Towards Coherent Visual Storytelling with Ordered Image Attention

Anonymous ACL submission

Abstract

We address the problem of visual storytelling, *i.e.*, generating a story for a given sequence of images. While each story sentence should describe a corresponding image, a coherent story also needs to be consistent and relate to both future and past images. Current approaches encode images independently, disregarding relations between images. Our approach learns to encode images with different interactions based on the story position (i.e., past image or future image). To this end, we develop a novel message-passing-like algorithm for ordered image attention (OIA) that collects interactions across all the images in the sequence. Finally, to generate the story's sentences, a second attention mechanism picks the important image attention vectors with an Image-Sentence Attention (ISA). The obtained results improve the METEOR score on the VIST dataset by 1%. Furthermore, a thorough human study confirms improvements and demonstrates that order-based Interactions significantly improve coherency (64.20% *vs*. 28.70%).

1 Introduction

Visual Storytelling (VST) (Park and Kim, 2015; Huang et al., 2016) – the task of generating a story based on a sequence of images – goes beyond a basic understanding of visual scenes and can be applied in many real-world scenarios, *e.g.*, to support the visually impaired. Moreover, VST reflects on the creative ability of intelligent systems. Although similar in concept to other cognitive tasks such as image captioning and visual question answering, VST differs as it requires to reason over a *sequence* of images while simultaneously ensuring coherence across multiple generated sentences. To achieve this, VST methods need to address two major challenges: the first is visual and relates to grounding the story's text to the images. The second is linguistic and relates to the quality of the story. Both challenges can be described in terms of coherency: the story should be coherent by itself, and coherent with the images.

Prior research on VST started to address the aforementioned challenges. Early works expand captioning (Vinyals et al., 2014; Xu et al., 2015; Chen and Zitnick, 2015), focusing sentence generation mainly on the current image. This limits the ability to incorporate complex semantic information, which is necessary for visual reasoning. Prior work also makes limited use of temporal dependence and history, *e.g.*, sentences that have already been generated are not used. Consequently, the output lacks narrative consistency and is prone to linguistic errors such as *repetitiveness* and *incoherence* (Modi and Parde, 2019). To mitigate these issues, later works strive to generate more meaningful stories via adversarial and reinforcement learning (Wang et al., 2018; Huang et al., 2018), which remain delicate to train.

Importantly, images are not independent. For example, if the first image in a sequence shows a protest, the model may want to focus on signs in later images. Conversely, if the last image shows a ring on a finger, then the model should pay attention to wedding-related objects and activities in the preceding images. This is important for VST because sentences are created per image but are part of a story. Hence, objects that the model is focusing on in one image should be conditioned on the selection in other images.

To do this we develop a novel model which (1) implicitly reasons over objects, activities, and their temporal dependencies in each image; and which (2) improves the coherency of the narrative. To reason over objects and activities in each image, *i.e.*, to understand their dependencies and their temporal ordering, we introduce *ordered image attention* (OIA). As illustrated in Fig. 1, for each image, OIA accumulates representation information from objects detected within the corresponding image into an attended image representation. Importantly, accumulation factors depend on whether the image

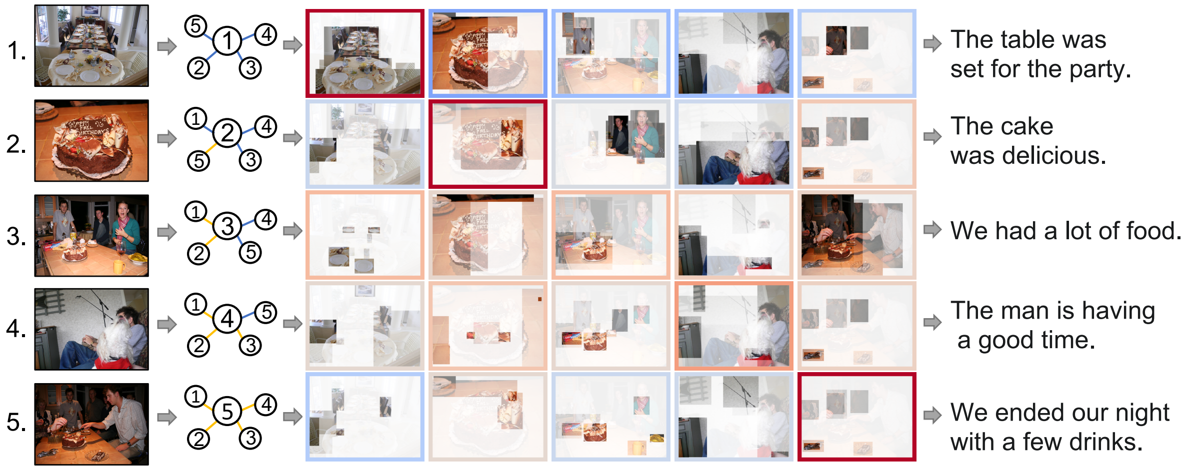


Figure 1: We propose Ordered Image Attention (OIA) to encourage coherency. In each row, we show the spatial attention maps. On the graph on the left, each colored edge indicates how an object is involved in a specific interaction. A yellow edge indicates a preceding interaction and a blue edge indicates a subsequent interaction. By collecting directional interactions, we can identify significant objects throughout the story. In total, five attention maps are calculated, one for each image. The border of attended images indicates how important an image is according to the Image-Sentence Attention (ISA). *E.g.*, red indicates a high attention score, meaning the image is essential for generating that sentence. Our model performs this step for all five images simultaneously, creating 25 attention maps that are fed into the decoder to create the sentences sequentially.

precedes or succeeds the image for which we are currently generating the sentence, which permits to establish an order. The attended image representations are subsequently summarized into a context embedding via an Image-Sentence Attention (ISA) unit, before being used for sentence decoding.

In addition, to alleviate common linguistic mistakes like repetitiveness and to promote coherence in the story, we incorporate information from the story generated up to the current sentence into the sentence generation decoder. Specifically, the decoding strategy decays the probability of a word if it has already been used in the story. The decoder also maintains a separate prior over the output probability distribution, independent from the language generation unit. This prior is based on counts of the words that were already predicted in the story. Both the prior, and the Recurrent Neural Net (RNN) decoder output are combined to predict the next word in the sentence.

Empirical results on the challenging VIST dataset demonstrate that the proposed method generates stories with an improved narrative quality. The method outperforms prior state-of-the-art by 1% on the METEOR score. Examples of stories generated by the approach are shown in Fig. 1. We also present a user study demonstrating the advantage of the model in terms of coherency (64.20% *vs*. 28.70%).

2 Related Work

Vision+Language has been an active area of research for many years, addressing tasks such as image/video captioning, paragraph generation, and visual question answering. We briefly review those related areas in the following.

2.1 Visual Storytelling

Huang et al. (2016) introduce the visual storytelling task. Visual storytelling is similar to captioning. Thus, early methods adapted captioning mechanisms, introducing context between story sentences (Gonzalez-Rico and Pineda, 2018). Following, Kim et al. (2018b) used a seq2seq (Sutskever et al., 2014) approach built on a decoding sampling strategy to reduce repetition. Here, we use a dynamic data-driven approach where each word is penalized differently based on its average count. Next, Wang et al. (2018) discuss the difficulty of learning stories with imaginary details that do not appear in the imagery. To that end, an adversarial reward system is used to improve the output stories. Several works use a reinforcement learning approach based on the interrelationships between images (Huang et al., 2018). Recently, state-of-the-art results were obtained by generating scene graphs for each image in the sequence (Wang et al., 2019). Following, Li et al. (2019) and Zhang et al. (2020) rely on preprocessing to ground visual elements. Yang et al. (2019) and Hsu et al. (2020) enrich the data with an external word common-sense knowledge graph.Wang et al. (2019) model relations within the image with scene graphs, which requires expensive annotations. Hong et al. (2020) also uses scene graphs with global embeddings to achieve coherence. In contrast, we model ordered

