Attractive Storyteller: Stylized Visual Storytelling with Unpaired Text

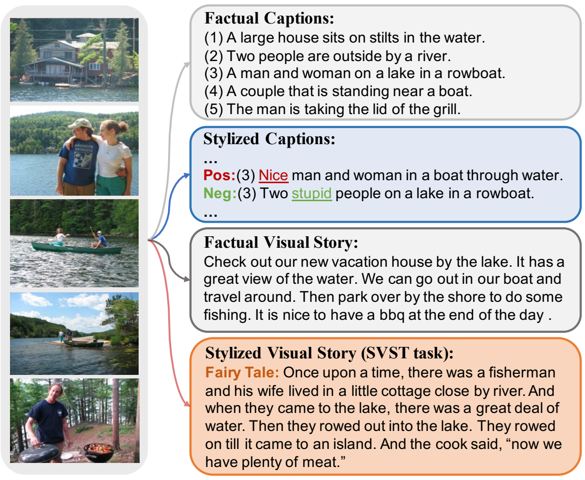
Dingyi Yang, Qin Jin*
School of Information, Renmin University of China
{yangdingyi, qjin}@ruc.edu.cn

Abstract

Most research on stylized image captioning aims to generate style-specific captions using unpaired text, and has achieved impressive performance for simple styles like positive and negative. However, unlike previous single-sentence captions whose style is mostly embodied in distinctive words or phrases, real-world styles are likely to be implied at the syntactic and discourse levels. In this work, we introduce a new task of *Stylized Visual Storytelling (SVST)*, which aims to describe a photo stream with stylized stories that are more expressive and attractive. We propose a multi-tasking memory-augmented framework called StyleVSG, which is jointly trained on factual visual storytelling data and unpaired style corpus, achieving a trade-off between style accuracy and visual relevance. Particularly for unpaired stylized text, StyleVSG learns to reconstruct the stylistic story from roughly parallel visual inputs mined with the CLIP[1] model, avoiding problems caused by random mapping in previous methods. Furthermore, a memory module is designed to preserve the consistency and coherence of generated stories. Experiments show that our method can generate attractive and coherent stories with different styles, such as fairy tale, romance, and humor. The overall performance of our proposed StyleVSG surpasses state-of-the-art methods on both automatic and human evaluation metrics [2].

1 Introduction

Over the years, Image Captioning has made remarkable progress (Xu et al., 2015; Guo et al., 2020; Hu et al., 2022a). Factual image captioning focuses on generating objective and neutral descriptions of image content without considering style characteristics. However, when describing images, people are likely to include specific styles, which can make captions more attractive and emotionally appropriate. Therefore, Stylized Image Captioning (Mathews et al., 2016; Gan et al., 2017) has recently attracted increasing research attention.



Factual Captions:
(1) A large house sits on stilts in the water.
(2) Two people are outside by a river.
(3) A man and woman on a lake in a rowboat.
(4) A couple that is standing near a boat.
(5) The man is taking the lid of the grill.

Stylized Captions:
...
Pos:(3) Nice man and woman in a boat through water.
Neg:(3) Two stupid people on a lake in a rowboat.
...

Factual Visual Story:
Check out our new vacation house by the lake. It has a great view of the water. We can go out in our boat and travel around. Then park over by the shore to do some fishing. It is nice to have a bbq at the end of the day .

Stylized Visual Story (SVST task):
Fairy Tale: Once upon a time, there was a fisherman and his wife lived in a little cottage close by river. And when they came to the lake, there was a great deal of water. Then they rowed out into the lake. They rowed on till it came to an island. And the cook said, "now we have plenty of meat."

Figure 1: Given an ordered image sequence from a photo album, SVST aims to generate an attractive story with specific styles. This story better reflects linguistic styles at both syntactic and discourse levels than single-sentence stylized captions.

Existing works on *SentiCap* (Mathews et al., 2016) and *FlickrStyle10K* (Gan et al., 2017) can generate stylized descriptions comprised of factual caption and additional stylistic words or phrases, such as the positive and negative captions shown in Figure 1, with the stylistic words highlighted. However, real-world styles are likely to be entangled throughout the text, making it difficult to separate style from fact. In addition, long text is better to reflect linguistic preference, not only at the syntactic level, but also at the discourse level. To step forward from the previous stylized captioning task, we propose a new task called ***Stylized Visual Storytelling (SVST)***, which requires models to generate a stylized story to describe a sequence of images (5 images in our following experiments). As the ex-

* Corresponding Author.

[1] https://openai.com/blog/clip/

[2] The code will be released at https://github.com/DingyiYang/StyleVSG.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 11053–11066
July 9-14, 2023 ©2023 Association for Computational Linguistics

ample illustrated in Figure 1, stylized visual stories are much more attractive than stylized captions and factual visual stories.

The lack of large-scale parallel data is the main challenge for stylized image captioning (Li et al., 2021; Tan et al., 2022), as well as for the new SVST task. Current unsupervised methods either split stylized captions into style- and content-related tokens (Li et al., 2021; Zhao et al., 2020), or disentangle style and content implicitly (Tan et al., 2022; Chen et al., 2018; Mathews et al., 2018; Gan et al., 2017). But for many human-like styles, it is difficult to strip out a clear "style-related part", such as the stylized text "Once upon a time, there was a fisherman and his wife lived in a little cottage close by river". In such cases, token-level separation is not effective; latent-level split might also fail due to a lack of supervision (Liu et al., 2022), leading to incorrect alignment between visual content and stylized descriptions.

In this paper, we propose a new framework called Stylized Visual Story Generator (**StyleVSG**), to generate attractive visual stories with target styles. StyleVSG is trained on the factual visual storytelling task using a paired dataset VIST [3], and the stylized story reconstruction task using an unpaired style corpus. Instead of applying latent-level split methods, StyleVSG aligns roughly parallel visual pairs for unpaired stylized text, avoiding problems caused by random mapping (Liu et al., 2022). Particularly, we leverage the large-scale language-image pre-trained model CLIP (Radford et al., 2021) to mine the most relevant visual content as input. In our story generator, we apply style-dependent layer normalization and style-dependent cross-attention to constrain specific styles; we further design a memory unit to model the relations among successive textual sequences and images, in order to address the challenge of generating coherent and fluent stories.

We carry out experiments to validate our proposed method, using both objective metrics and human evaluations. We consider multiple quality aspects of the generated stylized stories, including visual relevance, style appropriateness, and overall coherence. StyleVSG outperforms previous methods in terms of overall performance on both objective and human evaluations.

In summary, our contributions are as follows:

- To the best of our knowledge, it is the first work to generate stylized stories for image sequences without paired (images, stylized story) data.
- We propose the StyleVSG framework to train the model jointly by leveraging both paired factual visual-story data and our unpaired stylized story data.
- Both objective metrics and human evaluations verify that our proposed StyleVSG can generate coherent stylized stories for an image sequence, achieving better overall performance than other strong baselines.

