SCO-VIST: Social Interaction Commonsense Knowledge-based Visual Storytelling

Anonymous EMNLP submission

Abstract

Visual storytelling aims to automatically generate a coherent story based on a given image sequence. Unlike tasks like image captioning, visual stories should contain factual descriptions, worldviews, and human social commonsense to put disjointed elements together to form a coherent and engaging human-writeable story. However, most models mainly focus on applying factual information and using taxonomic/lexical external knowledge when attempting to create stories. This paper introduces SCO-VIST, a framework representing the image sequence as a graph with objects and relations that includes human action motivation and its social interaction commonsense knowledge. SCO-VIST then takes this graph representing plot points and creates bridges between plot points with semantic and occurrence-based edge weights. This weighted story graph produces the storyline in a sequence of events using Floyd-Warshall's algorithm. Our proposed framework produces stories superior across multiple metrics in terms of visual grounding, coherence, diversity, and humanness, per both automatic and human evaluations.

1 Introduction

Beyond interpreting the factual content of scenes with expressions, like image captioning, Visual Storytelling (VST) aims to conduct a human-like understanding of the idea of a sequence of images and generate more complicated visual scenarios with human-like textual expressions (Huang et al., 2016). In order to achieve this aim, the AI agent is required to model relationships between the images while remaining visually grounded, identify concepts that are implied (but not explicitly shown) in the images, as well as generate coherent, conversational language resembling how a human would tell a story in a social setting.

Numerous past studies have employed encoder-decoder frameworks that first utilise a computer vision algorithm to extract image-specific features, which are then fed into a language generation model to decode the story (Gonzalez-Rico and Pineda, 2018; Kim et al., 2018; Jung et al., 2020; Smilevski et al., 2018). Although these methods can yield reasonable stories to some extent, they often lack common sense reasoning, thus producing stories that are "generic" sounding with limited vocabulary, and irrelevant to the images. To alleviate these issues, more recent approaches adopt content planning methods that try to explicitly predict textual concepts from the images via detecting objects in the image by using external knowledge data sources to identify implicitly related concepts (Chen et al., 2021; Hsu et al., 2020, 2021a; Xu et al., 2021). Those external knowledge data sources mainly comprise taxonomic, lexical and physical relations, whereas human-like storytelling tends to use the social-aspect relations of everyday human experiences. Social-interaction relations comment on socially-triggered states and behaviours. It is crucial to gauge people's intentions and purpose and predict situationally-relevant human reactions and behaviours, which is directly aligned with the aim of human-like storytelling.

This paper proposes a new social-interaction commonsense-enhanced VST framework, SCO-VIST, for producing human-like stories by interpreting socially-triggered situations and reactions. We introduce a three-stage commonsense enhanced framework that attempts to construct a reasonable plot of story events from the given image stream for story decoding. Stage 1 focuses on constructing a story graph representing causal and logical relationships between social interactions and events. Motivated by the idea that captions may already have embedded social commonsense within them, we first generate a caption for each image to literally capture the event depicted in the photo. Additionally, we further extract commonsense from external data related to social situations, interactions and behavioural responses (i.e. character's intentions, desires or needs). Each extracted caption and commonsense is thus considered a different event or plot point, and we connect the plot points (nodes) with causal ordering. In stage 2, we convert the story graph to be weighted by conducting a comprehensive analysis on different edge weight assignment methods based on semantic similarity between nodes and graph learning. Intuitively, this weighted story graph reflects the branching space of plausible event continuations where the edge weights indicate the likelihood of transition between connected plot points. Given the weighted story graph, the optimal storyline is the path of nodes that yields the largest sum of weights from the left to the right-most nodes in the graph. Therefore, Stage 3 negates the edge weights and employs Floyd-Warshall's shortest path search algorithm to extract the optimal sequence of story events which is later fed into a Transformer for story generation. The main contributions of this research are: 1) We introduce a social-interaction commonsense enhanced VST framework that improves understanding of social situations and characters' feelings, 2) We design a heterogeneous story graph and conduct a comprehensive analysis of the role of node and edge construction and learning over the visual storytelling dataset, 3) We show that our model outperforms state-of-the-art when comparing automatic metrics, especially when analysing recently proposed metrics designed for VST, and 4) For robust evaluation, we also conduct human evaluation studies and demonstrate that our framework consistently and significantly outperforms several strong baselines.

2 Related Work

Earliest works on VST consist of an encoder-decoder structure incorporated in an end-to-end model (Gonzalez-Rico and Pineda, 2018; Kim et al., 2018; Smilevski et al., 2018). Recently, there has been increasing interest in reinforcement learning architectures which include a reward model to evaluate the generated stories (Hu et al., 2020; Wang et al., 2018). However, the training process of such methods are inherently unstable. Other approaches first translate images to semantic scene graphs to capture image features and then employ Graph Convolutional Networks (GCN) to enrich regions and object representations (Hong et al., 2020; Wang et al., 2020). Instead, we use literal text descriptions of images which can better explicitly represent the image contents.

To promote more diverse stories, newer works have also used knowledge graphs to assist the storytelling process, allowing for richer stories capable of expressing imaginative concepts that are not explicitly shown in the image scene. Most of these methods involve querying ConceptNet (Speer et al., 2017) with detected image objects or predicted key image concepts to find a set of related candidate concepts (Chen et al., 2021; Xu et al., 2021; Yang et al., 2019). While these methods show promising improvements in outputs, ConceptNet mainly comprises of taxonomic and physical relations, whereas our framework leverages commonsense that are more social-interaction focused and event-centred. Finally, most related to our work, recent studies try to form the story plot by first using external knowledge to connect concepts between images to reason about potential temporal relationships (Hsu et al., 2020, 2021a,b). However, these methods often employ complex network architectures to iteratively predict subsequent events. We alleviate these complexities and present a simple yet effective approach for storyline construction.

3 Method

Figure 1 depicts an overview of SCO-VIST's three stages. The following sections will describe each step in detail.

3.1 Stage 1: Story Graph Construction

Node Construction The story graph contains 3 types of nodes: caption, commonsense and theme nodes. The caption nodes are obtained by using a pre-trained image captioning model to generate a textual description for each image in the photo sequence. That is, given the sequence of 5 images, captions $\{C_1, C_2, ...C_5\}$ are generated where C_i is the caption for the i^{th} image. The intuition behind using captions is that literal descriptions of an image can provide more specific and accurate details about image contents compared to the raw visual features extracted from the image itself. Moreover, this step mimics how a human would tackle the VST task, as one would usually first consider what is visually represented in the image and its context before forming the premise of the story.

