the main commonality among the ground truth sentences is the noun mention. This is most likely because nouns (in particular, tangible nouns) tend to offer the most visual information and is the common element that people would recognize when observing an image. An example of this case is in Figure 2 where we can see that the nouns '*dart*' and '*game*' tends to appear in multiple gold sentences, even though each sentence is quite different in structure.

Our visual grounding scorer is inspired by the phrase localization task (Plummer et al., 2015) which involves learning to align sentence entities with image regions. We note that we could have just employed typical image-text matching models like SCAN (Lee et al., 2018) to calculate a similarity score between image and text. However, such models are trained on image captioning sentences and do not explicitly focus on the more fine-grained task of word-region alignment. Moreover, retraining these models with VIST images and whole sentence pairs would be challenging as previously mentioned, story sentences tend to differ significantly in semantics and structure due to human imagination. This is in contrast to image captions where ground truth sentences typically tend to be similar to each other even across different human annotators (e.g. see description in isolation sentences in Figure 2).

Inspired by CLIP (Radford et al., 2021), we create a model that learns the image region and text embeddings such that the noun mention corresponding to an image region will have similar vector representations in geometric space. Let I_i be an image of a bounding box region and T_i be the matching noun. For the image encoder, we follow Radford et al. (2021) and leverage the Vision Transformer (ViT) (Dosovitskiy et al., 2020) to first

Figure 2: Example ground truth description in isolation (dii) and story in sequence sentences (sis) sentences corresponding to an image from the VIST dataset.

Algorithm 1 RoViST-VG

Input: 1)A mini-batch of image regions I_n with shape $(m \times 3 \times 224 \times 224)$ where m is the batch size, and the last 3 dimensions correspond to the image channels, height and width respectively. **2)**A mini-batch of matching noun pairs T_n with shape $(m \times 300)$ where 300 represents the dimensions of the GLoVe vectors. **Output:** Symmetric loss for the mini-batch.

Initialization: Pretrained ViT Model with linear head for the image encoder, and a single linear layer for the text encoder.

```
1: h_n = VisionTransformer(I_n)
2: I_e = tanh(W_i h_n + b_i)            ▷ image embeddings; shape = [m,1024]
3: T_e = tanh(W_t T_n + b_t)            ▷ text embeddings; shape = [m,1024]
4: logits = T_e × I_e^T                 ▷ shape = [m, m]
5: I_sim = I_e × I_e^T                  ▷ shape = [m, m]
6: T_sim = T_e × T_e^T                  ▷ shape = [m, m]
7: labels = (I_sim + T_sim)/2           ▷ shape = [m, m]
8: L_image = cross_entropy_loss(labels^T, logits^T)
9: L_text = cross_entropy_loss(labels, logits)
10: L_symmetric = (L_image + L_text)/2
```

extract the image features from I_i. An additional linear head is further added to project the features to a vector embedding of dimension 1024. For the text encoder, T_i is first converted to 300 dimensional GLoVe vectors (Pennington et al., 2014). If T_i is composed of more than one word, the GLoVe vectors of each token are simply averaged. These vector representations are then passed through a single linear layer to project the text features into the 1024-dimensional joint embedding space. We train the model in a contrastive manner to minimize the symmetric loss. The psuedocode for each batch iteration is provided in Algorithm 1.

To compute the visual grounding score, we extract all nouns from the output story sentences and the top 10 bounding box regions for each image in the story based on the confidence scores generated from Faster R-CNN (Ren et al., 2015). This results in 50 regions for a 5-image story. Each extracted noun and image region is fed through our trained text and image encoder respectively to obtain the image and text embeddings which we denote by I_e and T_e. For each noun, the cosine similarity (cos) is calculated between its text embedding with all other region image embeddings. It is noted that a noun mention from a sentence can match with a region from other images and not necessarily just with regions from its corresponding image as we find that words in story sentences may refer to concepts in other images of the sequence. We then use a greedy matching approach to obtain the maximum similarity score for each noun. Following Zhang et al. (2019), we further experiment by multiplying the similarity score by the inverse document frequency (idf) of the noun calculated from the corpus. This is to put less emphasis on abstract nouns that are not visually grounding but frequently occur in stories (such as '*time*' and '*today*'). Given N stories, the idf score of a token T_i is:

$$idf(T_i) = log(\frac{N}{1 + df(T_i)}) \tag{1}$$

where $df(T_i)$ is the number of stories containing token T_i. Finally, inspired by Lee et al. (2018), a recall score is computed by using LogSumExp (LSE) pooling:

$$S_{VG} = \log \sum_{i=1}^{|T_e|} \exp(idf(T_{e,i}) \max_{I_{e,j} \in I_e} (cos(T_{e,i}, I_{e,j}))) \tag{2}$$