2 Related Works

2.1 Stylized Image Captioning

Stylized Image Captioning (Mathews et al., 2016) aims to describe an image with target styles, in order to make image captions more expressive and attractive. As it is laborious to construct large-scale parallel data, most existing works explore unsupervised methods. Some works explicitly divide stylized sentences into semantic parts and style-related phrases. Mathews et al. (2018) propose to generate visually relevant semantic terms, which are then translated into stylistic captions. Zhao et al. (2020) propose a memory module to extract style knowledge within content-related and style-related phrases. Li et al. (2021) extend the existing dataset using factual captions and possible stylized phrases. These methods work well for simple styles like positive and negative, but they fail to work when the target style is not implied in distinctive style-related tokens. Other approaches attempt to incorporate style information when generating captions from a shared intermediate image-text space. Gan et al. (2017) and Chen et al. (2018) propose to learn two groups of matrices to capture factual and stylized knowledge. Guo et al. (2019) propose an adversarial learning framework to enhance overall performance. Chen et al. (2019) apply style-dependent layer normalization to control different styles. Tan et al. (2022) detach text style representations in stylized textual space, and then attach them with visual content representations. Lovenia et al. (2022) propose a mapping network to align the visual and semantic spaces of large-scale pre-trained models, and apply style-related adapters to guide the generation of stylized stories that describe an image. These implicit methods are prone to failure due to the lack of supervision, leading to incorrect alignment through random mapping (Liu et al., 2022).

[3] http://visionandlanguage.net/VIST/

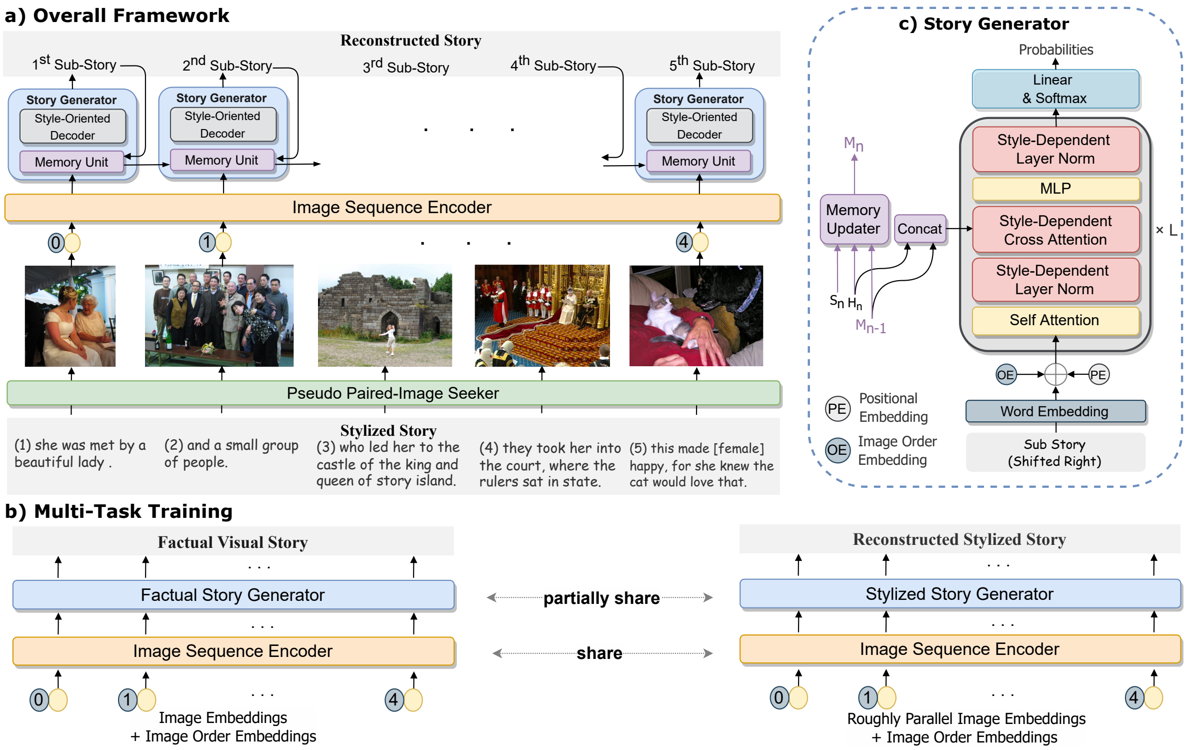

a) Overall Framework
Reconstructed Story
1st Sub-Story
2nd Sub-Story
3rd Sub-Story
4th Sub-Story
5th Sub-Story
Story Generator
Style-Oriented Decoder
Memory Unit
Image Sequence Encoder
0
1
4
Pseudo Paired-Image Seeker
Stylized Story
(1) she was met by a beautiful lady .
(2) and a small group of people.
(3) who led her to the castle of the king and queen of story island.
(4) they took her into the court, where the rulers sat in state.
(5) this made [female] happy, for she knew the cat would love that.
c) Story Generator
Probabilities
Linear & Softmax
Style-Dependent Layer Norm
MLP
Style-Dependent Cross Attention
Style-Dependent Layer Norm
Self Attention
× L
Mn
Memory Updater
Concat
Sn Hn
Mn-1
OE
PE
Word Embedding
Sub Story (Shifted Right)
PE Positional Embedding
OE Image Order Embedding
b) Multi-Task Training
Factual Visual Story
Factual Story Generator
Image Sequence Encoder
Image Embeddings + Image Order Embeddings
partially share
share
Reconstructed Stylized Story
Stylized Story Generator
Image Sequence Encoder
Roughly Parallel Image Embeddings + Image Order Embeddings

Figure 2: Overview of our proposed StyleVSG, which is trained on paired factual VIST data and unpaired stylized story corpus. (a) We illustrate the stylized story reconstruction task (Section 3.1.2) to show the overall framework. The factual visual storytelling task (Section 3.1.1) does not include the pseudo paired-image seeker, instead taking a real image sequence to generate a factual story; (b) During training, both tasks share the same set of parameters, apart from style-dependent components (red) in the Story Generator; (c) The Story Generator is constructed of a Memory Unit (Section 3.2.1) which records history information, and a Style-Oriented Decoder (Section 3.2.2) which constrains different styles (factual/target style).

2.2 Visual Storytelling

Visual storytelling (VST) (Huang et al., 2016) aims to generate a human-like story that describes an ordered image sequence from a photo album. VST is a challenging task because models need to understand not only the semantic meaning of each image, but also the relations among images, and generate fluent paragraphs and imaginary concepts for storytelling. Huang et al. (2016) release a large-scale benchmark dataset called VIST, which has inspired many following works in this area. Some of them (Yang et al., 2019; Hsu et al., 2020, 2021; Xu et al., 2021) attempt to incorporate extra commonsense knowledge to generate more interesting stories. Hsu et al. (2019) collect human edits of machine-generated visual stories, helping large visual storytelling models generate more huma-like stories. Other variants of VST include generating visual stories that incorporate emotional categories (Li et al., 2019), personalities (Prabhumoye et al., 2019), and specific topics (Zhang et al., 2022). However, there is no attempt to generate more fascinating stories with real-world writing styles. In this paper, we propose to generate stylized visual stories, which faces challenges in both storytelling and text style injection.