Next, we specifically focus on generating commonsense related to social interactions and dynamic aspects of everyday events. As such, Comet-

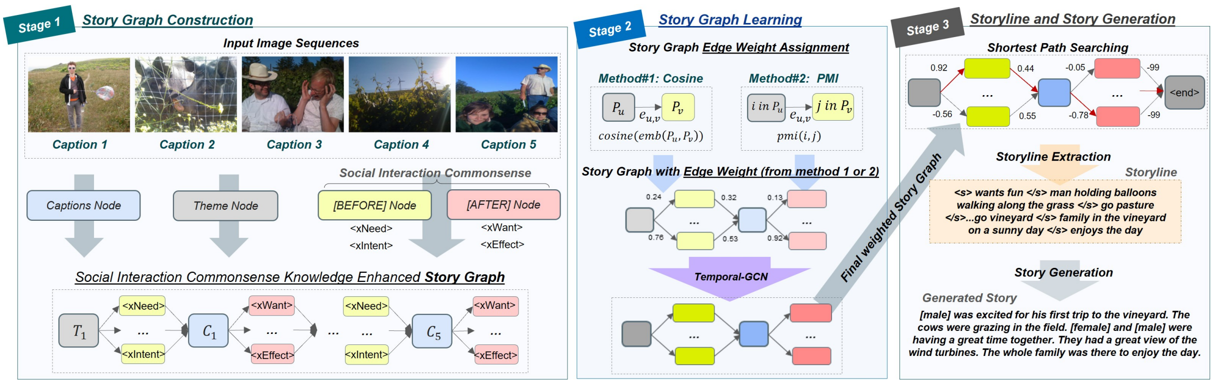


Figure 1: SCO-VIST's proposed framework. In Stage 1, the caption, theme and commonsense nodes are created and connected with causal ordering to form the story graph. In Stage 2, edge weights are assigned using cosine similarity or point mutual information and further refined through graph learning. Stage 3 takes the final story graph, negates the weights and constructs the storyline by finding the shortest path between the left and right-most node. The storyline is then fed to a Transformer for story generation. The corresponding detailed view of the final story graph for this example is depicted in Appendix F.

ATOMIC2020 is utilised, a '*neural knowledge model*' trained on the ATOMIC commonsense knowledge graph dataset (Hwang et al., 2021) which contains information on common human everyday experiences and mental states. Given a head/source phrase and relation (e.g. eat a cake `Intent`), Comet-ATOMIC2020 is capable of producing a tail phrase on-demand (e.g. celebrate birthday). Thus, out of the available 9 social interaction relations that Comet-ATOMIC2020 offers, we select 4 relations that primarily focus on causal and behavioural relationships: `xNeed`, `xIntent`, `xEffect` and `xWant`. More specifically, the `xNeed` relation indicates what event is needed to happen *before* a following event occurs while the `xIntent` relation indicates a character's intention *before* an action takes place. Conversely, `xEffect` are social actions that occur *after* an event while `xWant` represents a character's postcondition desires *after* an event.

We append the 4 relation tokens to each caption phrase C_i to provide as input for querying Comet-ATOMIC2020. Five commonsense inferences are generated per relation r, $\{ck_1^r, ck_2^r, ..., ck_n^r\}$, resulting in 20 commonsense altogether for each caption. The commonsense produced for each caption are then grouped into `BEFORE` and `AFTER` events. The `BEFORE` events category contains the knowledge extracted from the `xNeed` and `xIntent` relation while the `AFTER` events contains the `xEffect` and `xWant` commonsense.

Finally, the theme nodes contain a sequence of concepts that represent the theme depicted in each image. We use Clarifai [1], a pretrained object and concept detector model capable of predicting 11,000 unique concepts. We extract a sequence of 20 concepts for each of the 5 images to create 5 theme nodes $\{T_1, T_2, T_3, T_4, T_5\}$.

Connecting Nodes Let $CK_B = \{ck_1^r, ck_2^r, ..., ck_m^r\}$ where $r \in \{\texttt{xNeed}, \texttt{xIntent}\}$ be the `BEFORE` commonsense inferences for caption C_i. Similarly, we denote $CK_A = \{ck_1^r, ck_2^r, ..., ck_m^r\}$ where $r \in \{\texttt{xEffect}, \texttt{xWant}\}$ to be the `AFTER` commonsense inferences. To construct the story graph, we add directed edges between T_i (the theme node for image i) with the commonsense nodes in CK_B. Each node in CK_B is then connected to C_i which is further connected to each node in CK_A. Finally, each node in CK_A is connected with the theme nodes for the next image, T_{i+1}. Consequently, a directed acyclic graph S_G representing the branching space of possible story events for each image stream is constructed as seen in the graph in Stage 1 of Figure 1.

3.2 Stage 2: Story Graph Learning

This stage conducts an analysis on the importance and role of each node in the story graph by converting S_G into a weighted graph, $S_{G,weighted}$. Two main methods for edge weight assignment based on semantic similarity is experimented with and weights are further refined with graph learning.

Cosine Similarity Firstly, we use the cosine sim-

[1] www.clarifai.com

ilarity between plot points as an indicator of their level of association. Given connecting nodes u and v which contain words or a phrase denoted by P_u and P_v respectively, we convert P_u and P_v to a sentence embedding using a pretrained transformer model. The cosine similarity score between the two embeddings at node u and v is then simply assigned to their connecting edge $e_{u,v}$.

Pointwise Mutual Information (PMI) The second method computes the PMI between each pair of words in P_u and P_v where a high PMI implies high semantic correlation between words. Formally, the PMI between word i in P_u and word j in P_v is:

$$\mathrm{PMI}(i,j) = \log \frac{p(i,j)}{p(i)p(j)} \tag{1}$$

Here, $p(i,j) = \frac{\#S(i,j)}{\#S}$, $p(i) = \frac{\#S(i)}{\#S}$ and $p(j) = \frac{\#S(j)}{\#S}$ where $\#S(i)$ is the number of sentences in the corpus that contain word i, $\#S(i,j)$ is the number of sentences that contain both words and $\#S$ is the total number of sentences in the corpus. Finally, a normalized version of the PMI score is calculated:

$$\mathrm{NPMI} = \frac{\mathrm{PMI}}{-\log(p(i,j))} \tag{2}$$

The final weight assigned to $e_{u,v}$ is the maximum NPMI score out of all scores calculated from the possible word pair combinations.

Graph Learning We further refine the cosine or PMI-weighted story graph through graph learning. Specifically, the weighted graph is fed into a Temporal Graph Neural Network (TGCN). Such networks combine the advantages of GCNs and Recurrent Neural Networks to learn the graph's complex topological structure as well as its temporal changes. We use an implementation of the Gated Graph Convolution Long Short Term Memory Layer (Taheri and Berger-Wolf, 2019) which encodes the graph and yields embeddings for each node. We then extract the 5 embeddings from the caption nodes and feed them through the BART Transformer (Lewis et al., 2020) to decode the story. The TGCN and Transformer are trained end-to-end to minimise the cross-entropy loss:

$$L(\theta) = -\sum_{t=1}^{T} \log(p_\theta(y_t^*|y_1^*, ..., y_{t-1}^*)) \tag{3}$$

where θ is the parameters of the model, y^* is the ground-truth story and y_t^* denotes the t-th word in y^*. Finally, we extract the learnt node embeddings and compute the cosine similarity between the embeddings of each pair of connected nodes to obtain the edge weight in between.

3.3 Stage 3: Storyline and Story Generation

Storyline Extraction Given $S_{G,weighted}$, we consider the optimal storyline as the path from the left-most node to the right-most node that produces the highest sum of weights. To find this path, we negate each weight in $S_{G,weighted}$ and add a dummy end node D_E which is connected with the right-most nodes in $S_{G,weighted}$ with an edge weight of -99. An example of the final graph is depicted in Appendix F. Floyd–Warshall's algorithm (Floyd, 1962) is then adopted to find the shortest path starting from T_1 to D_E to produce the storyline containing a sequence of events $e_1, ..., e_L$ taking only the caption and commonsense nodes.

Story Generation The last stage consists of decoding the story. We separate each event e_i using a separator token `</s>`. The events are then fed through BART for story generation which we train with the cross-entropy loss from Equation 3.