For interpretability, one can optionally scale the score between 0 and 1 using a shifted and scaled version of the sigmoid function:

$$S_{VG(scaled)} = \frac{1}{1 + \exp(-0.5 \times S_{VG})} \times 2 - 1 \tag{3}$$

3.2 RoViST-C: Coherence Scorer

To measure the story's inter-sentence coherence, we leverage the ALBERT model to perform sentence order prediction (SOP) (Lan et al., 2019). The SOP task is a binary classification task, whereby positive samples are consecutive sentences while negative samples are simply constructed by swapping the two sentences around. This forces the model to primarily focus on learning coherence properties rather than topic prediction. We fine-tune the ALBERT model with adjacent story sentences extracted from the VIST and ROCStories dataset. In total, 822,920 training samples were created where 15% was used in the validation split.

Let $\{s_{i-1}, s_i\}_{n=1}^{N}$ denote the training data where s_{i-1} and s_i are adjacent segments. The input sequence fed into ALBERT is in the format $\mathbf{s}_n$ = '[CLS], s_{i-1}, [SEP], s_i, [SEP]', where [CLS] and [SEP] are special tokens. Then, the pooled 1024-dimensional vector representation $\mathbf{h}_n$ of the input sequence is obtained by the output of ALBERT:

$$\mathbf{h}_n = \text{ALBERT}(\mathbf{s}_n) \tag{4}$$

To perform SOP, we add a task-specific linear layer on top of ALBERT to predict the probability that s_i follows s_{i-1}:

$$\hat{p}_n = \text{softmax}(\mathbf{W}_c\mathbf{h}_n + \mathbf{b}_c) \tag{5}$$

where $\mathbf{W}_c$ and $\mathbf{b}_c$ are the trainable weights and bias. For the loss function, we optimize the binary cross-entopy loss as follows:

$$\mathcal{L} = -p_n \log(\hat{p}_n) - (1 - p_n)\log(1 - \hat{p}_n) \tag{6}$$

To obtain the final coherence score for each story, we compute $\hat{p}_n$ for each adjacent sentence pair in the story and average the probabilities across all sentence pairs.

3.3 RoViST-NR: Non-redundancy Scorer

A common problem faced by system output stories is redundancy of words in the form of whole sentences or phrases. While existing methods (Hu et al., 2020) for assessing word diversity and repetition do consider inter-sentence repetition, they do not address repetition *within* sentences. Therefore, to calculate the inter- and intra-sentence non-redundancy score, we propose calculating the Jaccard Similarity (JS) between and within sentences. The JS is defined as the intersection size divided by the union size of two sets (Singh and Singh, 2021). That is, in our problem, the intersection would be the number of co-occurring words between two texts, while the union is the total number of words in both texts. In particular, we compute the Jaccard Similarity with sentence $\hat{y}_i$ and all its preceding sentences $\{\hat{y}_1, ..., \hat{y}_{i-1}\}$ as in Eq. 7. Here, $C(\hat{y}_i)$ and $C(\hat{y}_j)$ are the count of unique words in sentence $\hat{y}_i$ and $\hat{y}_j$ respectively. The inter-sentence repetition score is then just simply the average JS scores across the $\binom{n}{2}$ sentence pairs where n is the number of sentences in the story.

$$JS(\hat{y}_i, \hat{y}_j) = \frac{C(\hat{y}_i) \cap C(\hat{y}_j)}{C(\hat{y}_i) \cup C(\hat{y}_j)} \tag{7}$$

We also measure the intra-sentence redundancy by first splitting each sentence into non-overlapping n-grams and then calculating the JS score between consecutive n-grams within sentences. The intra-sentence repetition score for a story is then the average JS scores across all consecutive n-gram computations. Lastly, we take the mean of the final inter- and intra-sentence score to obtain the final repetition score for the story and subtract from 1. The result is a score between 0 and 1 where a value closer to 1 means that the story tends to contain less redundancy.