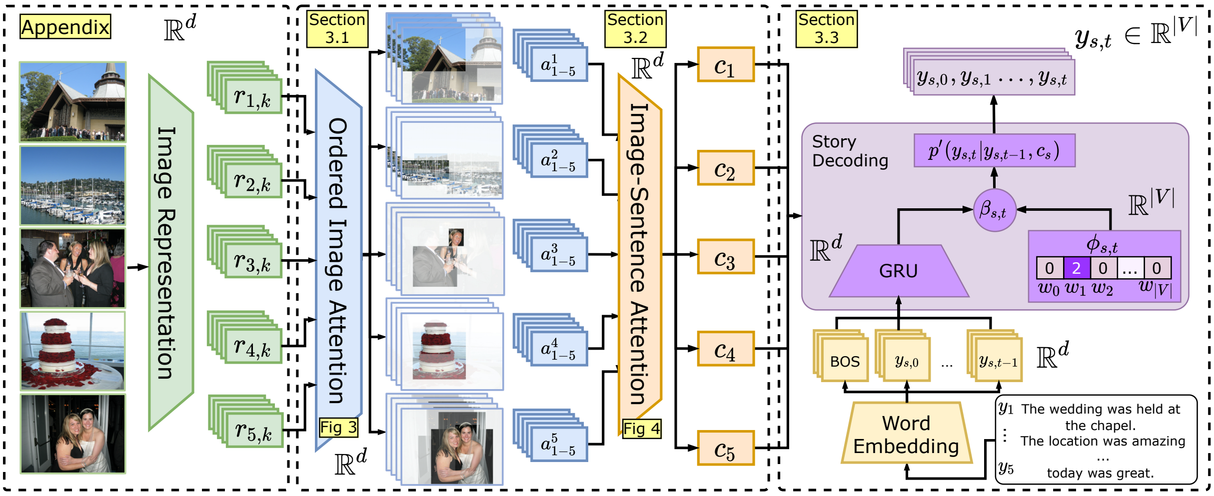


Figure 2: Our architecture for Visual Storytelling synthesis.

interaction for coherence. A recent study by Yu et al. (2021) employed large-scale pre-trained models for visual storytelling using auxiliary adaptation loss. Our research aims to establish a novel visual storytelling model that models ordered interactions without external knowledge.

Recently, Yu et al. (2021) employ large pre-trained models to visual storytelling via auxiliary adaptation loss. Our research focuses on creating a novel visual storytelling model instead of training strategies, loss functions, and pre-trained models. Encoding ordered images improves the coherence between sentences, one of the main challenges of visual storytelling.

2.2 Image Captioning

Barnard et al. (2003) first explored annotating images with text. Since then, image/video captioning has seen a surge of research activity. Initial work utilized pre-trained image embeddings from a CNN network. The success of attention mechanisms for language translation quickly transferred to image captioning as well (Xu et al., 2015). Later work leveraged advances in object detection and proposed a bottom-up/top-down attention approach to attend to specific objects in the image instead of fixed spatial regions (Anderson et al., 2017). Different from image captioning, for visual storytelling, both story coherency and visual grounding are important.

2.3 Multimodal Attention

Multimodal problems are characterized by input data that comes from different domains, *e.g.*, visual and linguistic. This raises two challenges: 1) how to model interactions between different domains, and 2) how to manage the large input data. Considering those challenges, attention has been a prominent tool as it models interactions to select the important elements. In early work, Xu et al. (2015) used interaction-based attention with the image at each caption generation step. This idea was later extended to visual question answering (Xu and Saenko, 2016). To imitate multi-step reasoning, Yang et al. (2015) stacked attention modules sequentially. Later, many works concentrated on better vector-fusion modeling (Fukui et al., 2016; Kim et al., 2017; Ben-Younes et al., 2017; Yu et al., 2018). Importantly, Lu et al. (2016) suggested attending to the visual and textual modalities separately. Afterward, Kim et al. (2018a) proposed a bilinear module that efficiently generates attention for every pair. Following Lu et al. (2016), Schwartz et al. (2017, 2019) suggested a general framework that extends attention to any number of utilities via local and interaction-based factors. We improve upon those ideas by suggesting an ordered attention. This ensures that interaction modeling is affected by the image position in a sequence.

3 Method

The goal of visual storytelling is to generate a story, composed of N *ordered* sentences $\{y_s | 1 \leq s \leq N\}$, given an *ordered* sequence of images $I = \{I_s | 1 \leq s \leq N\}$. Each sentence $y_s = (y_{s,0}, \dots, y_{s,t}, \dots)$ is composed of words $y_{s,t} \in \mathcal{Y}$ from vocabulary $\mathcal{Y}$.

The order in which the images are given is essential as it defines the plot line of the story. The story should be focused, *i.e.*, each sentence should be related to the remainder of the story. Importantly, the sentences should form a coherent body of text describing the set of images, and not only a set of related information. For instance, the story *“The*

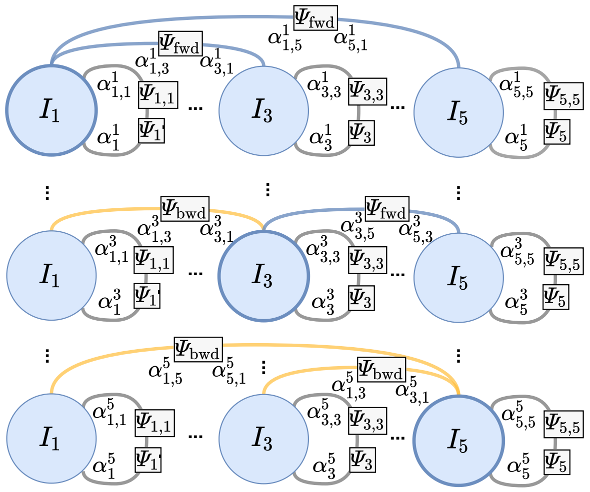


Figure 3: Illustration of Ordered Image Attention. Each node represents an image attention belief. For each sentence, we connect all the images with the sentence-corresponding image. The relative position to this image determines whether the connection is modeled with the Ψ_{bwd} factor (for preceding images) or the Ψ_{fwd} factor (for subsequent images; see Eqs. [8-10]). We infer the attention belief by collecting interactions and local object information within the image see Eqs. [2-4]). We use scalars to calibrate the importance of each factor. In total, we generate 25 attention maps, one per image for every sentence.

church was beautiful. The bride and groom walk down the aisle. The cake was amazing." is less coherent than: *"We went to the church for the wedding today. The bride and groom were excited for the day. Both cut the cake together."* **Overview:** To address this challenge, we develop the model illustrated in Fig. 2. It infers conditional probabilities $p'(y_{s,t}|y_{s,t-1}, c_s)$ for the t-th word $y_{s,t} \in \mathcal{Y}$ in sentence y_s given the previous word $y_{s,t-1}$ and the context embedding c_s for sentence s. The context embedding c_s summarizes region representations $r_{i,k}$ of all K object regions across all N images I_i ($i \in [1, N]$, $k \in [1, K]$) via Ordered Image Attention (OIA) (Sec. 3.1) and Image-Sentence Attention (ISA) (Sec. 3.2). Specifically, when generating sentence s, OIA computes an attended image representation a_i^s for every image I_i by attending to the K region representations $r_{i,k}$ (Sec. 3.1). These attended image representations a_i^s are subsequently summarized into the context embedding c_s via an image-sentence attention (Sec. 3.2).

Below we first discuss computation of the attended image representation a_i^s (Sec. 3.1), before detailing computation of the context embedding c_s (Sec. 3.2) and computation of the conditional probabilities $p'(y_{s,t}|y_{s,t-1}, c_s)$ (Sec. 3.3).

Ordered Image Attention (OIA) is designed to 1) form a structure across ordered images and to 2) select the relevant objects per image. For this we model preceding and subsequent interactions

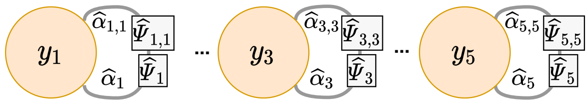


Figure 4: Illustration of ISA. The attention selects the attended image representation per sentence. We model interactions between attended images of the same sentence to compute each image's importance. Note, each node represents a sentence attention belief over the attended images.

separately using different attention factors. We calibrate each factor's importance with trainable scalars, which forms a graph of dependencies between the images. For each sequence of N images, the model infers a total of N^2 attention maps, one per image for each sentence. We detail this module next.