4 Evaluation Setup[2]

4.1 Data

VIST The Visual Storytelling Dataset (VIST) (Huang et al., 2016) consists of 210,819 unique images obtained from Flickr albums. The dataset is split into training/validation/testing with 8,031/998/1,011 albums where each album contains a set of similar image sequences with each sequence made up of 5 photos. Each album also has 5 human written stories where each story is usually comprised of one sentence per image. The unique number of stories in the training, validation and testing set is 40,155, 4,990 and 5,055 respectively.

4.2 Baseline Models

We compare our model with 6 state-of-the-art baselines: 1) **AREL** (Wang et al., 2018), 2) **GLACNet** (Kim et al., 2018), 3) **KG-Story** (Hsu et al., 2020), 4) **ReCo-RL** (Hu et al., 2020), 5) **PR-VIST** (Hsu et al., 2021a), and 6) **TAPM** (Yu et al., 2021). Details of each model is outlined in Appendix E.

[2] Implementation details can be found in Appendix C

4.3 Ablation Study Models

We also conduct ablation studies to compare different variants of our proposed model:

- **SRL-caption**: A story graph is not created and the 5 image captions are used as the storyline.
- **SRL-pmi/cosine**: The storyline is extracted from the story graph using weights obtained from the cosine similarity or PMI approach.
- **TGCN/TGCN-SRL**: TGCN-cosine/pmi is an end-to-end model where the story graph is fed to the TGCN and node embeddings are then inputted into BART for story decoding. The story graph input uses weights obtained from either the cosine or PMI approach. TGCN-SRL-cosine/pmi further uses the trained TGCN to extract the node embeddings and their similarities are then used to refine the story graph weights for storyline and story generation.

4.4 Automatic Metrics

Numerous past literature have shown that traditional automatic metrics like BLEU correlate poorly with human judgement and are unreliable for evaluating VST (Wang et al., 2018; Hsu et al., 2019). These metrics mainly focus on comparing n-gram similarity between hypothesis and references, thus are insufficient for evaluating open-ended text generation tasks like storytelling, where there are multiple plausible outputs for the same input which are not fully reflected in the references. Therefore, we focus on metrics specifically designed for '*open ended text generation*' which consider the plausibility of diverse outputs. The first is RoViST (Wang et al., 2022), an unreferenced metric set for VST consisting of three scores that target three criteria: visual grounding (RoViST-VG), coherence (RoViST-C) and no redundant repetition of concepts/words (RoViST-NR). An overall single score (RoViST) can be calculated by averaging RoViST-VG, C and NR. In addition to RoViST, we consider other learnt *'unreferenced'* metrics such as Perplexity and the storytelling metric, UNION (Guan and Huang, 2020) which assigns a score based on important story criteria like coherence, no conflicting logic and non-repeating plots. Finally, for completeness and maintaining consistency with other works, we further compute reference-based metrics including the classic ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016). For analysing semantic similarity, the BERT-based metric BLEURT (Sellam et al., 2020) is further adopted as well as the embedding-based metric, MoverScore (Zhao et al., 2019).

4.5 Human Evaluation

We finally conduct human evaluation studies and create 3 surveys where each survey conducts a pairwise comparison between our model and a baseline. In the survey, participants are given 100 randomly selected unique photo sequences from the test data (same sequences are used for each survey) and the corresponding generated story from our model and the baseline. They are then asked to choose which of the two stories are better based on 3 criteria: 1) **Visual Grounding**: the generated story must relate to concepts depicted in the image sequence, 2) **Coherence**: story sentences need to flow while remaining logical and topically consistent, and 3) **Non-Redundancy**: sentences are diverse and there are no unnatural-sounding repetition of words/phrases in the story. A final question also asks the annotator to choose which story is better out of the two based on their opinion. 15 respondents (5 per survey) were recruited where each participant answered 400 questions, resulting in 6000 instances collected in total.

5 Results

5.1 Overall Performance

Table 1 summarises several metrics for the 6 baselines and for the 7 different variations of SCO-VIST. After filtering out broken images in the test set and missing stories from the baseline models, a sample of 890 albums was used to calculate these metrics. Considering our best model based on the visual storytelling metric RoViST (SRL-pmi), RoViST-VG performs on par with the more recent baselines and significantly outperforms in RoViST-C when considering all our 7 model variants. RoViST-NR however underperforms, but we strongly emphasize that this is most likely attributed to the short story lengths which have a lower chance of repeating words as can be seen by KG-Story which has a repetition score of 99.9 but average story length of only 32. Furthermore, we note that studies in Wang et al. (2022) emphasized that humans considered coherence to play the most significant role when judging a story, followed by visual grounding and non-redundancy. Nevertheless, our models still achieve noticeably

Model	RoViST-VG	RoViST-C	RoViST-NR	RoViST (R)	SPICE (S)	BLEURT (B)	MoverScore (M)	UNION (U)	Perplexity	R+S+B+M+U	Story Len.
AREL ((Wang et al., 2018))	66.2	57.1	83.4	68.9	9.0	32.6	55.1	17.1	15.3	182.7	44.8
GLACNet (Kim et al., 2018)	61.6	68.6	95.1	75.1	7.0	33.5	54.9	75.9	24.6	246.3	35.2
KG-Story (Hsu et al., 2020)	58.7	65.1	**99.9**	74.6	7.2	32.3	54.9	65.8	46.1	234.8	32.3
ReCo-RL (Hu et al., 2020)	67.8	57.3	91.9	72.3	11.2	31.9	55.4	23.8	28.3	194.6	49.3
PR-VIST (Hsu et al., 2021a)	70.0	60.4	96.1	75.5	9.6	31.0	54.7	30.3	42.3	201.1	52.2
TAPM (Yu et al., 2021)	70.3	67.0	90.5	75.9	9.9	33.4	55.6	56.0	18.3	230.8	51.2
SRL-caption	65.2	73.9	91.4	76.8	6.1	31.7	53.3	76.5	16.0	244.5	49.7
SRL-cosine	69.6	72.1	91.9	77.9	11.2	34.6	**56.0**	78.8	15.1	258.4	48.0
SRL-pmi	**70.4**	72.8	91.6	**78.3**	**11.5**	34.7	**56.0**	75.9	14.7	256.3	51.2
TGCN-SRL-cosine	70.3	72.3	90.5	77.7	10.9	**34.9**	**56.0**	84.0	14.9	**263.4**	**52.3**
TGCN-SRL-pmi	69.0	71.9	91.6	77.5	11.2	34.7	**56.0**	80.6	**13.6**	259.9	51.5
TGCN-cosine	65.7	75.5	91.8	77.6	9.2	33.9	55.6	84.3	16.5	260.6	39.1
TGCN-pmi	65.7	**75.9**	91.3	77.6	9.4	33.8	55.7	**87.0**	15.5	**263.4**	40.5

Table 1: Automatic metrics and average story length (Story Len.) for the 6 baselines vs. our 7 model variants.

better performance than the baselines when comparing the overall RoViST metric with SRL-pmi considered as the best model as it achieved a good balance of high scores across RoViST-VG, C and NR.