3 Method

Given an image sequence $\boldsymbol{I} = \{I_n\}_{n=1}^{5}$, Stylized Visual Storytelling (SVST) aims to generate a story $\boldsymbol{s} = \{s_n\}_{n=1}^{5}$ in a specific style, where each sub-story $s_n = \{w_1, \ldots, w_{K_n}\}$ consists of K_n words in the word vocabulary. Please note that our task is fully unsupervised because there is no paired stylized data (i.e. in the form of $(\boldsymbol{I}, \boldsymbol{s})$) for training. To constrain the target style and fully leverage the auxiliary paired data of factual stories, we apply a *multi-task training framework* (Figure 2 (b)), which attempts to achieve a trade-off between style accuracy and visual relevance. Specifically, we utilize the VIST dataset $\{(\boldsymbol{I}^{(i)}, \boldsymbol{y}^{(i)})\}_{i=1}^{N}$ which contains pairs of an image sequence $\boldsymbol{I}$ and its factual story $\boldsymbol{y}$; and a stylized corpus $\{\boldsymbol{t}^{(j)}\}_{j=1}^{M}$ that only consists of several stylistic stories $\boldsymbol{t}$.

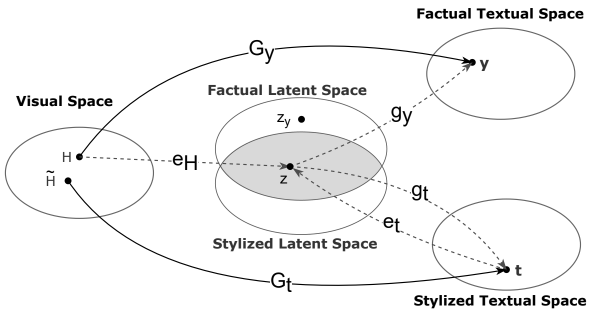


Figure 3: Previous methods assume that there exists a shared latent space $\mathcal{Z}$ (the shaded area in this figure), which can be further mapped into factual/stylized textual space via g_y/g_t.The visual encoder e_H is applied to encode visual content H into the intermediate latent space $\mathcal{Z}$, and the textual encoder e_t is employed to encode stylized text t into $\mathcal{Z}$. Ideally, e_H and e_t could align H and related t via the latent vector z in $\mathcal{Z}$. However, it is hard to learn such an ideal shared space and make accurate alignment (Liu et al., 2022). Instead, our proposed StyleVSG learns G_y and G_t that directly map from visual space to textual space.

The overall architecture of StyleVSG is illustrated in Figure 2 (a), which employs stacks of transformers (Vaswani et al., 2017). We use the pre-trained CLIP (Radford et al., 2021) to extract high-level image features, and sum them with image order embeddings as the model inputs. The *Image Sequence Encoder* E_y/E_t is employed to get contextual visual representation $\boldsymbol{H} = \{H_n\}_{n=1}^{5}$, which are then fed into the *Story Generator* G_y/G_t to generate each sub-story. As illustrated in Figure 2(c), the *Story Generator* consists of: the *Memory Unit* M_y/M_t, which records history information to preserve the coherence and consistency of generated stories; and the *Style-Oriented Decoder* D_y/D_t which constrains different styles.

3.1 Multi-Task Training Framework

The assumption of previous latent-level split methods (Chen et al., 2019; Tan et al., 2022) is that there exists a shared intermediate image-text space,from which factual descriptions can be generated via g_y or stylistic descriptions can be generated via g_t, as shown in Figure 3. In the ideal condition, after training, the visual encoder e_H and textual encoder e_t could align images and related stylized texts in a shared latent space; the decoder g_y and g_t share the ability to describe the same semantic content but with different styles. Therefore, in the inference step, they can firstly encode the image into a latent code z using e_H, and then apply g_t to generate a stylized description.

However, for complicated styles, stylized descriptions might be completely different from factual captions, making it difficult to achieve the desired results (Tan et al., 2022). There are two types of problems as illustrated in Figure 3: (1) The learned stylized and factual latent space are partially overlapping. While making predictions, some instances are encoded into the non-overlapping space, making g_t confused to generate stylized texts with such latent vectors (like z_y). (2) After training, e_H encodes visual content H into latent vector z, while e_t encodes stylized text t into the same z. However, H and t have different semantic meanings. Therefore, g_t will generate stylistic descriptions unrelated to the visual content.

Instead, StyleVSG learns G_y and G_t that directly generate text from shared visual space. This guarantees that similar input hidden states represent similar semantic content. With paired factual data, we could learn the factual story decoder G_y, corresponding to the task of *Supervised Visual Storytelling*. For unpaired style corpus, we consider reconstructing stylized stories with roughly parallel visual information, corresponding to the task of *Unsupervised Stylized Story Reconstruction*.

3.1.1 Supervised Factual Visual Storytelling

With the VIST dataset, StyleVSG learns the image sequence encoder E_y, memory unit M_y (described in 3.2.1), and factual decoder D_y by minimizing:

$$\mathcal{L}_Y = \sum_{n=1}^{5} \sum_{k=1}^{n_K} -log\left(p_{\theta_{E_y},\theta_{M_y},\theta_{D_y}}(w_k^n | w_{1:k-1}^n, y_{1:n-1}, \boldsymbol{I})\right), \tag{1}$$

where θ_{E_y}, θ_{M_y}, and θ_{D_y} are the parameters of E_y, M_y and D_y; y_n is the n-th sub-story and w_k^n is the k-th word in a sub-story.

3.1.2 Unsupervised Stylized Story Reconstruction

Since there is no paired image sequences for stylized stories, we mine roughly parallel visual information using CLIP (Radford et al., 2021), a pre-trained model trained on 400 million image-text pairs. Benefiting from the large-scale training data collected online, it has the power to find the closest visual sample for natural language, even with a specific style. In a sample story $\boldsymbol{t}$ from the stylized corpus, for each sentence t_n, we propose to seek the closest image $\tilde{I}_n$ in the source photo set $\mathcal{V}$ [4], considering both overall and local similar-

[4] $\mathcal{V} = \{I_n^{(i)}\}_{n=1,i=1}^{5,N}$ is the photo set of VIST.

ity. Concretely, we apply spacy[5] to extract noun chunks $\{N_c\}_{c=1}^{C_n}$ in each sentence, forming a sentence set $\boldsymbol{S}_n$ as $\{S_{nc} = \text{“a photo of } N_c\text{”}\}_{c=1}^{C_n}$. The final similarity alignment score is computed by:

$$\begin{aligned} \mathrm{S}(t_n, I) &= \mathrm{Sim}(\mathrm{CLIP}_{\text{text}} t_n, \mathrm{CLIP}_{\text{image}} I) \\ &+ \frac{\sum_{c=1}^{C_n} \mathrm{Sim}(\mathrm{CLIP}_{\text{text}} S_{nc}, \mathrm{CLIP}_{\text{image}} I)}{|\boldsymbol{S}_n|}, \end{aligned} \tag{2}$$

where Sim refers to cosine similarity. An example of a mined image sequence is shown in Figure 2.

