Generating a Temporally Coherent Visual Story by Multimodal Recurrent Transformers

Anonymous ACL submission

Abstract

Story visualization is a challenging text-to-image generation task for the difficulty of rendering visual details from abstract text descriptions. Besides the difficulty of image generation, the generator also needs to conform to the narrative of a multi-sentence story input. While prior arts in this domain have focused on improving semantic relevance between generated images and input text, controlling the generated images to be temporally consistent still remains a challenge. Moreover, existing generators are trained on single text-image pairs and fail to consider the variations of natural language captions that can describe a given image, causing poor model generalization. To address such problems, we leverage a cyclic training methodology involving pseudo-text descriptions as an intermediate step that decouples the image's visual appearance from the variations of natural language descriptions. Additionally, to generate a semantically coherent image sequence, we consider an explicit memory controller which can augment the temporal coherence of images in the multi-modal autoregressive transformer. To sum up all components, we call it **C**yclic **S**tory visualization by **M**ultimod**A**l **R**ecurrent **T**ransformers or C-SMART for short. Our method generates high-resolution, high-quality images, outperforming prior works by a significant margin across multiple evaluation metrics on the Pororo-SV dataset.

1 Introduction

Story visualization is a challenging task involving generating a sequence of images given natural language paragraph. A story consists of a sequence of pairs of texts and images where the pairs are temporally coherent as a story. Our task is to reproduce the images given the multi-sentence paragraph input. It is more challenging than the conventional text-to-image generation task owing to additional objectives such as understanding narrative in text, semantic relevance and temporal consistency, *e.g.* foreground and background consistency, in the generated sequence of images, as depicted in Fig. 1.

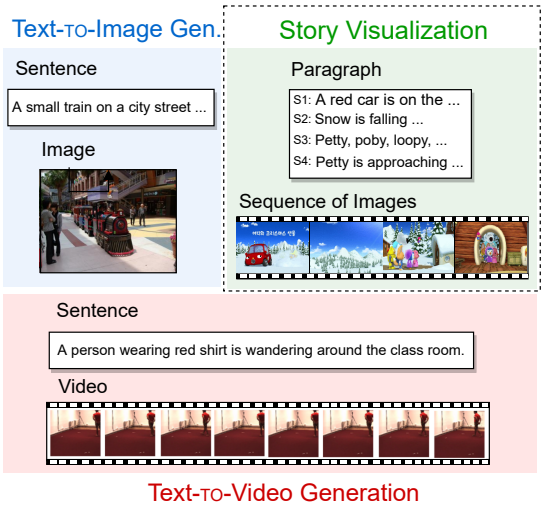

Figure 1: **Comparing visual generation tasks from texts.** Story visualization task aims to generate a sequence of images to describe a given story written in a natural language paragraph and is different from text-to-image or video generation.

The common approaches to tackle this task are to build sequential GAN framework (Li et al., 2019) and deep recurrent context encoder to track the story flow. However, all these efforts are limited to training with a single input text description per image, even though many possible descriptions may depict the target image.

Multiple descriptions for a paired image are essential for the generalization of text-to-image generators. They account for the variation in natural language descriptions that may be used to describe a scene. Thus, most image captioning datasets (Lin et al., 2014) comprise multiple natural language directives, obtained by a rigorous human annotation, in both the training and evaluation splits. Nevertheless, the Pororo-SV dataset for story visualization task (Li et al., 2019) consists of a single text-image pair in the training and evaluation splits.

To address this, we propose the cyclic pseudo-

text approach comprising an Image to Pseudo-Text to Image stream for training the text-to-image generator. The pseudo-text provides contextual information absent in the human-annotated label inputs, making the text-to-image generator to be more generalizable to natural language variations. (Sec. 3.3) We implement this cyclic optimization using a bi-directional generative model, *e.g*. text-to-image and image-to-text in a unified architecture, to produce multimodal outputs, as shown in Fig. 2. Because the cross-modal generation requires the model to fully understand the source modality, this iterative forward/reverse generative process (*i.e*. Forward: text-image, Reverse: image-text) could help generate the target modality with high semantic consistency with the source (Huang et al., 2021). In addition, we address the temporal consistency by developing a dynamic gated-memory module in a multimodal recurrent autoregressive transformer, inspired by (Maharana et al., 2021; Lei et al., 2020b).