Although classic automatic metrics are known to correlate poorly with human judgement for VST, it is still noteworthy to analyse them in conjunction with RoViST. Hence, ROUGE-L, METEOR and CIDEr are shown in Table 4 of Appendix D where we observe that SRL-pmi resulted in lower scores. This could be due to our model using knowledge from COMET-Atomic2020 to enrich lexical diversity which results in lower performance in n-gram matching between the generated and reference stories. However, SRL-pmi still outperforms the baselines when comparing less classic metrics like SPICE which focuses on semantic propositional content, BLEURT which is based on semantic meaning and slightly on MoverScore which compares distances of word embeddings between reference and hypothesis stories. The unreferenced metrics for evaluating open-ended text generation, Perplexity and UNION also show significant improvements. Most noticeably, UNION which scores based on coherence, conflicting logic and chaotic scenes is able to reach an upper bound score of 87.0 with TGCN-pmi.

Finally, to gain a better overview of the overall performance, we sum RoViST, SPICE, BLEURT, MoverScore, and UNION and present the scores in the R+S+B+M+U column. When comparing the sum, the best performing models were the TGCN methods with TGCN-SRL-cosine and TGCN-pmi producing the highest scores.

5.2 Ablation Study

To analyse the effect of the storyline extraction stage and different edge weight assignment methods, an ablation study was conducted to compare the 7 different variations described in Section 4.3. We first compare just using the 5 captions (SRL-caption) as the storyline versus extracting the storyline from the commonsense story graph (SRL-cosine/pmi, TGCN-SRL-cosine/pmi). Surprisingly, competitive RoViST-C and NR scores was achieved from SRL-caption but underperforms substantially in the VG criteria. Additionally, SPICE, BLEURT, MoverScore, UNION and Perplexity were considerably worse. This implies that captions alone have sufficient commonsense embedded in them and can be useful features for generating plausible stories. However, the VG aspect can be further enhanced by exploiting extra social commonsense from external data.

Moreover, the TGCN-cosine/pmi approach consisting of the end-to-end model with a TGCN combined with the Transformer evidently produces lower RoViST-VG and NR compared to the SRL methods. SPICE, BLEURT, MoverScore and Perplexity scores were also mostly less optimal. This suggests that feeding the node embeddings into the Transformer for story decoding is not as good as extracting the storyline and explicitly using the words as input which can provide more fine-grained details about the image contents for generating richer stories. However, TGCN-cosine/pmi noticeably yielded the best RoViST-C scores out of the 7 methods (> 75). This could be attributed to the shorter outputs as it is often easier to stay coherent with shorter generic sentences.

Finally, it is interesting to note that higher UNION scores were obtained for all TGCN methods when compared to not using the TGCN. It is hypothesised that incorporating learnt temporal information in the node embeddings implicitly via TGCN training perhaps resulted in more logical stories, thus improving the UNION score.

5.3 Visualising Diversity

We visualise the number of distinct unigrams, nouns, verbs and adjectives outputted by SRL-pmi versus the 6 baselines. Figure 2 illustrates that our model can produce significantly more unigrams overall especially when comparing nouns, suggesting that leveraging social interaction commonsense and the captions can generate richer and diverse sentences with more novel expressions.

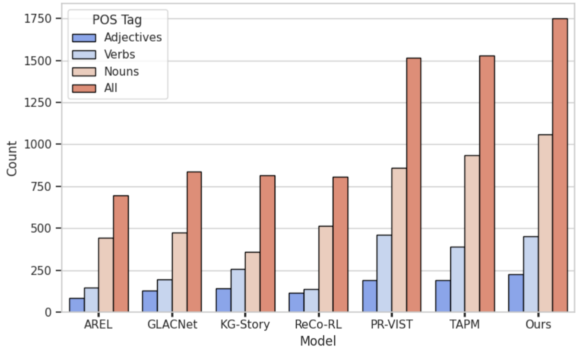


Figure 2: Count of unique unigrams for different part-of-speech (POS) tags for SRL-pmi vs. the 6 baselines.

5.4 Qualitative Analysis

To evaluate our model qualitatively, we show examples of generated stories from SRL-pmi versus the 5 baselines. Figure 4 illustrates that our model generates stories that are clearly more visually grounded. For instance, ReCo-RL in the first example mentions several irrelevant phrases like '*lot of fun*' while KG-Story incorrectly mentions '*gave another speech*' in the last sentence. On contrary, our model's stories are more detailed and less generic such as the phrase, '*ready to go on his mission*' and '*sights and sounds of the enemy*', thus highlighting the effectiveness of using captions and social commonsense to capture events depicted and implied by the images. By not solely relying on visual features and using literal descriptions and commonsense to construct storylines as input, our stories are also consequently more coherent and natural-sounding. Taking the last sentence from AREL in the second story as an example, '*This is the view from the top of the mountain*' sounds abrupt and is unrelated to the previous generated sentences. Conversely, our story is capable of capturing the changes between images while maintaining a strong focus on the topic of '*wine tasting*'.

5.5 Human Evaluation: Pairwise Comparison

Table 2 reports the results of the pairwise comparison between SRL-pmi with AREL, ReCo-RL and PR-VIST. The last column ('Agree') represents results from the Fleiss' kappa test used to assess inter-rater consistency (Fleiss, 1971). Agreement scores in the range [0.21, 0.40], [0.41, 0.60] and [0.61, 0.80] means fair, moderate and strong agreement between multiple annotators respectively.

When analysing all stories ('All Stories' sub-table), our generated stories evidently outperform the baselines by a large margin. All percentages in the first column are over 63%, indicating that the majority of annotators selected our story to be better across all criteria. Moreover, when comparing the 'Overall' criteria which asked evaluators to choose the better story, over 78% of the responses reported our stories to be better with the Fleiss' kappa test result showing a moderate to strong level of agreement between annotators. We believe the higher votes for the visual grounding criteria for our model is due to our method incorporating relevant social-interaction commonsense. Additionally, our constructed storyline is able to reflect the causal events implied by the image stream, resulting in improved story coherence and less repetition.

5.6 Human Evaluation: Story Categories

We analyse the human evaluation results by categorising the stories into 'event-based' and 'object-based'. Event-based stories refer to image streams that focus on people performing actions and there is a clear transition of events between images. Object-based consists of images that mostly picture landscapes and objects. Such instances have no clear event in the image, and thus require more imagination when creating the story. An example of an event-based story is the top sequence in Figure 3 where we can clearly see a man taking a photo and a girl running and sliding across the sand. Conversely, the second example is object-based as a majority of the images depict scenery and buildings. It is harder to generate a story from this input as the first 4 images are extremely similar while the last image is totally different.

Observing the last two sub-tables of Table 2, the first baseline AREL shows lower percentages and ties for object versus event-based stories. As AREL purely relies on generating stories from the visual features, it fails to create coherent output particularly when consecutive images are similar. We qualitatively analyse this in Figure 3 where AREL's story for the object-based example contains more monotonous sentences ('*This is a picture of a city*) and obvious repetition between consecutive sentences. On contrary, our model can generate a more