3.1 Ordered Image Attention (OIA)

3.1.1 Attention Belief

For each image $I_i = \{r_{i,1}, \dots r_{i,K}\}$ we consider a set of K regions, represented by their feature vectors $r_{i,k} \in \mathbb{R}^d$, where d is the objects' embedding dimension. Suppose we are currently generating sentence y_s ($1 \leq s \leq N$). To do this we first compute an attended image representation a_i^s as follows

$$a_i^s = \sum_{k=1}^{K} b_{i,k}^s r_{i,k}, \qquad (1)$$

where $b_{i,k}^s \geq 0$ is the attention belief highlighting the importance of the k-th object in the i-th image when generating the s-th sentence. Importantly, for every image I_i we require $b_{i,k}^s$ to be a valid probability distribution, *i.e.*, we also enforce $\sum_{k=1}^{K} b_{i,k}^s = 1 \; \forall s, i$.

The object attention belief $b_{i,k}^s$ is dependent on all the input data, *i.e.*, other objects and images. To avoid complex computation, we factorize the belief $b_{i,k}^s$ into two pairwise dependencies that preserve the order, and a local term. For the pairwise terms we use $\mu_{j \to i}^{\text{bwd}}$, which is a message from a preceding image I_j, or $\mu_{j \to i}^{\text{fwd}}$, which is a message from a subsequent image I_j. We also use $\mu_{i \to i}$ for self-messages. Additionally, we include a local factor $\Psi_i(r_{i,k})$ that considers the object representation. Unlike the messages mentioned before, the local factor does not rely on interactions with other objects. We aggregate all the messages along with the local factor as illustrated in Fig. 3. For normalization we employ a softmax.

Formally we compute the attention belief $b_{i,k}^s$ by distinguishing three cases. If $i = s$ we have

$$b^s_{i,k} \propto \exp(\alpha^s_i \Psi_i(r_{i,k}) + \alpha^s_{i,i}\mu_{i\to i}(r_{i,k}) + \sum_{j<i} \alpha^s_{i,j}\mu^{\text{bwd}}_{j\to i}(r_{i,k}) + \sum_{j>i} \alpha^s_{i,j}\mu^{\text{fwd}}_{j\to i}(r_{i,k})). \quad (2)$$

If $i < s$ we use

$$b^s_{i,k} \propto \exp(\alpha^s_i \Psi_i(r_{i,k}) + \alpha^s_{i,i}\mu_{i\to i}(r_{i,k}) + \alpha^s_{i,s}\mu^{\text{bwd}}_{s\to i}(r_{i,k})). \quad (3)$$

If $i > s$ we obtain

$$b^s_{i,k} \propto \exp(\alpha^s_i \Psi_i(r_{i,k}) + \alpha^s_{i,i}\mu_{i\to i}(r_{i,k}) + \alpha^s_{i,s}\mu^{\text{fwd}}_{s\to i}(r_{i,k})). \quad (4)$$

In all three cases $\alpha^s_i, \alpha^s_{i,i}, \alpha^s_{i,j} \in \mathbb{R}$ are scalars used to calibrate the importance of different messages for a given sentence. These scalars form a dependency structure between images for each of the generated sentence indices. Intuitively, when we generate the first sentence, the attention belief might depend more on subsequent images, to correctly identify the story event, *e.g.*, a wedding, a parade, *etc*. Thus, the scalars will promote interaction with later images. An analysis of these scalars is provided in the appendix. Next, we define the different types of messages.

3.1.2 Pairwise Messages and Factors

A message aggregates interaction scores from an image to an object. The three messages $\mu^{\text{bwd}}_{j\to i}$, $\mu^{\text{fwd}}_{j\to i}$ and $\mu_{i\to i}(r_{i,k})$ are computed as follows:

$$\mu^{\text{bwd}}_{j\to i}(r_{i,k}) = \sum_{k'=1}^{K} \Psi_{\text{bwd}}(r_{i,k}, r_{j,k'}), \quad (5)$$

$$\mu^{\text{fwd}}_{j\to i}(r_{i,k}) = \sum_{k'=1}^{K} \Psi_{\text{fwd}}(r_{i,k}, r_{j,k'}), \text{ and} \quad (6)$$

$$\mu_{i\to i}(r_{i,k}) = \sum_{k'=1}^{K} \Psi_{i,i}(r_{i,k}, r_{i,k'}). \quad (7)$$

Importantly, these messages collect three different types of order-dependent interaction factors: (1) A backward image interaction, namely $\Psi_{\text{bwd}}(r_{i,k}, r_{j,k'})$. This interaction models relations to the preceding j-th image in the sequence. (2) A forward image interaction, namely $\Psi_{\text{fwd}}(r_{i,k}, r_{j,k'})$. This interaction models relations to the subsequent j-th image in the sequence. (3) The self interaction factor, namely $\Psi_{i,i}(r_{i,k}, r_{i,k'})$, which takes into account interactions between objects within the image. We formally define the different factors next.

Interaction factors: A commonly used practice to capture interactions across attention mechanisms is to first embed the elements into a joint Euclidean space followed by a dot-product (Vaswani et al., 2017; Schwartz et al., 2017; Gao et al., 2019; Schwartz et al., 2019). While we follow the same practice, we define three types of interaction factors to preserve the order. Consider two objects, $r_{i,k} \in I_i$ from the sentence-corresponding image and $r_{j,k'} \in I_j$ from the interacting image. We describe three types of interactions: for interactions with subsequent images (*i.e.*, $j > i$) we use

$$\Psi_{\text{fwd}}(r_{i,k}, r_{j,k'}) = \left(\frac{L_{\text{fwd}} r_{i,k}}{\|L_{\text{fwd}} r_{i,k}\|_2}\right)^\top \left(\frac{R_{\text{fwd}} r_{j,k'}}{\|R_{\text{fwd}} r_{j,k'}\|_2}\right). \quad (8)$$

For interactions with preceding images (*i.e.*, $j < i$) we use

$$\Psi_{\text{bwd}}(r_{i,k}, r_{j,k'}) = \left(\frac{L_{\text{bwd}} r_{i,k}}{\|L_{\text{bwd}} r_{i,k}\|_2}\right)^\top \left(\frac{R_{\text{bwd}} r_{j,k'}}{\|R_{\text{bwd}} r_{j,k'}\|_2}\right). \quad (9)$$

For interactions within the image (*i.e.*, $j = i$) we have

$$\Psi_{i,i}(r_{i,k}, r_{i,k'}) = \left(\frac{L_{i,i} r_{i,k}}{\|L_{i,i} r_{i,k}\|_2}\right)^\top \left(\frac{R_{i,i} r_{i,k'}}{\|R_{i,i} r_{i,k'}\|_2}\right). \quad (10)$$

Note, $L_{\text{fwd}}, R_{\text{fwd}}, L_{\text{bwd}}, R_{\text{bwd}}, L_{i,i}, R_{i,i} \in \mathbb{R}^{d\times d}$ are trainable shared weights across the entire image sequence. Also, the object from the sentence-corresponding image will always be on the left side of the factor equation. Thus, the factor embeddings preserve the order.

Local factor: Differently from the previous interactions the following factor captures how important an object is based solely on the object representation. Given an object $r_{i,k} \in I_i$, we define the local factor as,

$$\Psi_i(r_{i,k}) = v^\top \text{ReLU}(V r_{i,k}), \quad (11)$$

where $v \in \mathbb{R}^d, V \in \mathbb{R}^{d\times d}$ are trainable weights.