To sum up all the components, we call our proposed model architecture as C-SMART (**C**yclic **S**tory visualization by **M**ultimod**A**l **R**ecurrent **T**ransformers). The experimental results manifest that we can improve the quality of visualized stories with a large margin on various evaluated metrics compared to prior works. (Sec. 5)

We summarize our contributions as follows:

- propose the first bidirectional generative model using multimodal self-attention on long-range input of text and image in a recurrent manner for generating a temporally coherent image (or text) sequence given text (or image) in an unified framework.
- exploit the nature of bi-directional multi-modal generation and cyclically generate pseudo-text for supplying contextual information deficit in the human-annotated label for improved generalization of the text-to-image generator.
- explicitly generate sequences of images at a higher resolution with higher quality than ever before on a benchmark dataset.
- significantly outperform prior works by a large margin along with various evaluation metrics for the image quality, temporal coherency, and global semantic matching between generated images and natural language descriptions.

2 Related Work

Text-to-Image generation. Text-based image synthesis has been widely studied recently. Most papers in this area focus on enhancing the semantic relevance of the generated image for the input text description and on resolution improvements. MC-GAN (Park et al., 2018) models both background and foreground information to generate photo realistic foreground objects for a background. StackGAN (Zhang et al., 2017) uses a two-stage process to enhance the resolution of the image conditioned on an input text description. Subsequent works focus on architectural enhancements over StackGAN. This is accomplished by either adding attention networks for improved semantic relevance, extending the two-stage process, or adding memory networks to improve the resolution of generated images and others (Xu et al., 2018; Zhang et al., 2018; Zhu et al., 2019; Gao et al., 2019). Most recently, text-based image synthesis has been studied in a zero-shot setting. DALL-E (Ramesh et al., 2021) proposes an autoregressive transformer to model the text and image as a single data stream. More recent approaches utilize the multimodal CLIP model to achieve the same objective (Radford et al., 2021).

Story Visualization. The story visualization task is a more complex counterpart of text-based image generation that has recently garnered research interest. StoryGAN (Li et al., 2019) was the first work in this direction and utilized a story-level discriminator to improve global consistency in generated images. CP-CSV (Song et al., 2020) disentangles figure and background information to enhance character consistency. DuCO-StoryGAN (Maharana et al., 2021) presents video captioning as an auxiliary task for story visualization along with other design improvements to StoryGAN. VLC-StoryGAN (Maharana and Bansal, 2021) uses constituency parse-trees and common sense knowledge to improve consistency and an object-level feedback loop to improve image quality. DuCO-StoryGAN and DALL-E are direct precursors of our work. While DuCO-StoryGAN utilizes MART (Lei et al., 2020b) to encode video captions, DALL-E presents a generation framework based on joint autoregressive modeling of text and images.

Recurrent Transformer. Although transformers have been shown to be effective and superior to RNNs (Vaswani et al., 2017) for sequential modeling, they are still unable to model historical information well. This problem is distinctive in the task of long-range sequential data modeling because of context fragmentation (Dai et al., 2019), *i.e*. each