	All Stories				Event-based				Object-based			
Criteria	**Ours**	**AREL**	**Tie**	**Agree**	**Ours**	**AREL**	**Tie**	**Agree**	**Ours**	**AREL**	**Tie**	**Agree**
Visual Grounding	88.0%	6.6%	5.4%	0.64	86.5%	8.5%	5%	0.60	96.9%	2.5%	0.6%	0.71
Coherence	90.0%	4.8%	5.2%	0.70	88.2%	5.3%	6.5%	0.66	93.8%	3.7%	2.5%	0.82
Non-Redundancy	83.6%	3.0%	13.4%	0.56	82.4%	3.2%	14.4%	0.54	86.3%	2.4%	11.3%	0.60
Overall	93.4%	4.4%	2.2%	0.78	91.8%	5.3%	2.9%	0.74	96.9%	2.5%	0.6%	0.88
Criteria	**Ours**	**ReCo-RL**	**Tie**	**Agree**	**Ours**	**ReCo-RL**	**Tie**	**Agree**	**Ours**	**ReCo-RL**	**Tie**	**Agree**
Visual Grounding	82.2%	10.0%	7.8%	0.49	82.3%	10.6%	7.1%	0.49	81.9%	8.7%	9.4%	0.48
Coherence	93.4%	4.2%	2.4%	0.78	94.7%	3.8%	1.5%	0.81	90.6%	5%	4.4%	0.70
Non-Redundancy	71.6%	11.0%	17.4%	0.30	72.3%	12.4%	15.3%	0.31	70.0%	8.1%	21.9%	0.28
Overall	92.2%	5.0%	2.8%	0.75	93.8%	4.1%	2.1%	0.79	88.8%	6.8%	4.4%	0.66
Criteria	**Ours**	**PR-VIST**	**Tie**	**Agree**	**Ours**	**PR-VIST**	**Tie**	**Agree**	**Ours**	**PR-VIST**	**Tie**	**Agree**
Visual Grounding	78.8 %	17.2 %	4.0 %	0.52	79.1%	16.2%	4.7%	0.52	78.1%	19.4%	2.5%	0.52
Coherence	77.8%	18.4%	3.8 %	0.44	79.4%	17.9%	2.7%	0.47	74.4%	19.3%	6.3%	0.35
Non-Redundancy	63.0%	24.6%	12.4%	0.28	64.4%	22.4%	13.2%	0.29	60.0%	29.4%	10.6%	0.23
Overall	78.0%	16.4%	5.6%	0.46	78.5%	16.2%	5.3%	0.48	76.9%	16.8%	6.3%	0.43

Table 2: Pairwise comparison between SRL-pmi with AREL, ReCo-RL and PR-VIST across the visual grounding, coherence, and non-redundancy criteria for all stories (500 instances) and when separated into event-based (340 instances) and object-based (160 instances) story categories. The 'Agree' column shows the Fleiss' Kappa results.

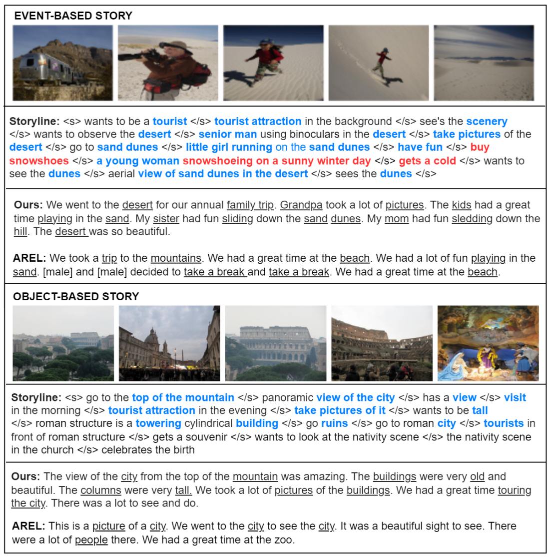


Figure 3: AREL vs. SRL-pmi for an event and object-based story. Blue words indicate concepts implicitly or explicitly used in the generated story while red represents irrelevant concepts. Underlined words in the story represent concepts relevant to the image stream.

visually grounded and coherent story by utilising the storyline. While this example shows several useful concepts in the storyline that are not used in the generated story ('*nativity scene*', '*roman structure*'), concepts such as '*tall*', '*take picture*', and '*tourists*' (highlighted in blue) did help in producing phrases related to these concepts, resulting in a story containing more interesting, diverse and relevant words. Furthermore, while there are error cases where the storyline contains irrelevant information such as the red words in the event-based example, this information was not included in the generated output. This is perhaps due to the advantages of the encoder-decoder cross-attentional mechanism of BART which allows the model to learn to select the more useful parts of the storyline.

Examining ReCo-RL, only the grounding and non-redundancy aspect received lower votes for object versus event-based instances. Compared to AREL, its better performance may be due to its framework incorporating RL rewards to directly align the outputs more to a human story in terms of the 3 criteria. PR-VIST however which first builds a storyline like ours, outperforms AREL and ReCo-RL and further, even yields slightly more votes for object-based stories compared to its proportion of votes received for event-based stories, thus highlighting the effectiveness of storyline and content planning. Despite PR-VIST's improvements, our approach and storyline construction method is evidently superior and substantially outperforms PR-VIST in all aspects across the 2 categories.

6 Conclusion

In this paper, we presented SCO-VIST, a multi-stage novel framework for visual storytelling that utilises social-interaction knowledge for enhancing commonsense reasoning in stories. We design a heterogeneous story graph with causal ordering that connects captions and commonsense extracted from external sources and employ shortest path algorithms to find the optimal storyline for story generation. Extensive experiments on the benchmark dataset, analysis of automatic metrics and human evaluations demonstrate that SCO-VIST outperforms existing baselines and is capable of generating diverse stories that are highly coherent with strong visual grounding.

Limitations

Benchmark Scope and Annotation Due to the lack of a high-quality visual storytelling dataset, most recent studies on visual story generation use only one publicly available dataset, VIST. The dataset size is large enough but the dataset used in most visual storytelling research publications, including this study, was limited in scope. The VIST consists of images from Flickr, which is an image/video-based social media platform and includes mostly personal images that captures people's daily lives or events. In addition, each Flickr album has 5 human written stories where each story is usually comprised of one sentence per image. Those human annotators are not the Flickr album owner and hence the gold standard annotations by annotators may not be perfectly matched with the intention of the original Flickr album. Future work should investigate how to mitigate this issue by establishing a new visual storytelling dataset via adopting the image album descriptions from the original authors, and providing better instructions for human annotators that map generated stories to objects/relations of images.

Adaptability to Low-Resource Languages Moreover, our model pipeline requires a pre-trained image captioning model in the first stage, which may not be available for low-resource languages that have relatively less data available for training natural language processing systems. The metrics used for evaluation are also only capable of judging English-written language. Nevertheless, our pipeline can be reproduced and future study should consider re-running experiments on other languages once models and data become available.

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.

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. *arXiv preprint arXiv:2102.02963*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Zi-Yuan Chen, Chih-Hung Chang, Yi-Pei Chen, Jijnasa Nayak, and Lun-Wei Ku. 2019. Uhop: An unrestricted-hop relation extraction framework for knowledge-based question answering. In *Proceedings of NAACL-HLT*, pages 345–356.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Robert W Floyd. 1962. Algorithm 97: shortest path. *Communications of the ACM*, 5(6):345.

Diana Gonzalez-Rico and Gibran Fuentes Pineda. 2018. Contextualize, show and tell: A neural visual storyteller. *CoRR*, abs/1806.00738.

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.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Xudong Hong, Rakshith Shetty, Asad Sayeed, Khushboo Mehra, Vera Demberg, and Bernt Schiele. 2020. Diverse and relevant visual storytelling with scene graph embeddings. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 420–430.

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.

Chi-yang Hsu, Yun-Wei Chu, Ting-Hao Huang, and Lun-Wei Ku. 2021a. Plot and rework: Modeling storylines for visual storytelling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4443–4453.

Chi-Yang Hsu, Yun-Wei Chu, Tsai-Lun Yang, Ting-Hao Huang, and Lun-Wei Ku. 2021b. Stretch-vst: Getting flexible with visual stories. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 356–362.