Through the above process, we get a roughly parallel image sequence $\tilde{\boldsymbol{I}} = \{\tilde{I}_n\}_{n=1}^{5}$, which are applied to reconstruct the stylized story $\boldsymbol{t}$. The loss function is formulated as:

$$\mathcal{L}_T = \sum_{n=1}^{5} \sum_{k=1}^{n_K} -log\left(p_{\theta_{E_t}, \theta_{M_t}, \theta_{D_t}}(w_k^n | w_{1:k-1}^n, s_{1:n-1}, \tilde{\boldsymbol{I}})\right), \tag{3}$$

where θ_{E_t}, θ_{M_t}, and θ_{D_t} are the parameters of E_t, M_t and D_t; other definitions are similar to those in Equation (1).

3.1.3 Training Process

Note that it is a challenging training task if E_y and E_t, or G_y and G_t are totally independent. To make constraint, the two tasks share the same image sequence encoder E and memory unit M, while the decoder has partially dependent parameters to constrain styles (factual/target style). In general, our training loss function is as follows:

$$\begin{aligned} \mathcal{L}\left(\theta_E, \theta_M, \theta_{D_y}, \theta_{D_t}\right) &= \lambda \mathcal{L}_Y\left(\theta_E, \theta_M, \theta_{D_y}\right) \\ &+ (1-\lambda) \mathcal{L}_T\left(\theta_E, \theta_M, \theta_{D_t}\right), \end{aligned} \tag{4}$$

where θ_E and θ_M are the set of parameters in shared E and M; $\theta_{D_y}, \theta_{D_t}$ are parameters for the style-oriented decoder; λ is the hyper parameter.

3.2 Memory-Augmented Style-Oriented Story Generator

We propose a Story Generator which consists of the Memory Unit and Style-Oriented Decoder.

3.2.1 Memory Unit

Understanding the history of a story can improve its coherence and reduce its redundancy. Inspired by MART (Lei et al., 2020), we design a memory unit to store history in previous images and sentences, serving as a latent story-line. Here we take the visual storytelling task as an example to describe the process of memory augmentation.

When generating a sub-story y_n, we aggregate visual hidden states H_n and the memory state M_{n-1} from the previous step. Concretely, we project H_n into an intermediate memory hidden state $\overline{H}_n$, then feed a multi-head memory attention module with the following inputs:

$$\begin{aligned} Q &= \overline{H}_n \\ K, V &= [M_{n-1}; \overline{H}_n] \end{aligned} \tag{5}$$

The memory-augmented hidden states will then pass through a feed forward layer, and be merged with the visual hidden states H_n using a residual connection and layer normalization. Finally, we obtain the augmented H_n' as input for the style-oriented decoder.

Meanwhile, we update the memory state with $\overline{H}_n$ and $\overline{y}_n$ (the memory intermediate state for the CLIP feature of sentence y_n):

$$\begin{aligned} U_n &= \mathrm{MultiHeadAtt}(M_{n-1}, [\overline{H}_n; \overline{y}_n], [\overline{H}_n; \overline{y}_n]) \\ C_n &= \tanh(W_{mc} M_{n-1} + W_{uc} U_n + b_c) \\ Z_n &= \mathrm{sigmoid}(W_{mz} M_{n-1} + W_{uz} U_n + b_z) \\ M_n &= (1 - Z_n) \odot C_n + Z_n \odot M_{n-1}, \end{aligned} \tag{6}$$

where $W_{mc}, W_{uc}, W_{mz}, W_{uz}$ are trainable weights, and b_c, b_z are trainable bias.

3.2.2 Style-Oriented Decoder

As shown in Figure 2 (c), the parameters of the layer normalization and cross-attention components are style-dependent to constrain the linguistic style (factual/stylized), inspired by Jin et al. (2020) and Chen et al. (2019). Specifically, the *style-dependent layer normalization* would transform the layers' activation x into a style-specific normalized activation:

$$\hat{x} = \gamma_s(\frac{x - \mu}{\delta}) - \beta_s, \tag{7}$$

where μ and δ are the mean and standard deviation of x. γ_s and β_s are style-specific parameters.

Our *style-dependent cross-attention* aims to apply diverse attention strategies for specific styles, as different styles might focus on different semantic content during prediction. The attention function among the cross-attention layer is defined as follows:

$$\begin{aligned} Q &= \text{query} \cdot W_q^s \\ K &= \text{key} \cdot W_k \\ V &= \text{value} \cdot W_v \\ \mathrm{Att}(Q, K, V) &= \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V, \end{aligned} \tag{8}$$

where the query refers to H_n'; key and value refer to embeddings of previous words; the parameter W_q^s

[5] https://spacy.io/api/doc#noun_chunks

is specified for different styles to pass information based on the current style; d is the dimension of the hidden states.

During inference, the style-specific parameters γ_s, β_s, and W_q^s of the target style are used to generate stylized stories.

4 Experiments

4.1 Datasets

VIST Dataset The visual storytelling (Huang et al., 2016) dataset consists of 210,819 images and 50,200 stories collected from 10,117 Flicker albums. Our experiments use the same split settings as in Huang et al. (2016), with 40,098 for training, 4,988 for validation and 5,050 for testing. Each sample contains five images and a factual story with five sub-stories.

Three Target Style Corpus We collect stories from the open-source Project Gutenberg[6] (Hart, 1992) in three different styles: fairy tale, romance and humor. We process the original long stories into several short stories, as detailed in Appendix A.1. Each story sample consists of five sub-stories.

Table 1: Data size of three target style corpora.

	Fairy Tale	Romance	Humor
Train	71,340	70,744	41,392
Valid	8,917	8,842	5,174

4.2 Implementation Details

We adopt the fairseq code base (Ott et al., 2019). For story generation, we build a vocabulary dictionary with 30,820 words and replace words that appear less than 5 times with [UNK]. Both our transformer-based encoder and decoder are composed of 6 stacks of multi-head attention layers and feed-forward layers, with a hidden size of 512 and attention heads of 8. For the memory module, the length of the memory state is set to 1 (Lei et al., 2020). During training, we apply the Adam optimizer (Kingma and Ba, 2014), with 64 factual stories and 64 stylized stories for each epoch. λ is set to 0.5 for multi-task training, which ensures a trade-off between style and sentiment. During decoding, we adopt beam search with a size of 5.

4.3 Baselines

We compare our proposed StyleVSG with the following strong baseline methods:

Seq2Seq+STRAP We first apply our framework to generate factual stories, which achieve comparable SOTA results in the VST task, and then use the unsupervised text style transfer method STRAP (Krishna et al., 2020) to obtain transferred stylized stories.

DLN It applies an intermediate latent space from the visual space to generate factual or stylized text (Chen et al., 2019; Tan et al., 2022), as shown in Figure 3. For the sake of fairness, we use the same architecture as StyleVSG. The only difference is that we apply CLIP text features as input for the stylized story reconstruction task.

SemStyle Mathews et al. (2018) employs visual-relevant semantic terms to generate stylized captions. This method can be ported to SVST task, by using the overall image sequence to generate semantic terms for each sub-story. Here we apply two methods of generating semantic terms: (1) **SemStyle-GRU** denotes to use a GRU-based decoder to predict story-like terms (Hsu et al., 2020). (2) **SemStyle-Transformer** denotes applying our memory-augmented framework, which can generate more concrete semantic terms.

5 Results and Discussions

5.1 Automatic Metric Results

We evaluate our results from three aspects: semantic relevance, style accuracy, and fluency.