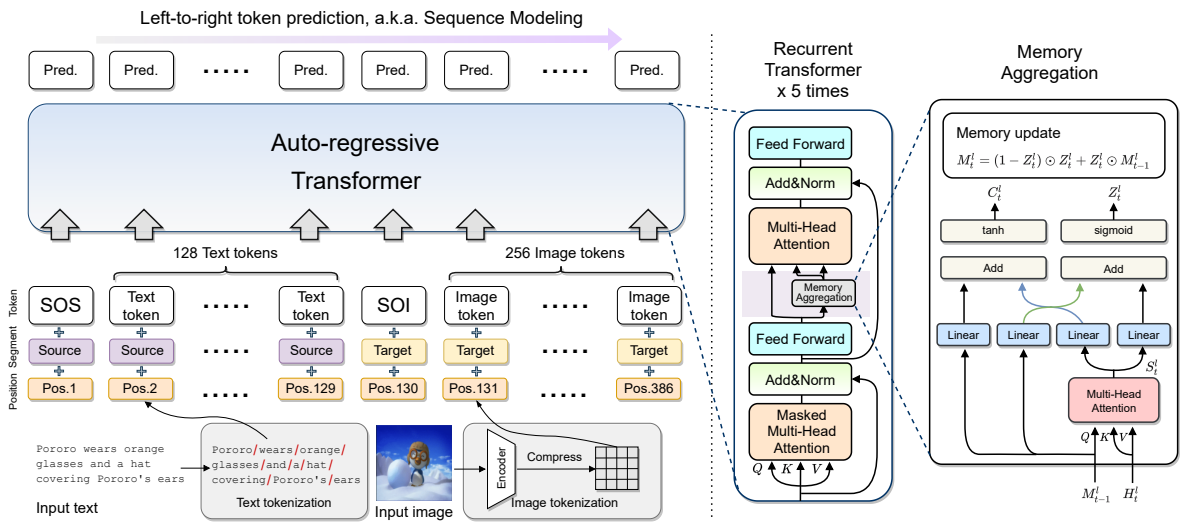

Figure 2: **Proposed Multimodal Recurrent Transformer for generating an image sequence given a multi-sentence paragraph. (Left)**: Illustration of the single text-to-image generation process. With auto-regressive transformer architecture, the training procedure is conducted using left-to-right token prediction, a.k.a. language modeling. **(Right)**: Basic building block of recurrent transformer. Considering historical information (*i.e.*, memory), multi-modal inputs are encoded in a recurrent manner.

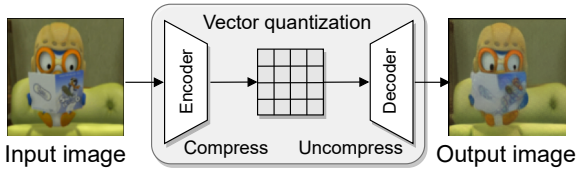

Figure 3: **Image tokenization using VQ-VAE.**

language segment is dealt with individually without knowing its surrounding context. This can lead to inefficient optimization and inferior performance because every segment is treated separately. To address this, (Lei et al., 2020a) transplanted a recurrent path into transformer architecture. Specifically, the modeling of new data segments is conditioned on historical hidden states produced in the previous layer and use highly summarized memory states. With this, they explicitly bridge useful prior semantic or linguistic cues to future segments.

3 Methods

3.1 Recurrent Text-to-Image Generation

Recently, researchers have explored the potential of transformer-based models for diverse generative tasks, *e.g.* image/video captioning (Cornia et al., 2020; Zhou et al., 2018), video prediction, text-to-image generation etc (Wu et al., 2021). Inspired by the autoregressive generative models for pixel-by-pixel image generation (Chen et al., 2018), the transformer-based methods (Ding et al., 2021) have shown promising results for text-to-image synthesis.

Considering this, we propose C-SMART to generate a semantically relevant and temporally consistent sequence of images corresponding to an input multi-sentence story. We train the model using a two-stage training procedure, similar to DALL-E (Ramesh et al., 2021). In contrast to the single-stream context-agnostic generation in DALL-E, our model utilizes a recurrent multimodal transformer architecture with dynamic aggregation of historical information for context-aware image sequence generation.

To train the model, we first compress the image into a discretized set of latent features called image tokens. This is achieved using a Vector Quantized Variational Autoencoder (VQ-VAE) (van den Oord et al., 2017) for improved computational efficiency. Second, we recurrently train the multimodal autoregressive transformer model with an infused dynamic gated-memory module to solve the story visualization task. If the source tokens $\{z_1, ..., z_n\}$ discretized from image compression using VQVAE and the textual tokens $\{t_1, ..., t_m\}$ tokenized from the text using the WordPiece tokenizer are concatenated and fed into the model, the loss for single generative model can be summarized as follows.

$$\mathrm{L}_{t2i} = \sum_{k=1}^{n} -\log P(z_k | t_1, ..., t_m, z_1, ..., z_{k-1}) \quad (1)$$

And, the loss for recurrent generative model for sequential story generation can be calculated by weight sum of sequence of single image loss.