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.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: on symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6384–6392.

Yunjae Jung, Dahun Kim, Sanghyun Woo, Kyungsu Kim, Sungjin Kim, and In So Kweon. 2020. Hide-and-tell: learning to bridge photo streams for visual storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11213–11220.

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. *CoRR*, abs/1805.10973.

DP Kingma and LJ Ba. 2015. 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.

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.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*.

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

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Marko Smilevski, Ilija Lalkovski, and Gjorgji Madjarov. 2018. Stories for images-in-sequence by using visual and narrative components. In *International Conference on Telecommunications*, pages 148–159. Springer.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Aynaz Taheri and Tanya Berger-Wolf. 2019. Predictive temporal embedding of dynamic graphs. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 57–64.

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.

Eileen Wang, Caren Han, and Josiah Poon. 2022. RoViST: Learning robust metrics for visual storytelling. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2691–2702, Seattle, United States. Association for Computational Linguistics.

Ruize Wang, Zhongyu Wei, Piji Li, Qi Zhang, and Xuanjing Huang. 2020. Storytelling from an image stream using scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9185–9192.

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.

Chunpu Xu, Min Yang, Chengming Li, Ying Shen, Xiang Ao, and Ruifeng Xu. 2021. Imagine, reason and write: Visual storytelling with graph knowledge and relational reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3022–3029.

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*, pages 5356–5362.

Youngjae Yu, Jiwan Chung, Heeseung Yun, Jongseok Kim, and Gunhee Kim. 2021. Transitional adaptation of pretrained models for visual storytelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12658–12668.

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)*.

Model	RoViST-VG	RoViST-C	RoViST-NR	RoViST	UNION	SPICE	Story Len.
CLIP-SRL-pmi	69.6	71.1	**91.1**	77.3	68.5	10.8	49.5
BLIP-SRL-pmi	70.8	69.8	90.5	77.0	72.1	11.3	51.2
VIST-SRL-pmi	**72.0**	**74.1**	90.6	**78.9**	**82.5**	**12.5**	**57.6**

Table 3: RoViST, UNION, and SPICE scores recorded when using different captioning models. All models are implemented using the SRL-pmi SCO-VIST variant.

A Qualitative Analysis

Figure 4 presents two example stories for our SRL-pmi model versus 6 visual storytelling baseline models: AREL, GLACNet, KG-Story, ReCo-RL, PR-VIST and TAPM.

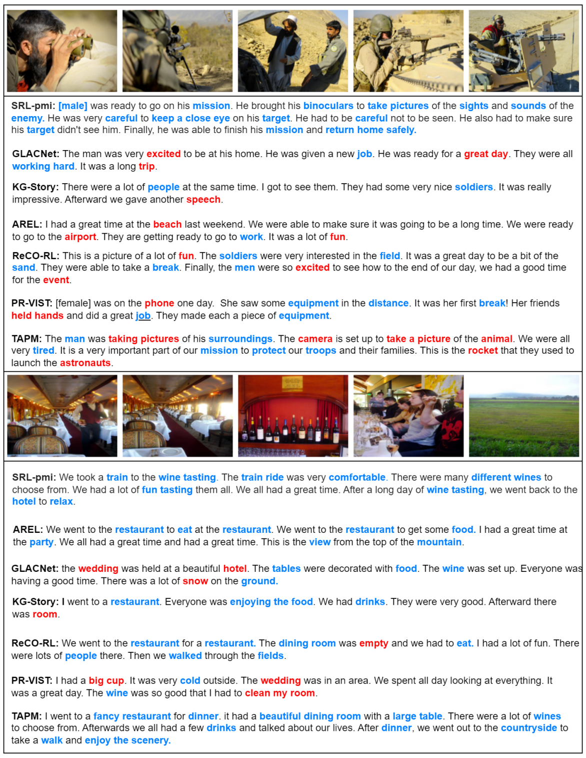


Figure 4: Generated stories for our SRL-pmi model versus the 6 baselines models. Blue/red words represent concepts relevant/irrelevant to the image sequence.

B Caption Ablation Study

We conduct a preliminary ablation study to examine the performance of the stories when using different captioning models. For the experiments in the main paper, we utilised ClipCap (Mokady et al., 2021) to generate the image captions. For this experiment, we additionally consider the BLIP captioning model (Li et al., 2022) which outperforms ClipCap on COCO captions (Chen et al., 2015). We also consider using the human-written captions which are provided as part of the VIST dataset. Note that for this experiment, we implement the SRL-pmi SCO-VIST variant for all models. Moreover, all models were trained on a substantially smaller dataset size (26939 instances for training, 3354 for validation and 3385 for testing) compared to the dataset used to retrieve the results in the main paper as ground-truth descriptions from VIST were only available for approximately half of the data. The CLIPCap and BLIP captions achieve a BLEU-1 score of 13.7 and 17.5 respectively when evaluated against the ground-truth VIST captions.

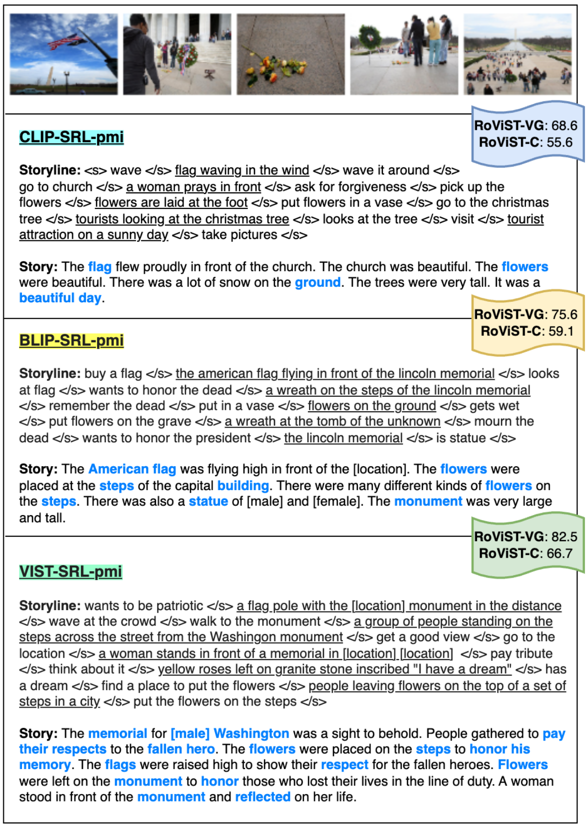


Figure 5: An example of a storyline and matching story generated using the SRL-pmi approach with different pre-trained image captioning models. Underlined words in the storyline are the image captions and blue words are visually relevant concepts to the image sequence.

The RoViST, UNION and SPICE scores using each captioning method is displayed in Table 3. Firstly, it is evident that using human-written captions in the story graph creation process results in a higher RoViST-VG, RoViST-C and RoViST score overall as observed by VIST-SRL-pmi. UNION and SPICE were also considerably higher, suggesting better captions lead to better stories and SCO-VIST's outputs can be perhaps further improved with a stronger pre-trained captioning model. However for this study, we did find that using the BLIP captions produces a similar overall RoViST score. Neverthless, BLIP-SRL-pmi did yield greater RoViST-VG, UNION and SPICE compared to CLIP-SRL-pmi. The higher RoViST-VG score could imply that the caption quality influences the visual grounding aspect the most. This is reasonable as an incorrect caption could cause irrelevant concepts to be generated in the storyline, which can directly negatively impact the visual grounding score (RoViST-VG).