3.2 Image-Sentence Attention (ISA)

In a next step we summarize the attended image representations a^s_i produced by OIA to compute the context embedding c_s for the sentence s that we wish to generate. For this we use the Image-Sentence Attention (ISA) unit. It picks the relevant image context for generating the specific sentence. Formally we obtain the context embedding via

$$c_s = \sum_{i=1}^{N} \hat{b}_{s,i} a^s_i, \quad (12)$$

Method	M	B-1	B-2	B-3	B-4	R	C	Img Feat
AREL[31]	35.0	63.8	39.1	23.2	14.1	29.5	9.4	FC
KS[34]	35.2	66.4	39.2	23.1	12.8	29.9	**12.1**	FC
HSRL[12]	35.2	-	-	-	12.3	29.5	8.4	Spatial
StoryAnchor[38]	35.5	65.1	40.0	23.4	14.0	30.0	9.9	FC
SGVST[30]	35.8	65.1	40.1	23.8	14.7	29.9	9.8	F-RCNN
SGEmb[10]	35.6	62.2	38.7	. 23.5.	14.8	30.2	8.6	F-RCNN
Ours	**36.8**±0.1	**68.4**±0.7	**42.7**±0.3	**25.2**±0.2	**15.3**±0.2	**30.2**±0.1	10.1±0.2	F-RCNN

Table 1: Quantitative results on the VIST dataset for METEOR, BLEU-1. . .4, ROUGE-L and CIDEr. The primary metric is METEOR. The 'Img Feat' column describes the pretrained image features. All models utilize a ResNet (He et al., 2015) backbone except CS&T which employs an Inception v3 model (Szegedy et al., 2015). FC and Spatial refer to features extracted from the penultimate layer and the preceding one accordingly. F-RCNN are bottom up features (Anderson et al., 2017).

where attention factors

$$\hat{b}_{s,i} \propto \exp\left(\hat{\alpha}_s \hat{\Psi}_i(a_i^s) + \hat{\alpha}_{s,s}\hat{\mu}_{s\to s}(a_i^s)\right), \quad (13)$$

and where $\hat{\alpha}_s, \hat{\alpha}_{s,s} \in \mathbb{R}$ are scalars. To avoid spurious correlations between sentences, we consider only self interactions and a local factor. This is illustrated in Fig. 4. The self-message of the attended image representation a_i^s is

$$\hat{\mu}_{s\to s}(a_i^s) = \sum_{j=1}^{N} \hat{\Psi}(a_i^s, a_j^s). \quad (14)$$

Finally, the self and local factors are defined with a different set of weights following Eq. (10) and Eq. (11) respectively.

3.3 Story Decoding

The goal at each timestep of decoding is to compute the conditional probability $p(y_{s,t}|y_{s,t-1}, c_s)$ where $y_{s,t} \in \mathcal{Y}$ is the t-th word in sentence y_s, $\mathcal{Y}$ is the vocabulary and c_s is the context embedding detailed in Sec. 3.2. For this we use a GRU recurrent unit, tasked with generating probabilities over the vocabulary conditioned on the context embedding c_s and the previously generated token $y_{s,t-1}$: $p(y_{s,t} = w|y_{s,t-1}, c_s) \propto$

$$\exp(\beta_{s,t} \cdot g_w(y_{s,t-1}, h_{s,t-1}, c_s) + (1-\beta_{s,t}) \cdot f_w(\phi_{s,t})), \quad (15)$$

where g_w is the output of a GRU unit for the word w. We set the GRU hidden dimension to d. $h_{s,t-1} \in \mathbb{R}^d$ is the hidden state at timestep $t-1$ for sentence s. $f : \mathbb{R}^{|\mathcal{Y}|} \to \mathbb{R}^{|\mathcal{Y}|}$ is a learned prior over the vocabulary based on a bag-of-words prior histogram $\phi_{s,t}$, which we describe in the next paragraph. The purpose of f is to reduce text repetitions. f_w denotes the value of f for a word w. We also incorporate a calibration gate $\beta_{s,t} : \mathbb{R}^d \to [0, 1]$ for functions f and g using

$$\beta_{s,t} = \sigma\left(v_\beta^\top \tanh(G_g h_{s,t} + G_f W_1(\phi_{s,t}))\right). \quad (16)$$

Here, $G_g \in \mathbb{R}^{d\times d}$ and $G_f \in \mathbb{R}^{\gamma\times d}$ are trained projections of the GRU hidden state and the bottleneck layer respectively, $v_\beta \in \mathbb{R}^d$ are learned weights and σ is the sigmoid function. W_1 is obtained from the prior as discussed next.

Bag-of-words (BOW) prior: Remembering history during storytelling permits to stay on topic and advance the story in the desired direction. Although quite intuitive, mimicking this ability is not trivial. *E.g.*, most approaches for VST generate all the sentences in parallel. Converting the parallel sentence generation into a sequential one implies a major computational overhead during training.

To address this, we propose a simple yet effective learnable framework that does not require sequential training while still exploiting information found in prior sentences. The history is represented via a bag-of-words histogram $\phi_{s,t}$, which includes all words that have been used until timestep t for the s-th sentence. During training, we initialize $\phi_{s,t=0}$ with the ground truth history counts found in the previous $s-1$ sentences. We update the statistics at each timestep with the predicted word $y_{s',t}$ for $s' < s$, and produce the next state of the counter $\phi_{s,t+1}$. At inference we generate sentences sequentially and update $\phi_{s,t}$ with the predicted words. $\phi_{s,t}$ is fed through a shallow bottleneck network to obtain the prior f, composed of two layers $W_1 \in \mathbb{R}^{|\mathcal{Y}|\times\gamma}$ and $W_2 \in \mathbb{R}^{\gamma\times|\mathcal{Y}|}$ without activation, where γ is the bottleneck dimension:

$$f(\phi_{s,t}) = W_2(W_1(\phi_{s,t})). \quad (17)$$

Also note the use of $W_1(\phi_{s,t})$ in the gate (Eq. (16)).

Intra-repetition regularization: To regularize intra-repetitions, we decay the probability of previously used words during sentence generation. A critical aspect of this approach is to exclude words that appear frequently in the language (*e.g.*, was, were, am). For this we pre-process the training set to calculate the average story frequency $\rho(w)$ of a word w via $\rho(w) = \frac{\#\text{ appearances of word } w}{\#\text{ stories } w \text{ was used}}$. The final count for word w at timestep t is calculated as $\phi'_{s,t}(w) = \max[0, (\phi_{s,t}(w) - \rho(w) + 1)]$. Intuitively, a word will not be penalized before it is used more than the prior belief average $\rho(w)$. The final probability for word w being used is given by

$$p'(y_{s,t} = w | y_{s,t-1}, c_s) = \frac{p(y_{s,t} = w | y_{s,t-1}, c_s)}{\pi \cdot \phi'_{s,t}(w) + 1}, \quad (18)$$

where $\pi \geq 0$ is a constant hyper-parameter. A penalty of 2 proved to work best on the validation set.

4 Results

Dataset: To train and test the model we use the VIST dataset. All images were collected from Flickr albums. All images from a story were taken from the same album. Each image sequence has five reference stories. Approximately 2.5 of the stories are based on human annotations, while the rest are rewrites. The overall numbers are 40,098 training stories, 4,988 validation stories, and 5,050 test stories.

Evaluation metrics: As suggested by the creators of VIST, METEOR correlates best with human judgement. We also report BLEU, ROUGE, and CIDEr and compare to prior work where available. The metrics are based on word correspondence with human references, which is unsuitable for measuring visual storytelling quantities such as coherence. For example, the ROUGE and CiDER scores are almost identical for all the recent years' baselines. While our experiments indicate statistically significant improvements across all metrics, we emphasize that human evaluation are currently the most reliable way to evaluate visual storytelling approaches. We conducted those in Sec. 4.2.

4.1 Quantitative analysis

Comparison to state-of-the-art: In Tab. 1 we compare the method to recent baselines. Early methods did not take into account visual-spatial information (*i.e.*, they employed FC features), which harms the performance (*e.g.*, 35.5% *vs.* 36.8% on METEOR). Wang et al. (2019) utilize image representations similar to our approach but do not consider relations between different images, resulting in a 1% drop on METEOR, showing that ordered structure encoding with OIA is beneficial. SGVST and StoryAnchor map images to distinct topics based on external knowledge. On the other hand, our approach is trained end-to-end. Furthermore, our image representations depend on all the images in a sequence. In contrast, SGVST uses scene graphs. Such models are pre-trained with an external model for generating scene graphs. Finally,

Model	M	B-4	R	C	#Params
attention					
w/o OIA	36.0	14.1	30.0	8.4	11M
w/o ISA	35.9	14.2	29.9	9.3	11M
w/o attention	35.8	13.6	29.7	7.2	11M
no-direction	36.1	14.5	28.9	8.4	12M
decoding					
w/o rep. regularization	36.2	14.5	29.8	8.7	13M
w/o count norm	36.2	14.6	29.9	9.4	13M
w/o BOW prior	36.2	14.5	30.0	9.7	13M
Transformer	36.7	**15.7**	30.0	9.9	13M
Ours	**36.8**	15.3	**30.2**	**10.1**	13M

Table 2: Components ablation analysis.