Methods	FID↓	FSD↓	Char. F1↑	Frame Acc.↑	BLEU2/3↑
StoryGAN (Li et al., 2019)	134.32	200.10	27.53	10.13	3.25 / 1.21
CP-CSV (Song et al., 2020)	140.24	184.52	21.33	8.78	3.23 / 1.26
DuCo (Maharana et al., 2021)	91.96	171.36	36.13	13.03	3.39 / 1.40
C-SMART (Ours)	**50.24**	**30.40**	**58.11**	**28.06**	**5.30 / 2.34**

Table 1: **Quantitative results on test split of Pororo-SV Dataset.** ↓ indicates 'lower the better' and ↑ indicates 'higher the better'. All experimental results of prior works are reproduced with author's codebase in Appendix.

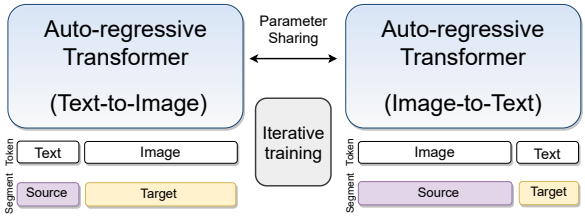

Figure 4: **The unified model architecture for bidirectional text-image and image-text generation.**

Image Tokenization. Image tokens are generated at the compression stage of training. Real images usually consist of millions of pixels which make the generative process extremely expensive. In the compression stage, we use a VQ-VAE (van den Oord et al., 2017) to transform the input images into a set of low-dimensional discrete latent features called image tokens. As shown in Fig. 4, this framework is the autoencoder structure that learns a discretized latent encoding for input data x in the training procedure.

Generating an Image given Text. A model designed for the story visualization task needs to (1) understand the cross-modal relationship between text and images, (2) interpret the narrative of the story from the text, and (3) generate temporally consistent images while maintaining semantic relevance with the input text.

Fig. 2 shows the proposed multimodal recurrent transformer for generating an image sequence given a multi-sentence story. First, we tokenize the text and image inputs for training and add a positional embedding. Both text and image tokens are treated equally and the auto-regressive transformer carries out a language modeling task, *i.e.*, left-to-right token prediction. We then decode the image tokens to form an image using a pre-trained VQ-VAE decoder.

The multimodal self-attention module helps preserve context even over long sequences of text and image tokens and leads to high resolution images. Additionally, we propose a dynamic memory aggregation module for improved narrative understanding, infused in the intermediate layers of the transformer as shown in Fig. 2 (right). The dynamic updates occur as follows (1) intermediate layer is modified for memory aggregation on current stage, and (2) aggregated information is passed through to next stage transformer. This module helps us improve temporal consistency and overall semantic relevance of the generated images by providing easy access to historically aggregated features.

$$\mathrm{M}_t^l = (1 - Z_t^l) \odot Z_t^l + Z_t^l \odot M_{l-1}^l. \qquad (2)$$

Auto-regressive Token Sampling. At the inference phase of the generative model, it is important to sample a plausible data from the model. Because we exploit an autoregressive transformer-based generative model, predicted data tokens can typically be sampled with various decoding strategies such as beam-search (Cohen and Beck, 2019), top-k and nucleus sampling (Holtzman et al., 2019), which are commonly used in many NLP literature (Ippolito et al., 2019). We use nucleus sampling strategy because it is a simple but effective method to draw considerably higher quality tokens out of the generative model (Holtzman et al., 2019). For most of our results, we use nucleus sampling with a rate of $p = 0.9$ unless stated otherwise. We empirically decide the values of the hyper-parameter of p using the FID score of the generated image sequences on the validation dataset.

3.2 Bi-directional Sequence Gen. for VL

Cross-modal generation, aiming at mapping or translating one modality to another, requires the model to fully "understand" the source modality and to faithfully "generate" the target modality with high semantic consistency with the source (Baltrušaitis et al., 2019). In the light of this, many researchers have explored the unification of bidi-

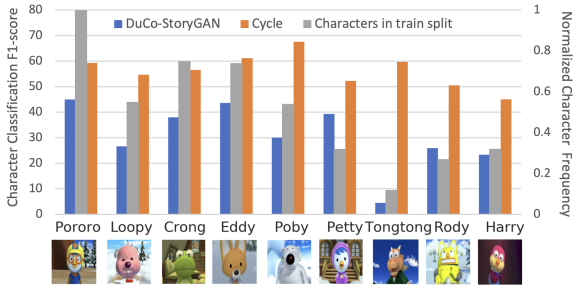

Figure 5: **Character classification F1-score on test split of Pororo-SV Dataset.**

rectional text-to-image generation tasks in a single model (Reed et al., 2016; Qiao et al., 2019) to improve the generative model performance.