4 Data

4.1 Supporting Datasets

VIST The Visual Storytelling Dataset (VIST) dataset (Huang et al., 2016) consists of 10,117 Flickr albums and 210,819 unique images. Each sample is one sequence of 5 photos selected from the same album paired with a single human constructed story, where each story is comprised of mostly one sentence per image.

ROCStories Corpora (Mostafazadeh et al., 2016) is used as additional data along with VIST to train the ALBERT model. It contains 98,161 stories where each story consists of 5 sentences written by humans after being given a prompt.

Flickr30K Entities (Plummer et al., 2015) is derived from the Flickr30K dataset (Young et al., 2014), consisting of 31,783 images each matched with 5 captions. The dataset links distinct sentence entities (i.e. a noun/noun phrase) to image bounding boxes, resulting in 70K unique entities and 276K unique bounding boxes. We use the Flickr30K Entities data to train our visual grounding scorer. After filtering out stopwords from the entity mention, we obtained 566K unique entity-region pairs.

4.2 VST Models

We evaluate our proposed metric on the output stories produced by 4 state-of-the art VST models: 1) **AREL (Wang et al., 2018)**: adopts an inverse reinforcement learning approach trained adversarially. The policy model is a CNN+GRU that generates sub-stories for each image, while the reward model is a CNN-based model designed to output the story reward. 2) **GLACNet (Kim et al., 2018)**: combines both local and global attention. Image features are fed sequentially to a bi-LSTM where the output is a global representation of the entire story. This is concatenated with local image-specific features to create *glocal* vectors which are passed to a decoder for story generation. 3) **KG-Story (Hsu et al., 2020)**: For each image, a word-form conceptual representation is created by predicting a set of terms which are then used to query Visual Genome (Krishna et al., 2017) and OpenIE (Pal et al., 2016) to identify links between sets of terms across images. Finally, a Transformer (Vaswani et al., 2017) takes in the term paths to decode the story. 4) **MCSM+BART (Chen et al., 2021)**: image concepts and related concepts extracted from ConceptNet (Liu and Singh, 2004) are used as input for generating richer stories with BART (Lewis et al., 2020). To incorporate the most appropriate concepts, their Maximal Clique Selection Module model learns a correlation map, reflecting co-occurrence probabilities of all candidate concepts.

5 Evaluation Setup[1]

Evaluation Metrics To assess the performance for RoViST, we analyze its correlation with reliable human judgements by recruiting many responders (26) whereas related works (Guan and Huang, 2020; Hu et al., 2020) have used 3-7 annotators. In total, the 26 responders analysed 400 machine generated sentences across 80 stories and 4 models, including AREL, GLACNet, KG-Story and MCSM+BART. A Likert scale was used to score 3 different criteria for each story based on what we believe defines a good story - 1) the story is visually grounded, 2) sentences are natural sounding and topically coherent, and 3) there is no repeating plots within the story. Annotators were additionally asked to vote for which of the 4 models produced the best story relating to the visual prompt based on no particular criteria. We follow existing literature and report the Spearman's correlation ρ, Pearson's correlation r and Kendall's correlation τ.

Baseline We select 11 baseline metrics to compare with our metric: BLEU-1,2,3,4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), CIDEr (Vedantam et al., 2015), SPICE (Anderson et al., 2016), WMD (Kusner et al., 2015), F_{BERT} (F1-measure version of BERTScore) (Zhang et al., 2019) and TIGEr (Jiang et al., 2019).

[1]The implementation details can be found in the Appendix

AREL: I went on a **boat trip** to the **lake** . This is a picture of a **lake**. This is a picture of a **field** . The **water** was **clear** and the water was **calm**. The boats were **docked** in the water .	**Grounding Score:** 3.4 / 5 **Coherence Score:** 2.7 / 5 **Non-redun. Score:** 2.7 / 5 **Proportion of votes:** 4%
GLACNet: The cruise **ship** was getting ready to go. They were off to a great spot. The **beach** was beautiful. There was a lot of **boats**. It was a very nice day.	**Grounding:** 2.5 / 5 **Coherence Score:** 3.3 / 5 **Non-redun. Score:** 4.6 / 5 **Proportion of votes:** 12%
KG-Story: We had a great time at the **lake**. There were so many **boats**. It was very beautiful. I spent all day out on this **field**. And even saw one **boat.**	**Grounding:** 2.8 / 5 **Coherence Score:** 3.3 / 5 **Non-redun. Score:** 3.9 / 5 **Proportion of votes:** 24%
MCSM+BART: We had a nice **summer** in [location]. The **fields** were absolutely gorgeous. We also had a **farm** that looked like a real **farm** all around. There were even **boat campsites**. Some of the **boats** were out at **night** to **camp**.	**Grounding:** 3.7 / 5 **Coherence Score:** 3.6 / 5 **Non-redun. Score:** 4.1 / 5 **Proportion of votes:** 60%