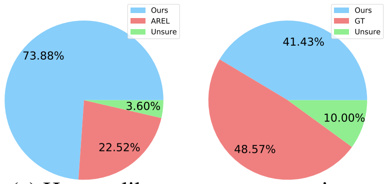
Local	Self	Directional	M	B-4	R	C
×	✓	✓	36.2	14.5	30.0	9.3
✓	×	✓	36.0	14.4	29.8	9.2
✓	✓	×	35.5	14.2	29.9	8.5
✓	✓	✓	**36.8**	**15.3**	**30.2**	**10.1**

Table 3: Factor ablation analysis.

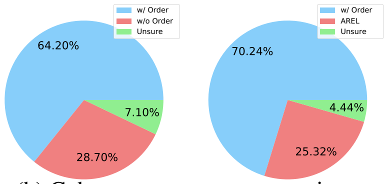
Yang et al. (2019) enhance the input with an external commonsense dataset. CIDEr scores are significantly higher, yet this improvement is not reflected in all metrics. Our work improves the state-of-the-art METEOR score from 35.8% to 36.8%. This increase is larger than the 0.8 increase of all advances since the 2018 VIST challenge (*i.e.*, 35.0% *vs.* 35.8%).

Ablation study: In Tab. 2 we conduct an ablation study for two novel components in our model: 1) Attention: In 'w/o OIA,' we replace the OIA module (Sec. 3.1) with simple averaging of the K object representations of image I_i, resulting in a 0.8% drop on METEOR. Similarly, in 'w/o ISA,' we replace the ISA unit (Sec. 3.2) with averaging, leading to a 0.9% drop on METEOR. In 'w/o attention,' we removed both OIA and ISA, which dropped the METEOR score to 35.8%. For the method referred to as 'no-direction,' we use the same factor for preceding and subsequent interaction (*i.e.*, $L_{\text{bwd}} = L_{\text{fwd}}$ and $R_{\text{bwd}} = R_{\text{fwd}}$). Here, METEOR results drop by 0.7%. Hence, ordered interactions are beneficial. 2) We assess the decoding components (Sec. 3.3). We first remove the intra-repetition regularization (*i.e.*, $\rho(w)$), which causes METEOR score to drop by 0.6%. Removing the popular words count ($\phi'_{s,t}$), results in a 0.4% drop on METEOR. The METEOR score drops by 0.4% when we remove the BOW prior. Last, we replace the GRU decoding layer with a Transformer, which did not change results a lot.

In Tab. 3 Further, we assess the necessity of different factors used in OIA. All factors contribute to the model's performance and the directional factors (*i.e.*, Ψ_{fwd} and Ψ_{bwd}) have the biggest impact.



(a) Human-like property comparison.



(b) Coherence property comparison.

Figure 5: Human evaluation to compare properties.

In Tab. 5, we show the ability to reduce repetitions. Text repetitiveness is measured by the repetition rate of non-singleton n-grams within each story. In our experiment, we use up to 4-grams. The use of intra-repetition regularization reduces text repetition (0.14 to 0.04). Combined with the trainable bag-of-words prior module, we further improve this measure (0.008 *vs*. 0.14). We also report sentence repetitiveness, *i.e*., the average number of repeated sentences in a story.

4.2 Human Evaluation

Due to the subjective nature of the VST task, a human evaluation is required. We randomly selected 150 image sequences from the test set and asked three MTurk annotators to rank them or assess them with other methods. We use the AREL as a baseline. Further, our method of generating coherent stories is tested using a slightly weaker variation of the method without directionality. Since the most recent baselines are not publicly available, we cannot compare them to recent approaches. We'll share the selected sequences to aid future comparisons.

In Fig. 5b, we assess stories coherency. To begin, we examine the importance of modeling direction-aware interactions. In our comparison, we changed only one aspect of our model. We used the same factor for preceding and subsequent interactions. We show that on VIST metrics the effect is relatively small (*i.e*., no-direction; see Tab. 2). However, the human-evaluation comparison shows a significant coherency improvement (64.2% *vs*. 28.7%), which is not revealed with classical evaluation. Also, a comparison against the AREL baseline demonstrates a more significant improvement (70.24% *vs*. 25.32%).

In Fig. 5a we provide the results when asking annotators to pick the most human-like story. We use the majority vote to decide the best model per story. The generated stories outperform the AREL baseline (73.87% *vs*. 22.53%). Surprisingly, in many cases, the annotators found the generated stories to be more human-like than the ground truth stories (41% *vs*. 48.57%).

Method	Focused	Coherent	Share	Human-like	Grounded	Detailed
AREL	3.49	3.18	3.18	3.26	3.32	3.15
Ours	3.67	3.52	3.20	3.56	3.54	3.32
GT	**3.72**	**3.57**	**3.34**	**3.64**	**3.56**	**3.53**

Table 4: Human evaluation results for rating survey (scores are between 1-5).

Model		Text Rep.	Sent. Rep.
AREL (Wang et al., 2018)		0.16	0.4
BOG prior	**Intra-repetition reg.**		
No	No	0.14	0.33
Yes	No	0.10	0.18
No	Yes	0.04	0.04
Yes	Yes	**0.008**	**0.0**

Table 5: Story generation ablation analysis.

To further evaluate the quality of the stories, we follow the criteria set by the Visual Storytelling Challenge[1] and conduct a survey where judges are asked to rate six categories between 1-5: *1. Focused*: the story contains information that is "naturally" relevant to the rest of the story; *2. Coherence*: the sentences in the story are related and consistent; *3. Share*: the inclination to share the story; *4. Human-like*: the story was likely written by a human; *5. Grounded*: the story directly reflects concrete entities in the image; and *6. Detailed*: the story provides an appropriate level of detail. To obtain the final score, we average the annotators' scores per sample, followed by averaging across the entire sample set. From Tab. 4 we observe: the model improved on all the criteria compared to the AREL model. Importantly, the generated stories are comparable to the ground-truth stories, indicating success in reducing the shortcomings found in prior methods. Nonetheless, the level of detail is still lacking, supporting the observation of Holtzman et al. (2020) that current decoding strategies tend to generate well-formed yet somewhat generic text.

5 Conclusion

We present a novel approach for VST, which encourages coherency of generated story. We incorporate structure between images with a new attention method that selects the important objects in an ordered image sequence. Human evaluation and quantitative analysis demonstrate that the approach outperforms existing methods. Further, we perform ablation analysis to show effectiveness.

[1] `http://visionandlanguage.net/workshop2018`

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Kobus Barnard, Pinar Duygulu Sahin, David A. Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan. 2003. Matching words and pictures. *JMLR*.

Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. 2017. Mutan: Multimodal tucker fusion for visual question answering. In *ICCV*.

Xinlei Chen and C. Lawrence Zitnick. 2015. Mind's eye: A recurrent visual representation for image caption generation. In *CVPR*.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*.

Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. 2019. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *CVPR*.

Diana Gonzalez-Rico and Gibran Fuentes Pineda. 2018. Contextualize, show and tell: A neural visual storyteller. In *Storytelling Workshop, NAACL*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. In *CVPR*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *ICLR*.

Xudong Hong, Rakshith Shetty, Asad Sayeed, Khushboo Mehra, Vera Demberg, and Bernt Schiele. 2020. Diverse and relevant visual storytelling with scene graph embeddings. In *CONLL*.

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 *AAAI*.

Qiuyuan Huang, Zhe Gan, Asli Çelikyilmaz, Dapeng Wu, Jianfeng Wang, and Xiaodong He. 2018. Hierarchically structured reinforcement learning for topically coherent visual story generation. In *AAAI*.

Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross B. Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *NAACL*.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018a. Bilinear attention networks. In *NeurIPS*.

Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2017. Hadamard product for low-rank bilinear pooling. In *ICLR*.

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

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

Jiacheng Li, Haizhou Shi, Siliang Tang, Fei Wu, and Yueting Zhuang. 2019. Informative visual storytelling with cross-modal rules. In *MM*.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *NeurIPS*.

Yatri Modi and Natalie Parde. 2019. The steep road to happily ever after: an analysis of current visual storytelling models. In *Workshop on Shortcomings in Vision and Language, NAACL*.

Cesc C. Park and Gunhee Kim. 2015. Expressing an image stream with a sequence of natural sentences. In *NeurIPS*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *IEEE PAMI*.

Idan Schwartz, Alexander G. Schwing, and Tamir Hazan. 2017. High-order attention models for visual question answering. In *NeurIPS*.

Idan Schwartz, Seunghak Yu, Tamir Hazan, and Alexander G Schwing. 2019. Factor graph attention. In *CVPR*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *NeurIPS*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. In *CVPR*.

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 *NeurIPS*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and tell: A neural image caption generator. In *CVPR*.

Ruize Wang, Zhongyu Wei, Piji Li, Qi Zhang, and Xuanjing Huang. 2019. Storytelling from an image stream using scene graphs. In *AAAI*.

Xin Wang, Wenhu Chen, Yuan fang Wang, and William Yang Wang. 2018. No metrics are perfect: Adversarial reward learning for visual storytelling. In *ACL*.

Huijuan Xu and Kate Saenko. 2016. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

Pengcheng Yang, Fuli Luo, Peng Chen, Lei Li, Zhiyi Yin, Xiaodong He, and Xu Sun. 2019. Knowledgeable storyteller: A commonsense-driven generative model for visual storytelling. In *IJCAI*.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J Smola. 2015. Stacked attention networks for image question answering.

Youngjae Yu, Jiwan Chung, Heeseung Yun, Jongseok Kim, and Gunhee Kim. 2021. Transitional adaptation of pretrained models for visual storytelling. In *CVPR*.

Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. 2018. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. In *NeurIPS*.

Bowen Zhang, Hexiang Hu, and Fei Sha. 2020. Visual storytelling via predicting anchor word embeddings in the stories. In *ICCV*. Workshop on Closing the Loop Between Vision and Language.

In this supplementary material, we provide explanations for the image representations (see Sec. B), scalars analysis (see Sec. C), and additional qualitative results (see Sec. D).

A Training Setup

We extracted the image features using a pre-trained F-RCNN model with a ResNet152 backbone (He et al., 2015; Ren et al., 2015; Anderson et al., 2017). We set the number of extracted objects $K = 36$. Bounding box coordinates were normalized between 0 and 1. Words that appear less than 3 times in the training set are represented by an <*UNK*> token. The vocabulary size is 12,210 words. Word representations were initialized using GloVe embeddings (Pennington et al., 2014). We set the decay parameter $\pi = 2$ and the image representation dimension $d = 512$. We set the dropout parameter to 0.3. We use cross-entropy loss to maximize likelihood of ground-truth stories. At decoding time we employ a beam search algorithm, with beam width set to 3. We use Adam (Kingma and Ba, 2014) optimizer with a learning-rate of 0.0004, which is decayed by a factor of 0.8 if the validation score (METEOR) does not improve after 4 epochs. The total amount of trainable parameters is 13,092,194. Training converges after ∼20 epochs. Each epoch needs 20 minutes on an Nvidia V100 GPU.

B Image Representation

An initial pre-processing step represents each of the input images I_i via K regional features $r_{i,k} \in \mathbb{R}^d, 1 \leq k \leq K$. For this we use bottom-up attention features (Anderson et al., 2017). Specifically, for each image I_i we first extract the top K region features $e_{i,k} \in \mathbb{R}^m$. Hereby, $e_{i,k}$ is an m-dimensional feature vector extracted from a pre-trained image classification network (He et al., 2015) along with their respective bounding boxes $b_{i,k} \in \mathbb{R}^4$, and classes $c_{i,k} \in \mathbb{N}$. The final d-dimensional representation $r_{i,k} \in \mathbb{R}^d$, of each region is defined by a combination of the extracted semantic features. Formally,

$$r_{i,k} = W_r[W_e e_{i,k} + W_b b_{i,k} + E_c(c_{i,k})], \quad (19)$$

where $W_r \in \mathbb{R}^{d\times d}$, $W_e \in \mathbb{R}^{d\times m}$, $W_b \in \mathbb{R}^{d\times 4}$, and E_c are trainable parameters shared between all images. We set $K = 36$ in our proposed model. Biases and normalization are omitted for readability.

C Factors Importance Analysis

In Fig. 6, we illustrate for each sentence, the value of the importance calibration scalars (*i.e.*, α_i^s and $\alpha_{i,s}^s$ in Eq. 2,3, and 4). Intuitively, these values indicate the importance of different image-to-image messages. We focus our analysis on the sentence-corresponding image (*i.e.*, $i = s$ in Sec. 3.1). We observe that the self-message scalars (*i.e.*, $\mu_{i\to i}$) of the sentences in the middle of the sequence, *i.e.*, sentences (2,3, and 4), are low. This indicates that the images in the middle of the sequence rely more on the other images. The beginning and the ending of the story depend more on the local factors. Notably, the most substantial influence is given to the following image (*i.e.*, $\alpha_{i,i+1}^i$). This means that while generating the current sentence, the OIA decision is based mostly on the next image. This is intuitive as it helps to advance the narrative in a desired direction.

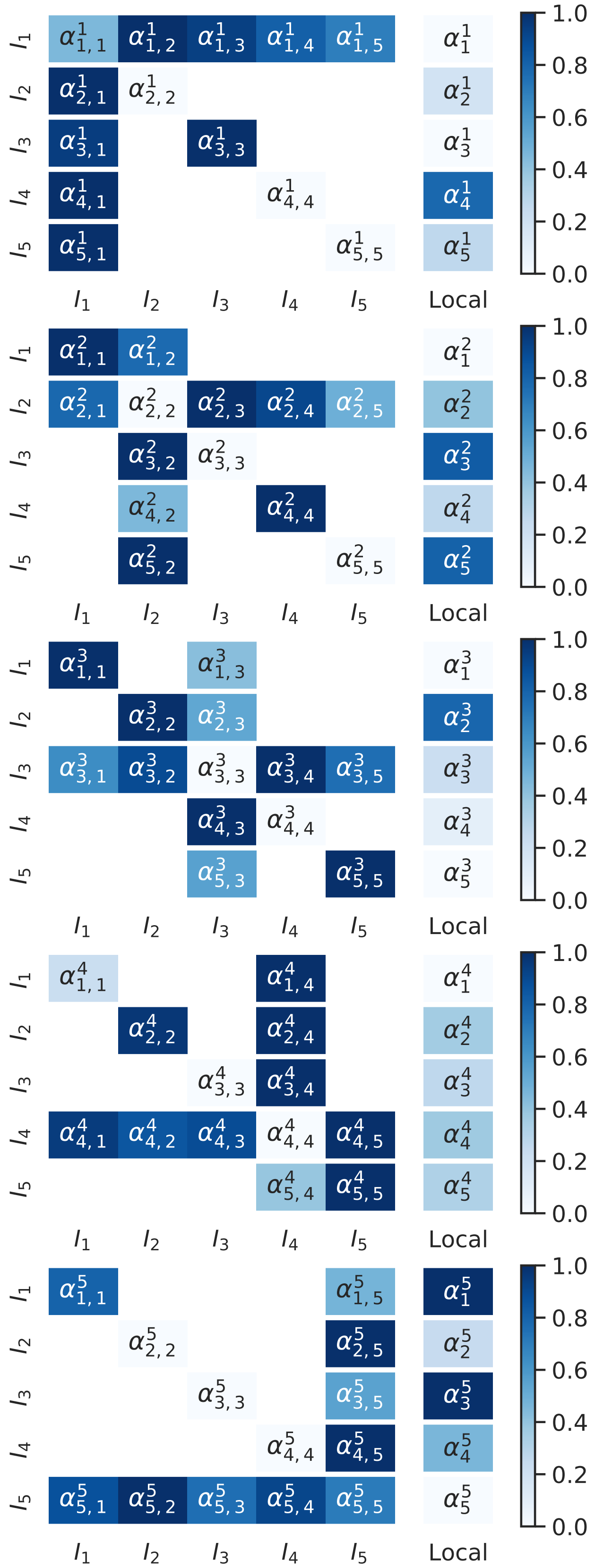


Figure 6: OIA scalar values (*i.e.*, α_i^s and $\alpha_{i,s}^s$ in Sec. 3.1.1). The top map corresponds to the first sentence (*i.e.*, $s = 1$) and bottom one to the last sentence (*i.e.*, $s = 5$)

.