Taking this into account, we propose a bi-directional image and text sequence generative model given sequence of other modality. We jointly train the image sequence generative model from a multi-sentence paragraph and multi-sentence paragraph generative model from an image sequence in unified framework.

For the generative model, we use autoregressive transformer-based model for both text-to-image and image-to-text generation task, where two-stage approach is applied as mentioned in Sec. 3.

$$\mathcal{L} = \mathcal{L}_{txt2img} + \mathcal{L}_{img2txt}, \tag{3}$$

$$\mathcal{L}_{txt2img} = \sum_{k=1}^{n} -\log P(z_k|t_1, ..., t_m, z_1, ..., z_{k-1}), \tag{4}$$

$$\mathcal{L}_{img2txt} = \sum_{k=1}^{n} -\log P(t_k|z_1, ..., z_n, t_1, ..., t_{k-1}). \tag{5}$$

3.3 Cyclic Pseudo-Text Generation

For image captioning task, there is a lot of avenue to describe input image with natural language sentence. Thus, many evaluation metric to measure model performance have been proposed (Vedantam et al., 2015; Post, 2018). Because there are different properties to evaluate quality of the generated caption, however, no single methodology has been proposed and many researchers have conducted human evaluation and combined it with automatic evaluation measures to evaluate the captioning model performance so far. Difficulty of evaluation is due to the possibility that one image can be expressed in various natural language sentences. For this reason, most image captioning datasets consist of multiple descriptive sentences achieved by multi-human annotation processing (Lin et al., 2014).

Nonetheless, the Pororo-SV dataset for story visualization task (Li et al., 2019) consist of one text-image pair in training and evaluation dataset. This can be detrimental from a model generalization point of view. To combat this, we propose to utilize the pseudo-text to train the text-to-image generation model. The purpose of this pseudo-text is to supply contextual information of the given image with train model.

Also, this formulation of cycle-consistency in text-image can be thought of as an online data-augmentation technique (Shah et al., 2019) where the model is trained on several generated pseudo-texts with one image and hence can be generalizable to the unseen source text during inference.

4 Experiments

4.1 Data and Evaluation Metrics

Dataset. We use Pororo-SV dataset proposed in (Li et al., 2019), which is a modified version of (Kim et al., 2017) for story visualization task. Each story sample consists of 5 image sequences and corresponding 5 descriptions. As mentioned in previous works (Maharana et al., 2021), there is a lot of data overlap between training and test samples in the original dataset split of Pororo-SV dataset (Li et al., 2019). To be more challenging, we follow the dataset split proposed in (Maharana et al., 2021), which contains 10191, 2334 and 2208 samples in training, validation and test splits, respectively. In this version, there is no data overlap between training and test split.

Evaluation Metrics [1] Due to the task complexity and its generative nature of the story visualization, evaluation is non-trivial. Evaluation method of generated image sequence needs to focus on the generated image quality, coherency between generated images and semantic matching generated image sequence with descriptions. Thus, we use diverse evaluation metrics to consider the complexity of the task following (Maharana et al., 2021). We also evaluate the generated caption quality with diverse automatic-evaluation metrics.

- **Fréchet Inception Distance (FID)**: Assessing the quality of generated image by calculating the

[1] To make a fair comparison with prior works, all pre-trained model for evaluating the performance are based on https://github.com/adymaharana/StoryViz.

Methods	FID↓	FSD↓	Char. F1↑	Frame Acc.↑	BLEU2/3↑
DuCo (Maharana et al., 2021)	91.96	171.36	36.13	13.03	3.39 / 1.40
Baseline (Transformer-based model)	66.51	40.34	48.38	18.38	4.34 / 1.77
+ Memory-Augmented Recurrent	65.89	36.81	57.53	27.65	4.90 / 2.01
+ Nucleus Sampling	56.04	33.27	**59.20**	**28.69**	5.18 / 2.18
+ Bi-directional	52.20	31.43	57.18	26.81	5.23 / 2.27
+ Cyclic Pseudo-Text (**C-SMART**)	**50.24**	**30.40**	58.11	28.06	**5.30 / 2.34**

Table 2: **Ablated results on test split of Pororo-SV Dataset.**

distance of the distribution between generated and real images used to train the generator.