- Semantic relevance is measured by **METEOR** (Banerjee and Lavie, 2005) and **CIDEr** (Vedantam et al., 2015). METEOR is reported to be most consistent with human evaluation in VST task (Huang et al., 2016); and CIDEr is efficient to measure visual relevance (Vedantam et al., 2015). Here we utilize the open source evaluation code[7] used in previous works.
- Style accuracy (**CLS**) is estimated by a style classifier, which is fine-tuned from the BERT-base classifier (Devlin et al., 2018). For each style, a binary classifier is trained on stories with the target style and stories from the VIST training set. The classifiers achieve an average accuracy of 98% on the validation set.

[6] https://www.gutenberg.org/

[7] https://github.com/lichengunc/vist_eval/

Table 2: Automatic evaluation of StyleVSG and several baselines (Section 4.3). w.r.t. METEOR (M), CIDEr (C), style accuracy (CLS), Geometric Mean (GM) and perplexity (ppl.). The main metrics are GM(·), which measures the overall performance of sentiment and style; and ppl. , which stands for fluency. We color the best score for the main metrics. The best result on each aspect is **bolded** and the second best is undelined.

Style	Model	M↑	C↑	CLS↑	GM(M, CLS)↑	GM(C, CLS)↑	ppl.↓
Factual	StyleVSG	36.6	12.1	-	-	-	-
Fairy tale	Seq2Seq+STRAP	30.0	4.2	54.38	40.39	15.12	41.55
	DLN	29.5	1.6	74.57	46.90	10.92	17.13
	SemStyle-Transformer	30.1	**5.0**	79.58	48.94	19.95	26.47
	SemStyle-GRU	28.9	3.9	97.54	53.09	19.50	19.63
	StyleVSG (Ours)	**30.2**	4.8	**98.35**	**54.50**	**21.73**	**9.48**
Romance	Seq2Seq+STRAP	30.6	5.2	20.63	25.13	10.36	45.13
	DLN	29.7	1.2	48.38	37.91	7.62	20.51
	SemStyle-Transformer	30.6	**6.1**	34.30	32.40	14.46	28.96
	SemStyle-GRU	28.4	3.9	79.24	47.44	17.58	22.53
	StyleVSG (Ours)	**30.8**	4.8	**83.46**	**50.70**	**20.02**	**12.03**
Humor	Seq2Seq+STRAP	30.6	**5.5**	13.71	20.48	8.68	40.67
	DLN	29.5	1.4	33.82	31.59	6.88	19.21
	SemStyle-Transformer	29.5	4.4	62.30	42.87	16.56	33.63
	SemStyle-GRU	29.0	3.2	75.60	46.82	15.55	33.47
	StyleVSG (Ours)	**31.0**	4.6	**80.23**	**49.87**	**19.21**	**12.75**

- Fluency is judged by the average perplexity score (**ppl.**) of three GPT-2 models (Lagler et al., 2013) fine-tuned on each stylized story corpus. Lower ppl. means more fluent and appropriately stylized (Zhao et al., 2020; Li et al., 2021).

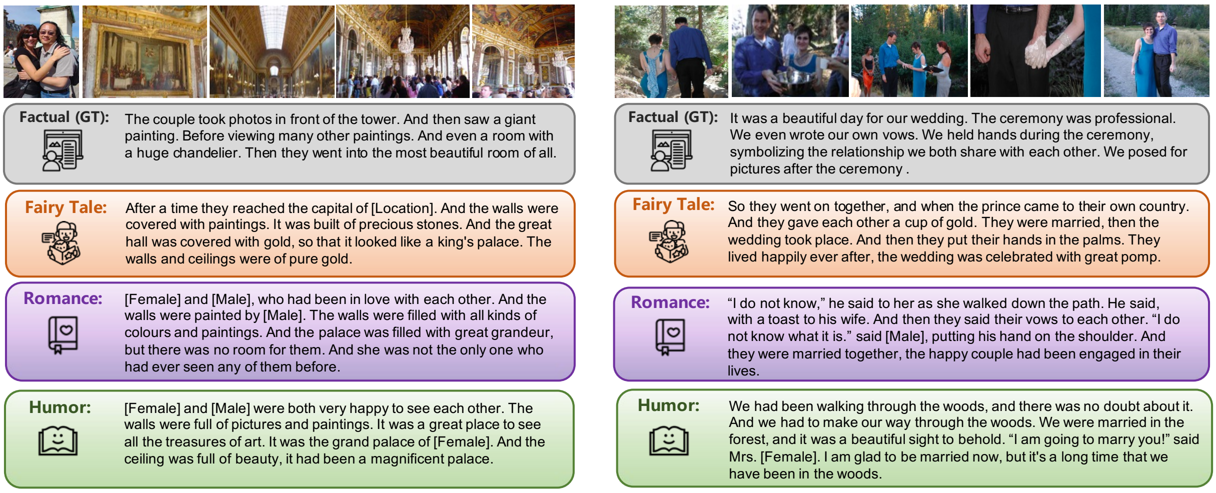
To measure the overall performance in terms of both semantics and style, we follow Hu et al. (2022b) and Tan et al. (2022) to compute the geometric mean score, which is denoted as **GM(·)**.

Table 2 summarizes the results from our StyleVSG and other compared models. For all styles, StyleVSG achieves the best score in the most important metrics, "Geometric Mean (GM)" and "perplexity (ppl.)". We observe that for the token-level split method SemStyle, performance is highly dependent on the quality of semantic terms. If the terms are inaccurate or insufficient, it will generate stories with stylistic imaginations but with bad visual relevance, as show by the results of SemStyle-GRU. If the terms are more accurate, the semantic scores are better. However, the generated stories will be limited by factual-like terms, resulting in lower style accuracy, as shown by the results of SemStyle-Transformer. Furthermore, for complicated styles such as fairy tale, there are many stylistic semantic terms, such as "king" and "fisherman", that may not be generated by a term generation model trained on the VIST dataset. For the latent-level split method DLN, even with the same structure and applying CLIP text features as input, the performance is much worse than StyleVSG, verifying the limitations of latent-split methods as described in Section 3.1.

Table 3: Results of three StyleVSG models, each holding a single style; and a single StyleVSG-Multi model holding multiple styles. The average score GM is calculated using the geometric mean of METEOR (M), CIDEr (C) and style accuracy (CLS). Overall, the performance is competitive.

Style	Model	M↑	C↑	CLS↑	GM↑
Fairy Tale	StyleVSG	30.2	4.8	98.35	**24.25**
	StyleVSG-Multi	30.2	4.3	99.1	23.43
Romance	StyleVSG	30.8	4.8	83.46	**23.11**
	StyleVSG-Multi	30.7	4.8	79.1	22.67
Humor	StyleVSG	31.0	4.6	80.23	22.53
	StyleVSG-Multi	31.2	5.2	83.08	**23.80**

To demonstrate the flexibility of our model, we expand it to include three target styles, which we call StyleVSG-Multi. Specifically, we simultaneously train the factual visual storytelling task on the VIST dataset and the stylized story reconstruction task on the three target style corpora. Except for style-specific parameters, all other parameters are shared. The results are shown in Table



Factual (GT): The couple took photos in front of the tower. And then saw a giant painting. Before viewing many other paintings. And even a room with a huge chandelier. Then they went into the most beautiful room of all.