To highlight a specific example, we further conduct a qualitative analysis in Figure 5 to assess how the caption quality can affect the generated storylines and stories. Taking CLIP-SRL-pmi for instance, the incorrect captions *'tourists looking at the christmas tree'* and *'a woman prays in front'* results in irrelevant concepts mentioned in the story such as *'church'* and *'snow'*. Conversely, using more detailed and accurate captions as depicted in BLIP-SRL-pmi and VIST-SRL-pmi clearly results in better storylines which in turn, translates to more visually grounding and detailed stories.

C Implementation Details

To generate the image captions for Stage 1, we use a pre-trained image captioning model called ClipCap (Mokady et al., 2021). For commonsense generation, we use the 'comet_atomic2020_bart' implementation of Comet-ATOMIC2020 (Hwang et al., 2021). Sentence embeddings of the nodes are then obtained with a Sentence Transformer using the 'all-mpnet-base-v2' model (Reimers and Gurevych, 2019) which outputs embeddings of size 768. Since some generated commonsense were found to be duplicated or similar, these similar or identical commonsense were filtered out based on if the sentence embedding cosine similarity score between the two phrases exceeded a threshold of 0.50 for each of the `BEFORE` and `AFTER` events produced by each caption. In Stage 2, the temporal GCN used to learn the node embeddings consisted of 1 layer and the chosen output dimension of the embeddings was 768. Furthermore, the Transformer model used to take in the 5 caption nodes to decode the story utilised the 'bart-base' configuration of the BART Transformer model (Lewis et al., 2020). This model was trained with a learning rate of 0.00001. In Stage 3, the story decoder using the storyline as input employed the 'bart-large' configuration and was trained with a learning rate of 0.00002. For all BART models, we initialise with the pretrained weights and finetune them on our VST task. All experiments also used a batch size of 8, weight decay of 0.00001, learning rate decay of 0.95 scheduled to decrease after every epoch and the Adam optimizer (Kingma and Ba, 2015). Early stopping was further employed to stop training after 3 consecutive epochs of no improvement on the validation set. At inference, we decode the story with nucleus sampling using the recommended values of p = 0.9 and temperature = 0.9 (Holtzman et al., 2019). All training of models was conducted using a Nvidia Tesla v100 16GB GPU which took approximately 15 hours to train.

D Story Generation Results for N-gram Metrics

ROUGE-L, METEOR and CIDEr results for the 6 state-of-the-art baselines versus our proposed SRL-pmi and TGCN-SRL-cosine variant of SCO-VIST.

Model	ROUGE-L	METEOR	CIDEr
AREL	**29.9**	35.2	9.1
GLACNet	27.2	33.5	4.4
KG-Story	25.2	31.5	3.8
ReCo-RL	29.3	**35.9**	**11.9**
PR-VIST	26.1	31.4	7.6
TAPM	21.7	27.0	4.5
SRL-pmi	22.1	27.5	5.9

Table 4: Classic n-gram metrics for our top model, SRL-pmi vs. the 6 baselines.

E Baseline Models

A brief description of each baseline model is as follows:

1. **AREL** (Wang et al., 2018) adopts an inverse reinforcement learning (RL) approach trained in an adversarial manner with a CNN-based reward model.

2. **GLACNet** (Kim et al., 2018) is another end-to-end model that combines both local and global attention mechanisms on the image features.
3. **KG-Story** (Hsu et al., 2020) attempts to enrich stories by leveraging external knowledge bases like Visual Genome (Krishna et al., 2017) and OpenIE (Pal et al., 2016). For story generation, a Transformer model is used.
4. **ReCo-RL** (Hu et al., 2020) proposes another RL method with composite rewards designed to target the *relevance*, *coherence* and *expressiveness* criteria of VST.
5. **PR-VIST** (Hsu et al., 2021a) is a newer model where similar to ours, attempts to link nouns together with verb relations extracted from Visual Genome and VIST to form a story graph. The optimal storyline is then extracted using UHop (Chen et al., 2019).
6. **TAPM** (Yu et al., 2021) introduces an auxiliary training task to harmonise the language generator and visual encoder before optimising the target objective. The task proposes to minimise the 'sequential coherence loss' which aims to enforce text representations to predict surrounding visual representations within a closed neighbourhood.

Note that this is not the end of the Appendix section. The following page includes Appendix F, G, and H.

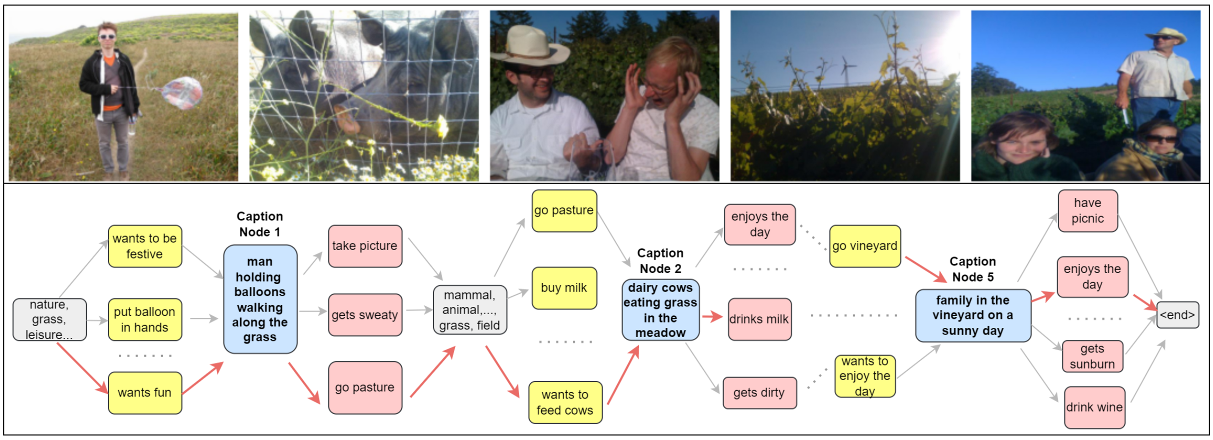


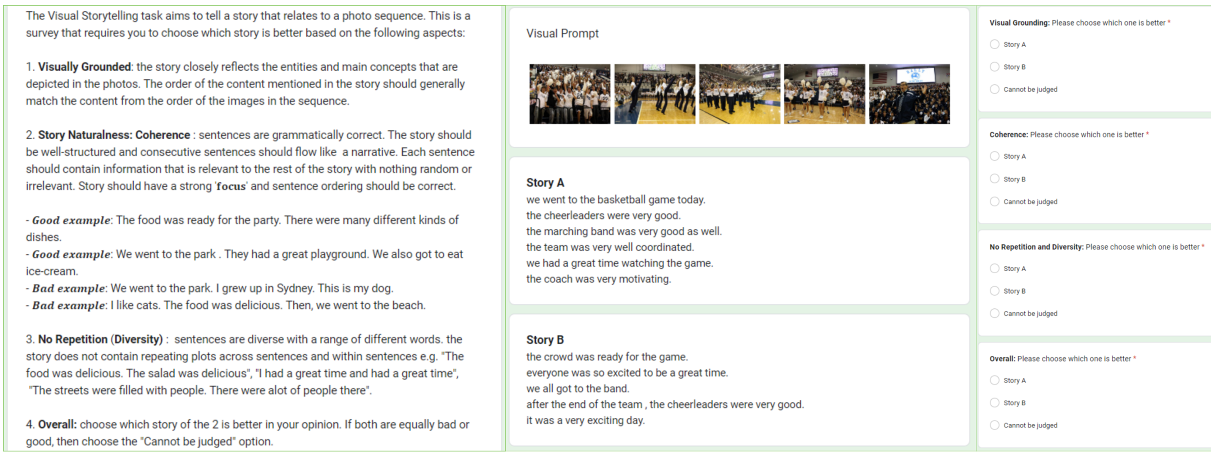
Figure 6: Final story graph generated from Stage 3 with red arrows indicating the optimal extracted storyline.