- **Fréchet Story Distance (FSD)**: Assessing the coherency of the generated sequence of images by calculating the distance of the distribution between generated and real stories used to train the generator proposed in (Song et al., 2020).
- **Character Classification**: Assessing the presence of character in generated image sequence. Using pre-trained Inception-v3 with a multi-label classification loss to identify characters in the generated image. In particular, we report micro-averaged F-score of character classification and exact matching using frame accuracy as done in prior work (Maharana et al., 2021).
- **Video Captioning Accuracy**: Assessing the global semantic matching between generated image sequence and captions. We report the BLEU2/3 scores of captions predicted using generated images with pre-trained video captioner.
- **Language Quality**: Assessing the quality of generated captions using C-SMART. We use the automatic evaluation metrics composed of BLEU 2/3, METEOR, ROUGE-L, and CIDEr score. [2]

4.2 Implementation Details [3]

We use a recurrent GPT-based paragraph-to-image sequence generator having a memory layer for story visualization. With this, we iteratively conduct bi-directional generation of text-to-image and image-to-text. In the first stage of training, we train a discrete variational autoencoder with only Pororo-SV dataset, which compresses each input image into 16 $\times$ 16 grid of image tokens having 8192 possible values for each element. Then, we use a simple text tokenizer[4] having vocaburary size of 49,408. Finally, we use 128 text token length and totally 386 (128 + 16 $\times$ 16 + 2) input tokens with two special tokens (*i.e.*, start of sentence token and start of image token) (Fig. 2). In the second stage of training, we train the autoregressive multimodal transformer in a recurrent manner, *i.e.* C-SMART, to produce image sequence given multi-sentence paragraph. We set the hidden dimension size to 512, the number of transformer layer to 16, and the number of attention heads to 16. For positional encoding, we use embedding layer to learn each relative token position. Moreover, because we train the generative network in a bi-directional way, we use special segment embedding to implement it. By indicating source and target individually with segment embedding, we can make the model to produce target data from source data, *i.e.* text-to-image or image-to-text) And, we apply attention mask similar to (Ramesh et al., 2021). For memory module, we set the length of recurrent memory state 1. For cyclic pseudo-text approach, we set it to be implemented after half time of total training times. This is because at the beginning of the training, model can not produce proper pseudo-text using the model. This can be harmful to train text-to-image generation. Thus, we experimentally decide to implement it after half time shown in Fig. 8

5 Results

5.1 Quantitative Results

In Table 1, we summarize the performance comparison to prior works and C-SMART on Pororo-SV (Li et al., 2019) in test split. In all metrics (*i.e.*, **FID**, **FSD**, **Char.F1**, **Frame Acc.**, **BLEU2/3**) used for evaluating image quality, temporal coherency on sequence of generated images and global semantic consistency between generated images and descriptions, C-SMART outperforms prior existing works by a large margin. Particularly, C-SMART shows a significant gain of **FSD**, which measures

[2] We use the nlg-eval package (Sharma et al., 2017) to evaluate the generated caption quality.

[3] You can find more details about implementation details in our code. We will release code soon.

[4] https://github.com/openai/CLIP/blob/main/clip/simple_tokenizer.py

Figure 6: **Comparative Qualitative Results to Prior Arts.** GT refers to ground-truth. We compare our method (C-SMART) to prior arts including StoryGAN, CPC-SV and DuCo. Ours generates a semantically more plausible and temporally more coherent image sequence than the prior arts. Note that our C-SMART generates 128×128 whereas other methods generate 64×64, thus the clarity of the images is an additional benefit of our method.

Figure 7: **Examples of generated sequence of images with and without recurrent memory module.** Ours (C-SMART) generates semantically and visually plausible image sequences.

the temporal coherency in the story, over existing works. Along with total averaged **Char.F1** score in Tab. 1, Fig. 5 shows comparison of per-character classification F1-score on test split of Pororo-SV dataset and normalized character frequency in train split dataset. With our C-SMART, we obtain superior performance on various characters shown in Fig. 5 Particularly, when compared with (Maharana et al., 2021), we see up to 55% improvement for less frequent character, *i.e.* the character name of Tongtong, in test split of Pororo-SV dataset. In addition, considering the upper bound of **BLEU2/3** score measured using ground truth test dataset, *i.e.* oracle score: BLEU2/3: 5.54/2.34, C-SMART significantly outperform prior works and approximately is closet to the oracle score. Considering prior work (Maharana et al., 2021) explicitly use the pre-trained video captioner as learning signal, this result shows the superior performance of C-SMART.