Fairy Tale: After a time they reached the capital of [Location]. And the walls were covered with paintings. It was built of precious stones. And the great hall was covered with gold, so that it looked like a king's palace. The walls and ceilings were of pure gold.

Romance: [Female] and [Male], who had been in love with each other. And the walls were painted by [Male]. The walls were filled with all kinds of colours and paintings. And the palace was filled with great grandeur, but there was no room for them. And she was not the only one who had ever seen any of them before.

Humor: [Female] and [Male] were both very happy to see each other. The walls were full of pictures and paintings. It was a great place to see all the treasures of art. It was the grand palace of [Female]. And the ceiling was full of beauty, it had been a magnificent palace.

Factual (GT): It was a beautiful day for our wedding. The ceremony was professional. We even wrote our own vows. We held hands during the ceremony, symbolizing the relationship we both share with each other. We posed for pictures after the ceremony .

Fairy Tale: So they went on together, and when the prince came to their own country. And they gave each other a cup of gold. They were married, then the wedding took place. And then they put their hands in the palms. They lived happily ever after, the wedding was celebrated with great pomp.

Romance: “I do not know,” he said to her as she walked down the path. He said, with a toast to his wife. And then they said their vows to each other. “I do not know what it is.” said [Male], putting his hand on the shoulder. And they were married together, the happy couple had been engaged in their lives.

Humor: We had been walking through the woods, and there was no doubt about it. And we had to make our way through the woods. We were married in the forest, and it was a beautiful sight to behold. “I am going to marry you!” said Mrs. [Female]. I am glad to be married now, but it's a long time that we have been in the woods.

Figure 4: Examples of stylized stories generated by our proposed StyleVSG. Note that the goal of Stylized Visual Storytelling is to describe a photo stream with an abstract story, which reflects a specific linguistic style and contains imaginary concepts.

3. StyleVSG-Multi can achieve almost the same performance as StyleVSG (three models for three styles), but with far fewer parameters.

5.2 Human Evaluation Results

We also conduct human evaluations using three metrics: (1) **Relevance** that measures relationship between the generated stories and the source photo stream. (2) **Style appropriateness** which means how well the stories express the target style. (3) **Coherence** that measures the inter-sentence coherency of the whole story. Following the standard in Guo et al. (2019), relevance is rated from 0 (unrelated) to 3 (very related), coherence from 0 (unreadable) to 3 (perfect), and style appropriateness from 0 (bad) to 3 (perfect).

We randomly select 50 image sequences from the testing set and generate stylized stories with different models, resulting in a total of 50 × 3 visual-story pairs. Each pair to be evaluated contains 5 images and 5 sub-stories. In this evaluation, we focus on fairy tale style, which is the most distinctive and attractive. The results are collected from 14 evaluators[8], and the average inter-rater reliability is 0.71 in terms of Pearson Correlation Coefficient (Fleiss and Cohen, 1973). As shown in Table 4, StyleVSG achieves the best score in all three metrics.

Table 4: Human evaluation results (Section 5.2) on three aspects.

Model	Relevance↑	Style↑	Coherence↑
STRAP	2.15	1.08	1.45
SemStyle	1.46	1.35	1.34
StyleVSG	**2.37**	**2.22**	**2.16**

Table 5: Ablation studies on fairy tale style. As in Hu et al. (2022b), the average score GM is measured by the geometric mean of METEOR, CIDEr, CLS and 1/ppl. .

Model	M↑	C↑	CLS↑	ppl. ↓	GM↑
StyleVSG	30.2	**4.8**	**98.35**	9.48	**6.23**
w/o OE	30.2	4.6	97.57	9.48	6.15
w/o Mem	**30.6**	4.0	98.10	9.83	5.91
w/o VisMem	29.9	4.1	98.09	9.26	6.00
w/o MultiTask	29.1	3.6	94.48	13.82	5.17
w/o ITM	29.2	2.7	97.38	**8.23**	5.53
obj input	29.0	2.0	91.70	8.89	4.95

5.3 Ablation Study

We conduct ablation studies to verify the effectiveness of different components.

w/o OE To evaluate the contribution of image order embedding, this feature is removed.

w/o Memory To evaluate the contribution of history memory, we remove the memory unit M.

w/o VisMem To verify the contribution of visual memory, we only use previous textual information when updating the memory unit.

[8] Graduate students who are fluent in English, and familiar with research of text generation.

w/o MultiTask To verify the effectiveness of our multi-task training strategy, we first train a factual visual storytelling model using the VIST dataset, and then use the stylized corpus to fine-tune it.

w/o ITM While mining roughly parallel visual pairs, instead of applying image-text matching, we try to retrieve the most similar sentence in factual VIST stories, and apply the corresponding image as pseudo image.

Obj input For each stylized sub sentence s_n, we find 10 most related visual objects in the photo set of VIST to reconstruct stylized story.

The results of ablation studies are reported in Table 5. We can observe that: (1) Without the image order embeddings, the style accuracy and fluency drop, which indicates that temporal information benefits the specific stylistic structure of a generated story. (2) Removing the memory unit reduces the fluency of generated stories, which demonstrates the effectiveness of our memory-augmented structure. (3) If we only apply textual history to update the memory state, the overall performance drops. This suggests that visual memory primarily benefits visual relevance, as judged by METEOR and CIDEr. (4) After fine-tuning on the stylized corpus, when we do inference on VIST image sequences, the domain gap will confuse the model. Our multitasking setting could guarantee both content preservation and stylization. (5) Without applying image-text alignment, the quality of pseudo images will drop significantly, leading to the decrease in semantic metrics. (6) Although the pseudo objects for one stylized sub-story are more closely related to this story, some of them are unlikely to come from the same photo, which can lead to a drop in final performance.

5.4 Qualitative Examples

Figure 4 represents some stories generated by StyleVSG in three different styles: fairy tale, romance, and humor. Our model can generate attractive stories for photo streams taken in our daily lives. The linguistic style is reflected throughout the entire story. More cases can be found in Appendix A.2.

6 Conclusion

In this paper, we propose a new task of stylized visual storytelling, aiming to generate attractive stylized stories for a photo stream. By applying style-dependent components and multi-task training, our proposed StyleVSG is able to generate stylistic stories without paired (images, stylized story) corpus. Furthermore, our memory unit can preserve the coherence of generated stories. Experiments demonstrate that StyleVSG achieves better overall performance for complicated styles.

Limitations

While imaginary concepts are encouraged in stylized visual storytelling task, it would be better if these literary imaginations are more related to visual contents. In order to improve semantic relevance, we could restrain models from generating visually unrelated descriptions, or make pseudo images more related to stylized stories. However, the former solution is likely to harm the style expression by decreasing stylistic imaginations. For the latter scheme, we have tried to generate pseudo visual inputs with pre-trained text2image model (Ramesh et al., 2022), however, there is a domain gap between photos in VIST and images generated with stylized sentences. It would be a challenging and interesting problem to be explored in the future.

Ethics Statement

We acknowledge the Code of Ethics and Professional Conduct and strictly adhere to the rules throughout this research. We would like to note that the style corpus might be further filtered by human beings to decrease the possibility of generating stories with offensive content.