Figure 3: Average human scores for an example story across 3 criteria for 4 different VST models. 'Proportion of votes' refers to the percentage of voters who voted that model's story as the best out of the 4. Blue highlighted words visually relate to the image.

6 Results

6.1 Human Scores versus Story Ranking

We first investigate whether there is any correlation between the human scores for each 3 criteria and the model that was voted as the best for each photo sequence. For each photo sequence, we rank each of the 4 models' stories based on the proportion of votes that it received. The correlations were then calculated between the mean human scores for each criteria and the model rankings, and the average correlation coefficients were finally taken across the unique stories to obtain the values in Table 2. We also sum up the human scores across the 3 criteria and measure its correlation with the rankings to further analyze at an *Overall* level.

Interestingly, we find that sentence coherence plays the most significant role when ranking stories whereas non-redundancy and visual grounding are less important. Figure 3 provides an example of this case where our human annotators preferred KG-Story and GLACNet over AREL which was more visually grounding but less coherent-sounding. We observe even stronger correlation when we sum the 3 criteria scores, suggesting that all 3 aspects combined can give better guidance when judging a story as can be seen in Figure 3 where most of the votes went to MCSM+BART which scored relatively well in all 3 areas.

	Grounding			Coherence			Non-redun			Overall		
	ρ	r	τ	ρ	r	τ	ρ	r	τ	ρ	r	τ
BLEU-1	0.233	0.209	0.146	0.045	0.035	0.028	0.016	-0.054	0.019	0.093	0.073	0.058
BLEU-2	0.306	0.295	0.209	0.046	0.045	0.030	-0.029	-0.166	-0.005	0.084	0.048	0.059
BLEU-3	0.314	0.314	0.209	0.107	0.143	0.074	-0.076	-0.188	-0.054	0.087	0.079	0.056
BLEU-4	0.280	0.221	0.183	0.107	0.066	0.068	-0.077	-0.224	-0.060	0.073	-0.012	0.040
ROUGE-L	0.287	0.274	0.192	0.187	0.152	0.122	-0.042	-0.152	-0.021	0.125	0.086	0.086
METEOR	0.412	0.392	0.272	0.271	0.238	0.199	0.201	0.057	0.138	0.353	0.295	0.242
CIDEr	0.332	0.201	0.238	0.186	0.094	0.130	0.011	-0.201	0.005	0.208	0.003	0.149
SPICE	0.353	0.345	0.240	0.043	0.059	0.025	0.018	-0.051	0.014	0.144	0.143	0.105
WMD	0.472	0.490	0.337	0.186	0.235	0.129	0.106	0.015	0.076	0.262	0.312	0.183
F_{BERT}	0.221	0.216	0.149	0.272	0.311	0.189	0.087	0.028	0.062	0.213	0.227	0.137
TIGEr	**0.519**	**0.504**	0.354	-0.03	-0.089	-0.027	-0.224	-0.325	-0.147	0.010	-0.005	0.010
RoViST(-VG/C/NR)	0.509	0.460	**0.365**	**0.446**	**0.456**	**0.308**	**0.531**	**0.736**	**0.397**	**0.554**	**0.579**	**0.387**

Table 1: Criteria level Spearman's ρ, Pearson's r and Kendall's τ correlations between automatic metrics and mean of human scores. Correlations for Grounding, Coherence, Non-redun and Overall are measured with RoViST-VG, RoViST-C, RoViST-NR and RoViST respectively.

	ρ	r	τ
Grounding	0.423	0.434	0.400
Coherence	0.663	0.698	0.618
Non-redun	0.379	0.484	0.328
Overall	**0.754**	**0.769**	**0.676**

Table 2: Criteria level Spearman's ρ, Pearson's r and Kendall's τ between human scores and story ranking.