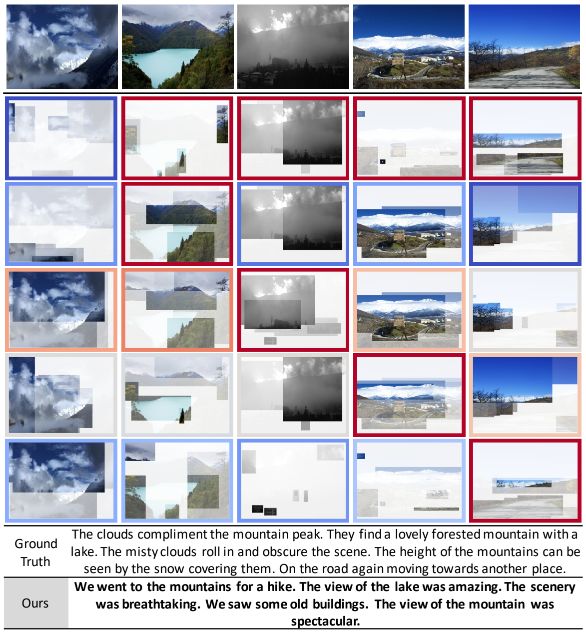


Figure 7: Illustration of OIA and ISA attention maps, the ground-truth story and the final generated story. Each row corresponds to a story sentence and shows objects OIA highlights. The attended images' border specifies the relevancy to sentence generation, from red (important) to blue (not important).

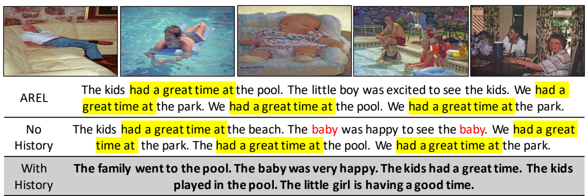


Figure 8: An illustration of an image sequence along with three different stories generated by: (1) AREL baseline (Wang et al., 2018), (2) No History: a model without intra-repetition regularization and BOW prior (see Sec. 3.3); and (3) With History: the final model. Repeated sentences are highlighted with a yellow colored marker. Repeated words in a sentence are emphasized in red color.

D Qualitative Results

In Fig. 7 we illustrate the attention maps along with the generated story. The first sentence, "We went to the mountains," sets the theme for the story, which requires the processing of subsequent images. Notably, the ISA module picked the subsequent images. In contrast, for the second sentence, the attention focuses mainly on the second image resulting in a description of the lake observed exclusively in this image. The third sentence relates to the scenery. Hence the attention focuses on preceding and subsequent images.

In Fig. 8, we show the ability of the method in reducing repetitions. We observe the AREL baseline to repeat the same sentences, for example, "...had a great time at...". We also observe this repetitiveness when we remove the bag-of-words prior and the intra-sentence regularization (*i.e.*, No History column). Nevertheless, the method remains on topic, *i.e.*, family in the pool.

In Fig. 9-12, we show samples used for human evaluation (see Sec. 4.3 of the main paper). We present two sequences for the coherence and human-like categories, where judges preferred our generated stories, and another two, where the judges chose the human formulated story.

In Fig. 13, we illustrate the attention output of the OIA and the ISA modules and show two stories. The first story is generated using OIA with direction-based factors. The other story is generated using the same interaction factors for both past and future interactions. In the first sequence, our direction-based model infers the topic of the story, *i.e.*, "hiking in the woods" and sets the theme in the first sentence by mentioning the word 'hike.' In contrast, our model without direction mentions park but misses the hiking topic.

In Fig. 14, we show two failure cases. The first story is incoherent, *e.g.*, the gender switches between male and female. In the second story, the model reduces word repetition by using synonyms. *E.g.*, the sentences, "I had a great time," "I was a lot of fun" and "... having a good time" have the same meaning.

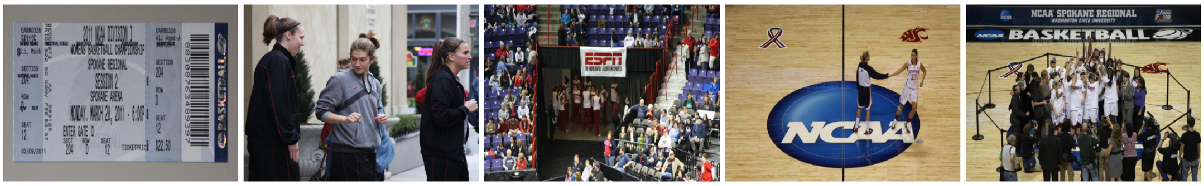
Ours: The girl was having a great time at the party. We had a lot of fun. We were so happy to see each other. The table was set for the reception. It was a great day.

GT: I went to the wedding last weekend. The tables were beautiful. I had a great time there. The entire family was there. It was so much fun.

Ours: Today was the day of the wedding. The bride and groom were ready to be married. The cake was amazing. The bride and groom pose for a picture. The ceremony was beautiful.

GT: Soon to be husband waiting on his bride. Here comes the newlyweds. Very plain birthday cake , looks delicious. Great photo of the wedding people. They bride is happy.

Figure 9: Sample of a sequence where the *coherence-score* of our story was rated higher than the human story. Note: the coherence-score assesses whether the sentences in the story are related and consistent.



Ours: Today was the day of the organization. There were a lot of people there. The crowd was ready to start. This is a picture of the game. The cheerleaders had a great time.

GT: The women 's basketball game was today. The team members were seen on the campus before the game. Everyone got into their uniforms to play. After the game two of the players were seen shaking hands. The team ended up winning the game.

Ours: It was a cold day. This is a picture of a sign. This is a picture of the building. After the night , we decided to go to the bar. I had a great time at the location.

GT: A group of friends decided to take a road trip through location location. They stopped in a town called location. It was n't very lively out , despite being friday night. They walked around looking for a place to eat. They finally found an open cafe and had some mexican-american grub.

Figure 10: Sample of a sequence where the *coherence-score* of ground truth story was rated higher than the our story. Note: the coherence-score assesses whether the sentences in the story are related and consistent.

Ours: The cake was decorated for the wedding. It was a great day to celebrate. The bride and groom were very happy. There was a piece of art. The bride had a great time.

GT: I love my beautiful wedding cake. Napkins with our wedding date on them. Everyone standing waiting on me to come down the aisle. I love the cross ! here i am with my sister

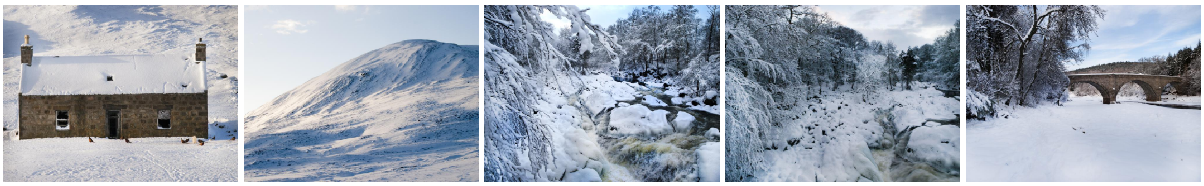
Ours: We went to the city to see the old buildings. The streets were beautiful. There was a lot of people in the market. There was a lot of food there. There is a variety of fresh vegetables.

GT: I love the architecture of the old city the streets are narrow and everyone drive small cars local food is very good and very inexpensive all organic fruit and vegetables fruit and vegetables were cheap and delicious

Figure 11: Sample of a sequence where the *human-like-score* of our story was rated higher than the human story. Note: the human-like-score assesses whether a human likely wrote the story.