Acknowledgements

This work was partially supported by the National Key R&D Program of China (No. 2020AAA0108600) and the National Natural Science Foundation of China (No. 62072462).

References

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.

Cheng-Kuan Chen, Zhufeng Pan, Ming-Yu Liu, and Min Sun. 2019. Unsupervised stylish image description generation via domain layer norm. In *Proceedings of* *the AAAI Conference on Artificial Intelligence*, pages 8151–8158.

Tianlang Chen, Zhongping Zhang, Quanzeng You, Chen Fang, Zhaowen Wang, Hailin Jin, and Jiebo Luo. 2018. "factual"or"emotional": Stylized image captioning with adaptive learning and attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 519–535.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3137–3146.

Longteng Guo, Jing Liu, Peng Yao, Jiangwei Li, and Hanqing Lu. 2019. Mscap: Multi-style image captioning with unpaired stylized text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4204–4213.

Longteng Guo, Jing Liu, Xinxin Zhu, Peng Yao, Shichen Lu, and Hanqing Lu. 2020. Normalized and geometry-aware self-attention network for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10327–10336.

Michael Hart. 1992. The history and philosophy of project gutenberg. *Project Gutenberg*, 3:1–11.

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*, pages 7952–7960.

Chi-Yang Hsu, Yun-Wei Chu, Ting-Hao'Kenneth' Huang, and Lun-Wei Ku. 2021. Plot and rework: Modeling storylines for visual storytelling. *arXiv preprint arXiv:2105.06950*.

Ting-Yao Hsu, Chieh-Yang Huang, Yen-Chia Hsu, and Ting-Hao'Kenneth' Huang. 2019. Visual story post-editing. *arXiv preprint arXiv:1906.01764*.

Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022a. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17980–17989.

Zhiqiang Hu, Roy Ka-Wei Lee, Charu C Aggarwal, and Aston Zhang. 2022b. Text style transfer: A review and experimental evaluation. *ACM SIGKDD Explorations Newsletter*, 24(1):14–45.

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.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, Lisa Orii, and Peter Szolovits. 2020. Hooks in the headline: Learning to generate headlines with controlled styles. *arXiv preprint arXiv:2004.01980*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. *arXiv preprint arXiv:2010.05700*.

Klemens Lagler, Michael Schindelegger, Johannes Böhm, Hana Krásná, and Tobias Nilsson. 2013. Gpt2: Empirical slant delay model for radio space geodetic techniques. *Geophysical research letters*, 40(6):1069–1073.

Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara L Berg, and Mohit Bansal. 2020. Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning. *arXiv preprint arXiv:2005.05402*.

Guodun Li, Yuchen Zhai, Zehao Lin, and Yin Zhang. 2021. Similar scenes arouse similar emotions: Parallel data augmentation for stylized image captioning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5363–5372.

Nanxing Li, Bei Liu, Zhizhong Han, Yu-Shen Liu, and Jianlong Fu. 2019. Emotion reinforced visual storytelling. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 297–305.

Ruibo Liu, Chongyang Gao, Chenyan Jia, Guangxuan Xu, and Soroush Vosoughi. 2022. Non-parallel text style transfer with self-parallel supervision. *arXiv preprint arXiv:2204.08123*.

Holy Lovenia, Bryan Wilie, Romain Barraud, Samuel Cahyawijaya, Willy Chung, and Pascale Fung. 2022. Every picture tells a story: Image-grounded controllable stylistic story generation. *arXiv preprint arXiv:2209.01638*.

Alexander Mathews, Lexing Xie, and Xuming He. 2016. Senticap: Generating image descriptions with sentiments. In *Proceedings of the AAAI conference on artificial intelligence*.

Alexander Mathews, Lexing Xie, and Xuming He. 2018. Semstyle: Learning to generate stylised image captions using unaligned text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8591–8600.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

Shrimai Prabhumoye, Khyathi Raghavi Chandu, Ruslan Salakhutdinov, and Alan W Black. 2019. " my way of telling a story": Persona based grounded story generation. *arXiv preprint arXiv:1906.06401*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Yutong Tan, Zheng Lin, Peng Fu, Mingyu Zheng, Lanrui Wang, Yanan Cao, and Weipinng Wang. 2022. Detach and attach: Stylized image captioning without paired stylized dataset. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4733–4741.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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.

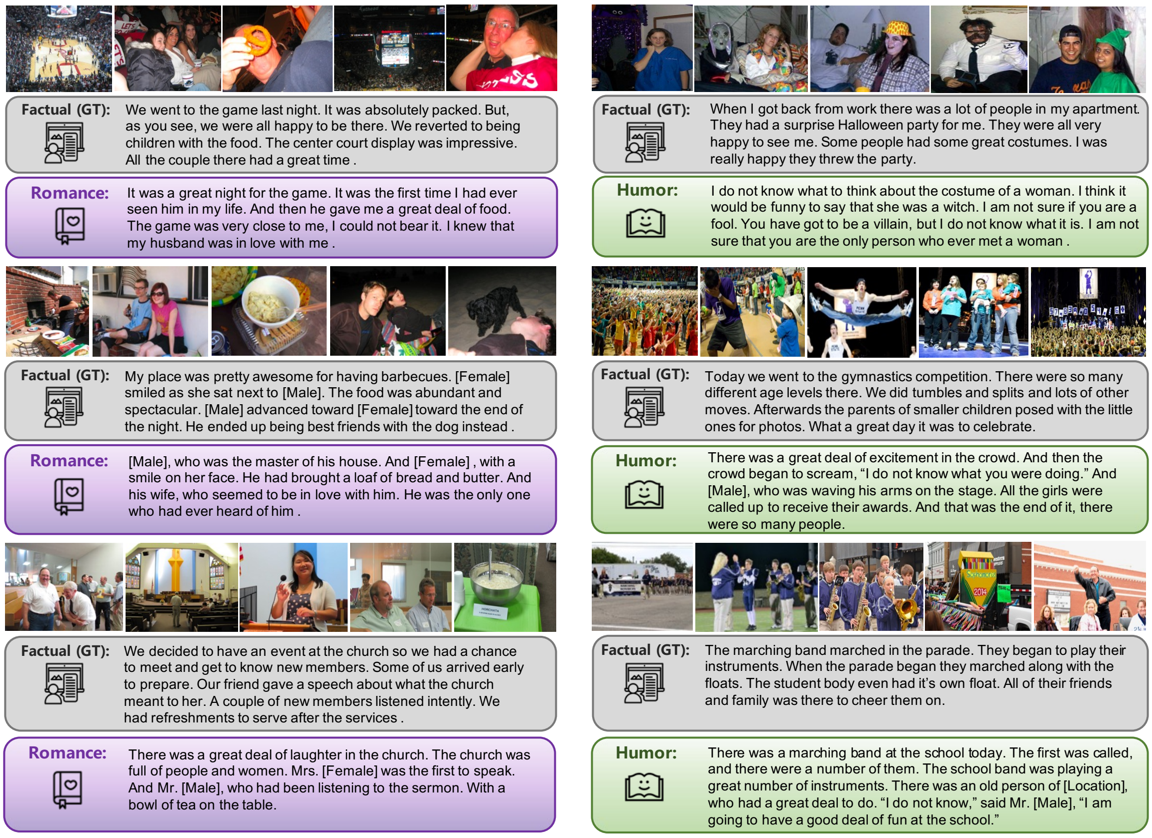
Chunpu Xu, Min Yang, Chengming Li, Ying Shen, Xiang Ao, and Ruifeng Xu. 2021. Imagine, reason and write: Visual storytelling with graph knowledge and relational reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3022–3029.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

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*, page 7.