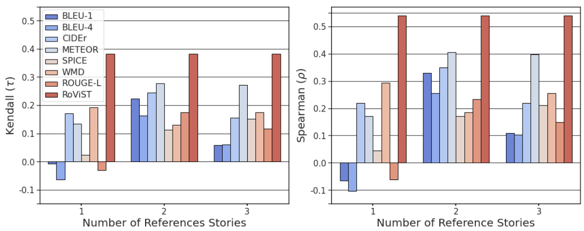

Figure 4: Kendall (left) and Spearman (right) correlation vs. Number of References.

6.2 Correlation Analysis with Human Scores

Table 1 displays the correlation between the metrics and the mean human scores. The results were analyzed at a criteria level by examining correlations between each criteria's scores with our metric which targets that criteria. We also analyze the *Overall* scores by summing up the 3 criteria scores and measuring its correlation with RoViST which represents the sum of the scores produced by RoViST-VG, RoViST-C and RoViST-NR.

With the grounding correlations, RoViST-VG outperforms the baselines for Kendall's correlation. However, it is slightly outperformed by TIGEr when comparing Spearman's correlation and by TIGEr and WMD when comparing Pearson's correlation. We note that all baseline metrics are reference-based and therefore, a likely explanation for the moderate correlations for even simple metrics like METEOR is that human references can already provide a good guideline when assessing text-to-image relatedness. Moreover, we hypothesize that image captioning metrics will perform well for the visual grounding aspect in the case when the model happens to output a sentence that sounds like an image caption. However, unlike image captioning, we emphasize that just having high correlation between image objects and text descriptions does not necessarily mean a good story as we highlighted in the previous section. Examining the coherence and non-redundancy aspect, we observe that a majority of the baselines correlate poorly. Conversely, our RoViST-C and RoViST-NR metric designed to specifically target these criteria generated significantly higher correlations. When comparing at the *Overall* level, we also achieved noticeably better results in terms of ρ, r and τ.

6.3 Changing Number of References

Figure 4 shows how the Spearman and Kendall correlations for some of the metrics vary with different number of human-written references versus our reference-less metric. The stories selected for our analysis each have a different number of reference stories ranging from 1 to 4. As there were not many stories with 4 references, we select those stories that had 3 references, resulting in 60 stories with 300 sentences for analysis. We then compute the correlations with the human judgement across the metrics using 1,2, and 3 references.

It is evident that the results from the reference-based metrics fluctuate significantly according to the number of references. However, the trend is unclear. Increasing the number of references from

Model	Input Image Sequence: 	Coherence Scores
AREL	1. I went to visit the old church. 2. This is the house we stayed at. 3. Then we went to the market to get some food . 4. We had a great time there. 5. The view from the porch was beautiful.	Sent 1-2: 0.956 Sent 2-3: 0.900 Sent 3-4: 0.942 Sent 4-5: 0.337 **Mean: 0.783**
MCSM + BART	1. We found an old looking house . 2. We had an apartment near our home. 3. The decor around the building was very old. 4. We saw people on the street with their dogs. 5. The houses are very old now .	Sent 1-2: 0.058 Sent 2-3: 0.999 Sent 3-4: 0.707 Sent 4-5: 0.460 **Mean: 0.556**
KG-Story	1. We visited a small town. 2. There were many buildings. 3. The dog was very old. 4. This is my home. 5. I also saw some really cool house .	Sent 1-2: 0.999 Sent 2-3: 0.387 Sent 3-4: 0.522 Sent 4-5: 0.149 **Mean: 0.514**
GLACNet	1. The house was very big. 2. It was a lot of work . 3. There was a lot of snow. 4. I had to go back home. 5. I was not happy.	Sent 1-2: 0.128 Sent 2-3: 0.212 Sent 3-4: 0.957 Sent 4-5: 0.101 **Mean: 0.349**

Figure 5: Predicted coherence probabilities from RoViST-C for 4 VST models.