Ours: The mountains were beautiful. The view from the lake was amazing. The view of the city was spectacular. The buildings were very tall. The mountains are breathtaking.

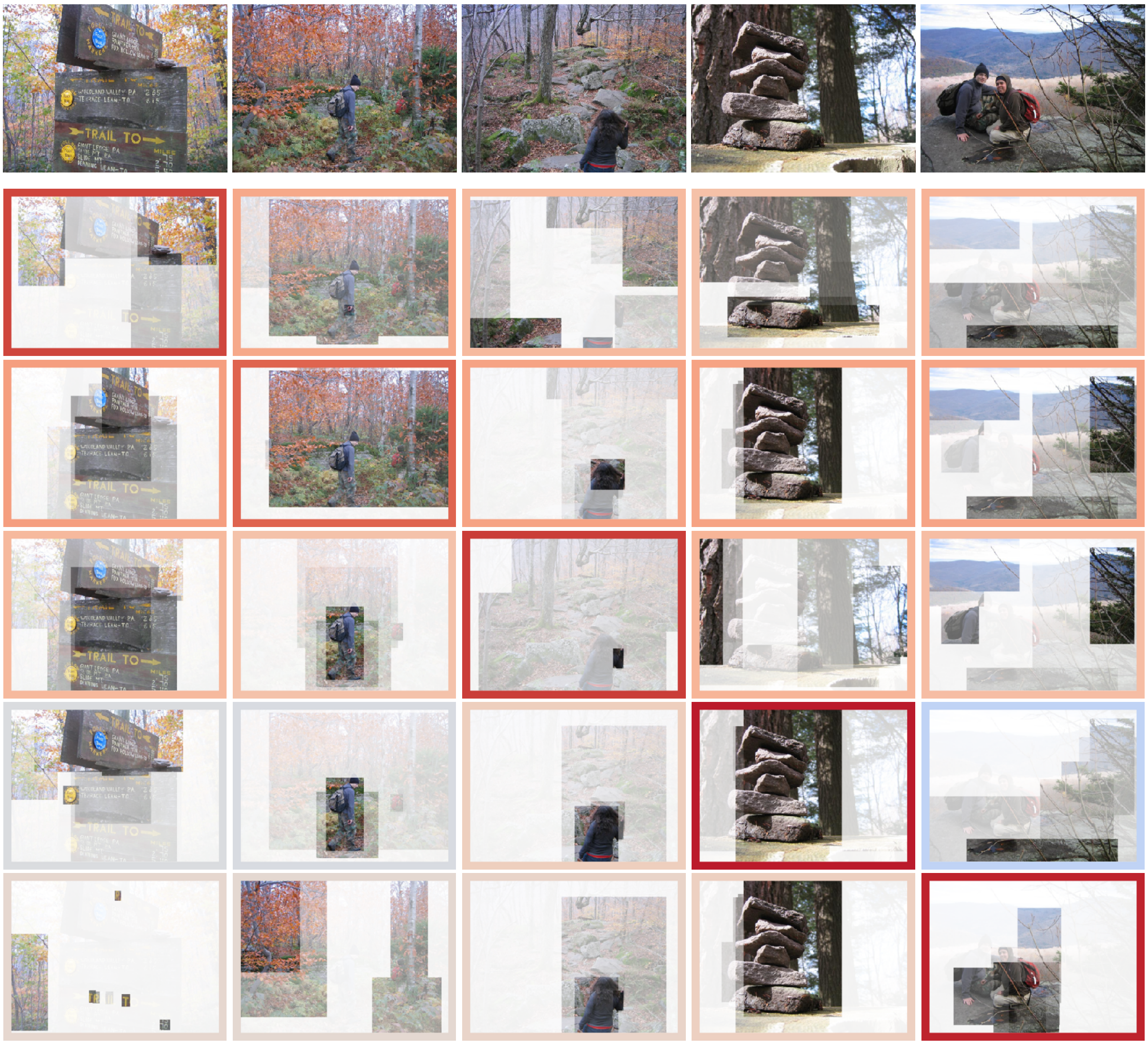
GT: So this is where all the beautiful landscape screenshots come from ! is this a place to live forever , or what ? oh , my gosh , just build me a tiny house and i 'm all here ! and castles - they have castles , too ! what an impressive mountain !



Ours: The house was covered in snow. The mountains were amazing. The mountain was beautiful. The trees were covered in snow. We took a picture of a bridge.

GT: We recently headed up to my family 's vacation house in the mountains. In our backyard there is this beautiful mountain. We often go for walks on the various trails around the house. We often see a lot of animals and all of their tracks on the trails. At the end of the trail in this beautiful bridge that goes over a frozen river.

Figure 12: Sample of a sequence where the *human-like-score* of ground truth story was rated higher than the our story. Note: the human-like-score assesses whether a human likely wrote the story.



No Direction: We went to the park. There were a lot of people in the woods. We took a picture of the trees. We saw some interesting things to see. It was a great day.

With Direction: We went to the park for a hike. We took a walk through the woods. We had to take a picture of the trees. There were many interesting statues in the woods. The view from the top was beautiful.

GT: There was a big sign that the guy passed. The girl also saw it near the big rocks before they took a photo.

No Direction: The bride and groom were ready for the wedding. The flower was beautiful. The bride and groom are having a great time. The couple of friends had a good time. The family was happy to be married.

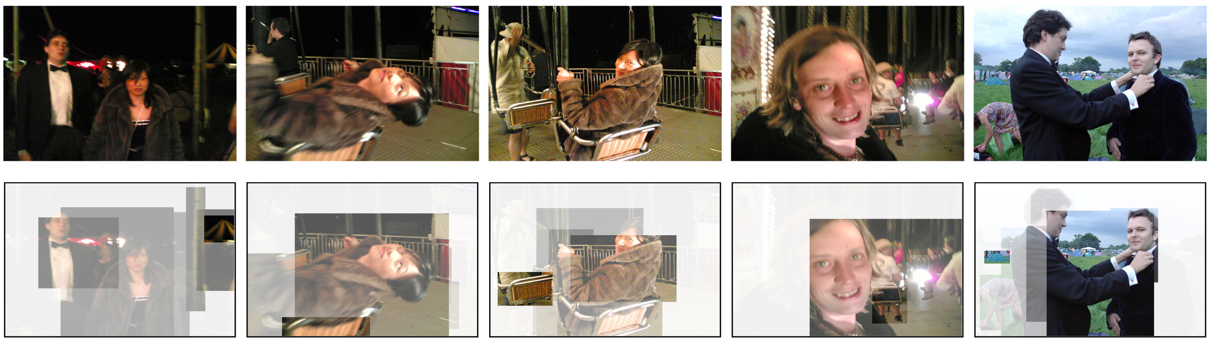
With Direction: The wedding party was beautiful. There were a lot of beautiful flowers. The party was a lot of fun. This man and his friend are having a good time. After the ceremony , we all got togethter for a picture.

GT: The wedding last week was beautiful. I brought a lot of flowers for the bride and groom. All of the friends and family were there to show their support. I took a ton of pictures while i was there. Everyone was dressed up very nicely.

Figure 13: Additional qualitative results. Each story is displayed along with 5 images with the attention maps generated by OIA for each sentence. Additionally, the border of each image indicates the attention score of each image in the image-sentence attention (ISA) module. Our model with and without order and ground-truth stories are also provided.

OURS: I was so excited to be at the school yesterday. [male] was very happy to see him. She had a great time. She did n't want to take a picture with her. [male] is the best of the day.

GROUND TRUTH: Mr. Green is a teacher at a local high school. Mr. Green teaches music. Mr. Green has enjoyed teaching music for over 30 years. Mr. Green makes sure each of his students thoroughly understands each subject. Mr. Green 's favorite part of any of his lessons are student questions.



OURS: We went to the party last night. This guy is getting ready to go. I had a great time. I was a lot of fun. The man and his friends were having a good time.

GROUND TRUTH: A pristine night at the carnival after the opera. It feels good to unwind after such a regal event. Makes you feel like a kid after all. And really brings out the smiles inside everyone. Before we return to the prim and proper world we are accustomed to.

Figure 14: Illustration of failure cases.