Zhexin Zhang, Jiaxin Wen, Jian Guan, and Minlie Huang. 2022. Persona-guided planning for controlling the protagonist's persona in story generation. *arXiv preprint arXiv:2204.10703*.

Wentian Zhao, Xinxiao Wu, and Xiaoxun Zhang. 2020. Memcap: Memorizing style knowledge for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12984–12992.



Factual (GT): We went to the game last night. It was absolutely packed. But, as you see, we were all happy to be there. We reverted to being children with the food. The center court display was impressive. All the couple there had a great time .

Romance: It was a great night for the game. It was the first time I had ever seen him in my life. And then he gave me a great deal of food. The game was very close to me, I could not bear it. I knew that my husband was in love with me .

Factual (GT): My place was pretty awesome for having barbecues. [Female] smiled as she sat next to [Male]. The food was abundant and spectacular. [Male] advanced toward [Female] toward the end of the night. He ended up being best friends with the dog instead .

Romance: [Male], who was the master of his house. And [Female] , with a smile on her face. He had brought a loaf of bread and butter. And his wife, who seemed to be in love with him. He was the only one who had ever heard of him .

Factual (GT): We decided to have an event at the church so we had a chance to meet and get to know new members. Some of us arrived early to prepare. Our friend gave a speech about what the church meant to her. A couple of new members listened intently. We had refreshments to serve after the services .

Romance: There was a great deal of laughter in the church. The church was full of people and women. Mrs. [Female] was the first to speak. And Mr. [Male], who had been listening to the sermon. With a bowl of tea on the table.

Factual (GT): When I got back from work there was a lot of people in my apartment. They had a surprise Halloween party for me. They were all very happy to see me. Some people had some great costumes. I was really happy they threw the party.

Humor: I do not know what to think about the costume of a woman. I think it would be funny to say that she was a witch. I am not sure if you are a fool. You have got to be a villain, but I do not know what it is. I am not sure that you are the only person who ever met a woman .

Factual (GT): Today we went to the gymnastics competition. There were so many different age levels there. We did tumbles and splits and lots of other moves. Afterwards the parents of smaller children posed with the little ones for photos. What a great day it was to celebrate.

Humor: There was a great deal of excitement in the crowd. And then the crowd began to scream, "I do not know what you were doing." And [Male], who was waving his arms on the stage. All the girls were called up to receive their awards. And that was the end of it, there were so many people.

Factual (GT): The marching band marched in the parade. They began to play their instruments. When the parade began they marched along with the floats. The student body even had it's own float. All of their friends and family was there to cheer them on.

Humor: There was a marching band at the school today. The first was called, and there were a number of them. The school band was playing a great number of instruments. There was an old person of [Location], who had a great deal to do. "I do not know," said Mr. [Male], "I am going to have a good deal of fun at the school."

Figure 5: More examples of romance and humor stories generated by StyleVSG.

A Appendix

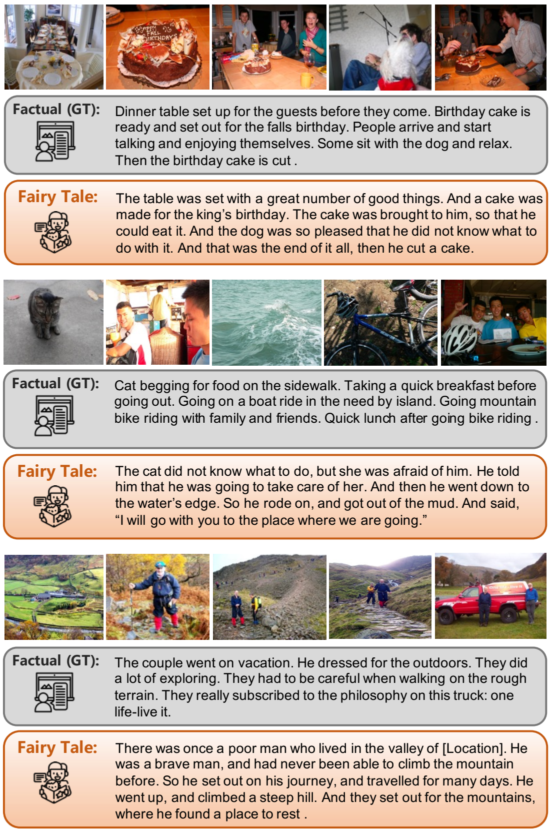
A.1 Style Corpus Processing

When processing long stories from the open-source Project Gutenberg[9] (Hart, 1992), we break them down into shorter sentences. If the sub-sentences in one long sentence have more than 3 noun chunks, we consider it to have enough visually-related information and aggregate them as a new sentence. After processing, each story sample consists of 5 aggregated sentences.

In addition, we apply a name entity recognition tagger [10] to replace the low-frequency words. The names of person, location, and organization are replaced by [Male]/[Female]/[Person], [Location], and [Organization], respectively.

A.2 More Generated Examples

In Figure 5 and 6, we represent more examples of stylized stories generated by StyleVSG.



Factual (GT): Dinner table set up for the guests before they come. Birthday cake is ready and set out for the falls birthday. People arrive and start talking and enjoying themselves. Some sit with the dog and relax. Then the birthday cake is cut .

Fairy Tale: The table was set with a great number of good things. And a cake was made for the king's birthday. The cake was brought to him, so that he could eat it. And the dog was so pleased that he did not know what to do with it. And that was the end of it all, then he cut a cake.

Factual (GT): Cat begging for food on the sidewalk. Taking a quick breakfast before going out. Going on a boat ride in the need by island. Going mountain bike riding with family and friends. Quick lunch after going bike riding .

Fairy Tale: The cat did not know what to do, but she was afraid of him. He told him that he was going to take care of her. And then he went down to the water's edge. So he rode on, and got out of the mud. And said, "I will go with you to the place where we are going."

Factual (GT): The couple went on vacation. He dressed for the outdoors. They did a lot of exploring. They had to be careful when walking on the rough terrain. They really subscribed to the philosophy on this truck: one life-live it.

Fairy Tale: There was once a poor man who lived in the valley of [Location]. He was a brave man, and had never been able to climb the mountain before. So he set out on his journey, and travelled for many days. He went up, and climbed a steep hill. And they set out for the mountains, where he found a place to rest .

Figure 6: More examples of fairy tale stories generated by StyleVSG.

[9] https://www.gutenberg.org/

[10] https://github.com/kamalkraj/BERT-NER

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations

☑ A2. Did you discuss any potential risks of your work?
Ethics Statement

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

4.3

☑ B1. Did you cite the creators of artifacts you used?
4.3

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
4.3

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
4.3

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
A.1

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
4.3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
4.1

C ☑ Did you run computational experiments?

5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
4.2

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4.2

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
5.1

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
4.2

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
5.2

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
5.2

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
5.2

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
5.2

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
5.2

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
5.2