1 to 2 appears to improve most of the correlations for the baseline metrics. This may be because having more references can better capture allowable variations in storytelling compared to a single reference. However, incorporating 3 references actually worsens the performance for some of the metrics like BLEU and CIDEr. A possible explanation could be that the additional reference added may have caused bias for some metrics. In particular, n-gram based metrics like BLEU and ROUGE focus on n-gram overlap. Thus, it is possible that the additional reference introduced may have a high n-gram overlap with the candidate but for unimportant filler words like '*the*' or '*and*'. Our RoViST metric on the other hand, alleviates this issue by first being a reference-free metric and secondly, by only focusing on important words (nouns) in the candidate story via our visual grounding scorer.

It is also noted that examining more amount of references could potentially reveal a better trend. However, this is challenging as the maximum amount of references in the VIST dataset is 5 with 82.50% of the stories having 3 or less. Moreover, collecting multiple human reference stories is an expensive process in most cases.

6.4 Qualitative Analysis

We conduct qualitative analysis on our visual grounding scorer (RoViST-VG) and coherence scorer (RoViST-C).

RoViST-VG Figure 6 in Appendix A displays an example gold story with noun mentions highlighted in blue, followed by the corresponding bounding box regions that gave the highest similarity score retrieved by our RoViST-VG model. We observe that the model performs well at matching a majority of the nouns. However, words that are less visually grounding like '*corner*' or intangible nouns such as '*visit*' are extremely challenging to ground. Consequently, RoViST-VG can sometimes retrieve a region that is not closely related for these types of words. This also occurs for words that are mentioned in the story but not explicitly shown in the images like the word '*photos*' in Example Story 2. A potential problem of this may be the presence of false positives if a story tends to mention many non-visual entities. This could lead to a higher grounding score compared to a story that only mentions a few entities that are visually grounded. Nevertheless, our model can still serve as guidance for analyzing how visually detailed a story is and can also reflect how many related entities a story mentions.

RoViST-C The qualititive results for 4 example machine stories is displayed in Figure 5. Noticeably, RoViST-C tends to assign higher probabilities to sentences that flow. These sentences do not necessarily need to be about the same topic. For instance, sentence 2 and 3 in AREL's story each have a different topic focus but the sentence transition is given a 0.90 coherence score as they follow a narrative style. Conversely, consecutive sentences with similar topics but are incoherent can be given low scores such as sentences 4-5 from KG-Story. It is clear that training ALBERT with sentence order prediction allows the model to capture inter-sentence coherence and is not just limited to modelling topic similarity across sentences.

7 Conclusion

We propose RoViST, a metric for evaluating VST tasks on 3 aspects: visual grounding, coherence and non-redundancy. RoViST correlates well with human judgement, outperforming other metrics when comparing the coherence and non-redundancy criteria as well as when combining all 3 criteria. While some existing metrics slightly outperform our method on visual grounding, we note that image-to-text similarity is just one aspect of VST and this aspect alone is insufficient in defining a good story. Unlike other metrics, RoViST is reference-free and hence, robust to the number of references which are costly to obtain for VST. It is also interpretable and can be used to gauge out where the model is underperforming. We hope that RoViST provides preliminary insight into future work on developing VST models and evaluations.

References

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

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

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Hong Chen, Yifei Huang, Hiroya Takamura, and Hideki Nakayama. 2021. Commonsense knowledge aware concept selection for diverse and informative visual storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 999–1008.

Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge Belongie. 2018. Learning to evaluate image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5804–5812.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16×16 words: Transformers for image recognition at scale.

Jian Guan and Minlie Huang. 2020. Union: An unreferenced metric for evaluating open-ended story generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9157–9166.

Chao-Chun Hsu, Zi-Yuan Chen, Chi-Yang Hsu, Chih-Chia Li, Tzu-Yuan Lin, Ting-Hao Huang, and Lun-Wei Ku. 2020. Knowledge-enriched visual storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7952–7960.

Ting-Yao Hsu, Chieh-Yang Huang, Yen-Chia Hsu, and Ting-Hao Huang. 2019. Visual story post-editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6581–6586.

Junjie Hu, Yu Cheng, Zhe Gan, Jingjing Liu, Jianfeng Gao, and Graham Neubig. 2020. What makes a good story? designing composite rewards for visual storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7969–7976.

Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239.

Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. 2019. Tiger: Text-to-image grounding for image caption evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2141–2152.