F Story Graph

Figure 6 shows the final directed story graph generated from Stage 2 with the additional dummy end node added in Stage 3. Grey and blue nodes are theme and caption nodes respectively. Yellow nodes are commonsense nodes from the BEFORE events group generated by the xNeed and xIntent relation while red nodes are the AFTER events commonsense nodes from the xWant and xEffect relation. Due to limited space, only the nodes corresponding to image 1, 2 and 5 are visualised and dotted lines are used to indicate nodes in the graph that are not displayed. The red highlighted arrows show the shortest path found by Floyd Warshall's algorithm where the caption nodes and commonsense nodes are taken in order to use as the storyline. For simplicity, edge weights are also not shown.

G Human Evaluation Survey

Figure 7 shows the survey instructions used in the human evaluation study and the format of the survey questions. The 15 participants recruited were volunteers from a variety of age groups (20-60 years old), occupation and gender (8 female, 7 male). All participants were proficient in English with at least a university education level. Note that we modified and used similar instructions from the study proposed in Wang et al. (2022). It is also emphasised that annotators do not know which model generated which story as for each example, we randomly swap the order of the baseline story and SCO-VIST's story to be presented as Story A and Story B.



The Visual Storytelling task aims to tell a story that relates to a photo sequence. This is a survey that requires you to choose which story is better based on the following aspects:

1. **Visually Grounded**: the story closely reflects the entities and main concepts that are depicted in the photos. The order of the content mentioned in the story should generally match the content from the order of the images in the sequence.

2. **Story Naturalness: Coherence** : sentences are grammatically correct. The story should be well-structured and consecutive sentences should flow like a narrative. Each sentence should contain information that is relevant to the rest of the story with nothing random or irrelevant. Story should have a strong '**focus**' and sentence ordering should be correct.

- ***Good example***: The food was ready for the party. There were many different kinds of dishes.
- ***Good example***: We went to the park . They had a great playground. We also got to eat ice-cream.
- ***Bad example***: We went to the park. I grew up in Sydney. This is my dog.
- ***Bad example***: I like cats. The food was delicious. Then, we went to the beach.

3. **No Repetition** (**Diversity**) : sentences are diverse with a range of different words. the story does not contain repeating plots across sentences and within sentences e.g. "The food was delicious. The salad was delicious", "I had a great time and had a great time", "The streets were filled with people. There were alot of people there".

4. **Overall:** choose which story of the 2 is better in your opinion. If both are equally bad or good, then choose the "Cannot be judged" option.

Visual Prompt

Story A
we went to the basketball game today.
the cheerleaders were very good.
the marching band was very good as well.
the team was very well coordinated.
we had a great time watching the game.
the coach was very motivating.

Story B
the crowd was ready for the game.
everyone was so excited to be a great time.
we all got to the band.
after the end of the team , the cheerleaders were very good.
it was a very exciting day.

Visual Grounding: Please choose which one is better *
Story A
Story B
Cannot be judged

Coherence: Please choose which one is better *
Story A
Story B
Cannot be judged

No Repetition and Diversity: Please choose which one is better *
Story A
Story B
Cannot be judged

Overall: Please choose which one is better *
Story A
Story B
Cannot be judged

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

H Event-based versus Object-based Stories

Figure 8 contains examples of more generated outputs from our SRL-pmi model versus AREL for event-based and object-based stories as described in Section 5.6 of the paper. Here, blue words in the storyline indicate concepts implicitly or explicitly used in the generated story while red words represent irrelevant or not useful concepts in the storyline. The underlined words in the generated story represent concepts relevant to the image stream.

EVENT-BASED STORY

Storyline: <s> be a pilot </s> helicopter flying in the sky </s> flying in air </s> learn how to fly </s> pilots in the cockpit of aircraft model </s> fly the model </s> wants to go on a mission </s> soldiers prepare to board a helicopter </s> gets ready to fly </s> go to military base </s> soldiers prepare to take part in a training exercise </s> to go to war </s> wants to go to war </s> soldiers take part in a military exercise </s> gets fatigued </s>

Ours: The helicopter was ready to take off. The pilots were in the cockpit. The soldiers were getting ready to board the helicopter. They were all lined up and ready to go. The helicopter took off. It was a long flight.

AREL: The men are getting ready for the show. The soldiers were getting ready to go. We were able to get a closer view of the fire. They were all very proud. They were all happy to be there.

EVENT-BASED STORY

Storyline: <s> find a bar </s> person and me at the bar </s> have a drink </s> have a gallery </s> person with some of her sculptures </s> person with some art </s> buy a bike </s> bicyclists ride along the boardwalk </s> enjoy the weather </s> wants to capture the beauty of nature </s> aerial 4k drone view of waves crashing on the rocky shoreline during a stormy sunset </s> see's the sunset </s> walk up the hill </s> person at the top of the hill </s> go to the top </s>

Ours: [female] and [female] were having a great time on their vacation. They saw a lot of interesting things. They rode their bikes along the beach. The waves were crashing on the shore. They had a great view of the ocean. They took a picture together at the end of the day.

AREL: We had a great time at the beach. We saw many interesting statues. We rode on the beach. The view from the top of the beach was beautiful. We had a great time at the beach .

OBJECT-BASED STORY

Storyline: <s> wants to enjoy the outdoors </s> a picnic basket full of food </s> eat food </s> buy corn on the cob </s> corn on the cob at the farmer's market </s> eat the corn </s> eat </s> foods and food product on a picnic blanket </s> have a picnic </s> turn on grill </s> cooking food on the grill </s> cook the food </s> read the text </s> a plate with the text written on it </s> read the text </s>

Ours: We went to the farmer's market today. They had a lot of different kinds of corn. They also had lots of different types of hot dogs. They even had hot dogs on the grill. We had a great time. We ended the day with a delicious meal.

AREL: We had a lot of food for the party. I made a lot of food to eat. I got to meet some of my friends. We had a lot of food to eat. I had a great time.

OBJECT-BASED STORY

Storyline: <s> wants to show his age </s> portrait of an old man in a straw hat </s> painting of an old man </s> have a garden </s> view of the garden </s> look at the flowers </s> have a garden </s> a view of the gardens </s> look at the flowers </s> wants to enjoy nature </s> tourist attraction which includes a garden and a pond or waterhole as well as a small group of people </s> have a picnic </s> visit </s> tourist attraction in the rain </s> gets wet from rain </s>

Ours: [male] was excited to visit the garden. The garden was very well kept. There were many different types of flowers. There was a pond in the back of the garden as well. [male] had a great time there. It was a great place to visit.

AREL: I went to the park today. The garden was very beautiful. The trees were very beautiful. The trees were very beautiful. The garden was a beautiful place to visit.

Figure 8: AREL versus our SRL-pmi model for event-based and object-based stories.