Taehyeong Kim, Min-Oh Heo, Seonil Son, Kyoung-Wha Park, and Byoung-Tak Zhang. 2018. Glac net: Glocal attention cascading networks for multi-image cued story generation.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

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

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. 2020. Vilbertscore: Evaluating image caption using vision-and-language bert. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 34–39.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart:

Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

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

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 13–23.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.

Harinder Pal et al. 2016. Demonyms and compound relational nouns in nominal open ie. In *Proceedings of the 5th Workshop on Automated Knowledge Base Construction*, pages 35–39.

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

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99.

Naeha Sharif, Lyndon White, Mohammed Bennamoun, and Syed Afaq Ali Shah. 2018. Nneval: Neural network based evaluation metric for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 37–53.

Ritika Singh and Satwinder Singh. 2021. Text similarity measures in news articles by vector space model using nlp. *Journal of The Institution of Engineers (India): Series B*, 102(2):329–338.

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

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

Xin Wang, Wenhu Chen, Yuan-Fang Wang, and William Yang Wang. 2018. No metrics are perfect: Adversarial reward learning for visual storytelling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 899–909.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. 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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049.

A RoViST-VG Example Output

Figure 6 shows the retrieved regions from the RoViST-VG model for an example gold story (top) and a machine-story generated from the MCSM+BART model (bottom). The blue highlighted words are the nouns while red highlighted words indicate words that do not explicitly appear in the image sequence or are less visually grounding words.

B Implementation Details

RoViST-VG We use the Adam optimizer (Kingma and Ba, 2014) with a 0.00001 weight decay. The learning rate was initially set to 0.00005 and was reduced by 5% with each consecutive epoch. For the ViT model, we use the '*vit-base-patch16-224*' style configuration which outputs image features as a 768 dimensional vector. Further, the linear layer used to project the text and embedding features to the joint embedding space (of dimension 1024) uses a $tanh$ activation function. No normalization of the image and text embeddings was done during the training process as we did not find any benefit from doing this. Finally, we set the mini-batch size to 64 and use early stopping to cease training after the validation loss fails to improve for 3 consecutive epochs. We note that 85% and 15% of the data was used in the training and validation set respectively. The model converged in 3 epochs, taking approximately 12 hours with a Nvidia Tesla P100 GPU.

RoViST-C For ALBERT, we use the '*albert-large-v1*' configuration and the Adam optimizer with a 0.00001 weight decay for training. The learning rate was 0.00001 which we schedule to reduce by 5% every epoch. Additionally, the batch size was 32 and a dropout layer with 40% probability was placed before the final linear layer. Early stopping was employed after the validation loss failed to improve for 5 epochs. We note that 85% and 15% of the data was used in the training and validation set respectively. In total, we trained the model for 5 epochs, taking 14 hours with a Nvidia Tesla P100 GPU.

RoViST-NR For assessing intra-sentence non-redundancy, n-grams of size 4 were used as we found that repetition of words within sentences usually occurred in fours.

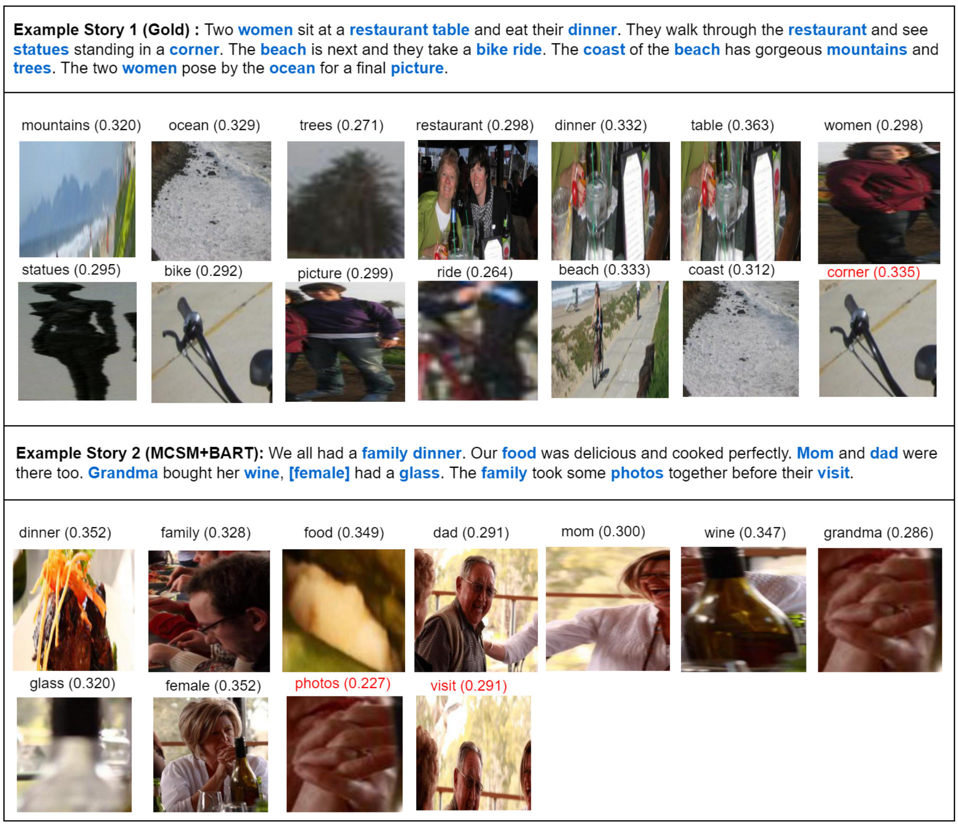

Figure 6: Retrieved regions from RoViST-VG for an example gold story (top) and machine generated story (bottom).

C Human Evaluation Survey

Figure 7 shows the survey instructions used in the human evaluation study and the format of the survey questions. Participants recruited were volunteers from a variety of age groups (20-60 years old), education level and gender (10 female, 16 male).

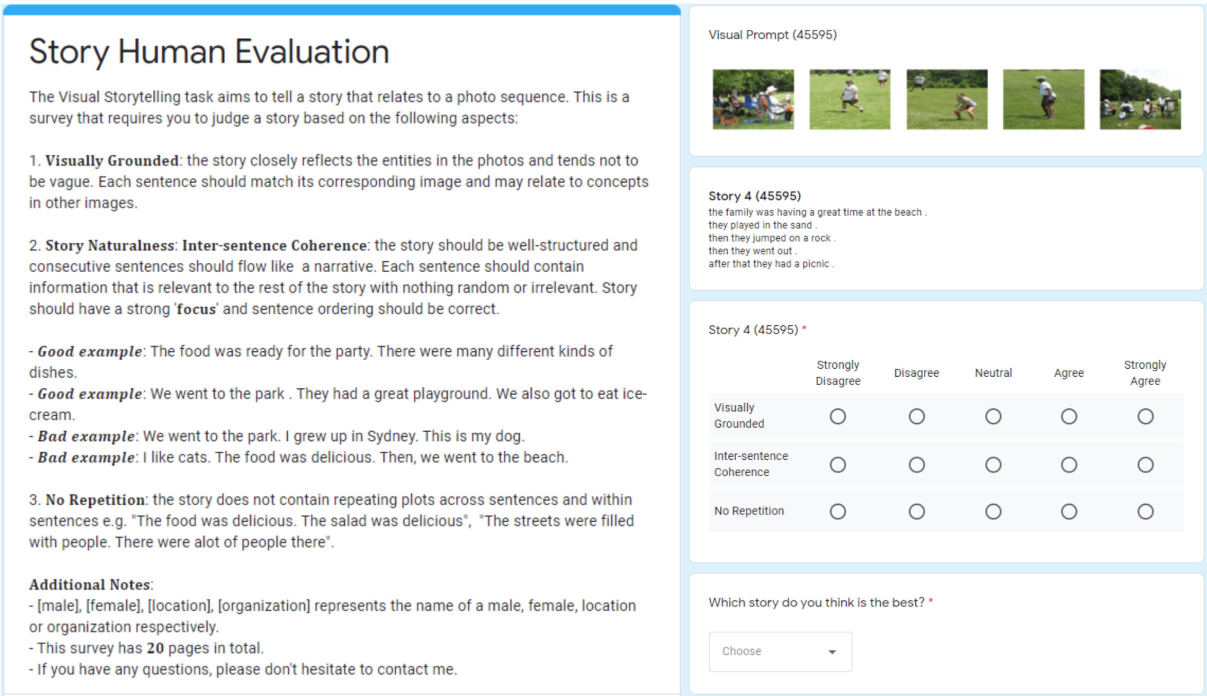

Figure 7: Survey instructions and form format for the human evaluation study.