Furthermore, to assess the contribution of various components added to C-SMART, we performed an ablation experiment with different configurations as shown in Tab. 2.

The baseline model consists of autoregressive transformer-based model only for single text-to-image generation. Although it lacks ability to globally understand sequential context because of limit of the architecture, the performance among various evaluated metrics is dramatically improved as compared to prior work (Maharana et al., 2021). We conjecture that this improvement can be attributed to the improved image quality by using transformer-based approach. To allow the baseline model to utilize the historical information and to be more temporally coherent between generated contents, we add recurrent path augmented with memory module as similar with (Lei et al., 2020b). We identify that the generative performance increase among all evaluation metrics and the increment of **FSD**, **Char.F1** and **Frame Acc.** is the highest among different components, *e.g.* nucleus sampling, bi-directional and cyclic pseudo-text. This improved performance can be atttributed to Next, we use the nucleus sampling strategy with the memory-augmented recurrent transformer. With the addition of this, we obtain more enhanced **FID** score, which means improved image quality. Next, we evaluate the addition of the bi-directional mechanism where

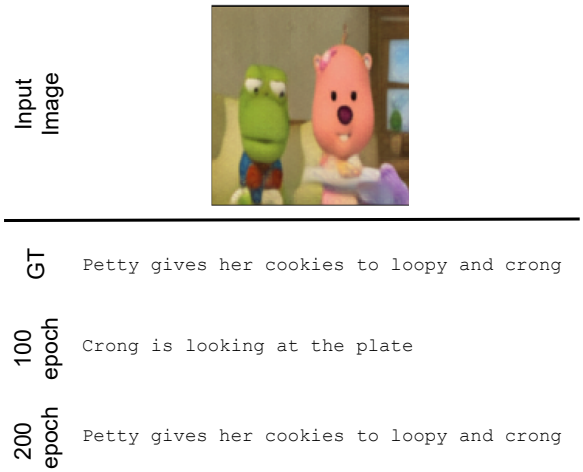

Figure 8: **Contextual information predicted from the model when increasing training time.**

generative process of text-to-image and image-to-text is conducted iteratively in a unified architecture. We observe the performance improvement except **Char.F1** and **Frame Acc.**. Lastly, we use cyclic pseudo-text approach where pseudo-text is produced by using autogressive transformer-based model to supply the generative model contextual information not described in ground-truth label. The use of cyclic pseudo-text approach makes the generative model to be more improved compared to prior version, and obtain increased performance.

5.2 Qualitative Results

Fig. 8 contains generated examples from the Pororo-SV dataset. The top row in images shows the ground truth sequence image, the three rows (2-4) contain prior works (Li et al., 2019; Song et al., 2020; Maharana et al., 2021) and the final row is the image generated by C-SMART. In this example, we demonstrate the superior visual quality and temporal cohrecy of our approch as compared prior works.

We empirically investigate the advantage of recurrent memory and summarize the results in Fig. 8. As shown in the examples, the proposed recurrent memory promotes to generate a semantically more plausible and temporally consistent image sequence (compare second rows to third rows).

5.3 Analysis of Linguistics

As the unified architecture for generating of text/image sequence given other modality input, we also have advantage of video description when using C-SMART framework. By using this, we apply it as pseudo-text generation and conduct cyclic approach on Image to Pseudo-Text to Image stream. In order to compare generated caption quality, we compare the generated caption results using C-SMART with MART (Lei et al., 2020b) which is only used for video caption task.

Pseudo-Text for contextual information. Fig. 8 shows various generated captions that change as training time increases using C-SMART. Looking at the results of captions inferred through C-SMART, it can be confirmed that although it does not generate proper captions at the beginning of training, it generates descriptive sentences to input image as training progresses, which is different from GT but it can be also thought of as another descriptive sentence, *i.e.* contextual information. However, when training procedure is completed (200 epoch), it can be shown that the generated captions are almost similar to the training data, indicating that it overfits to the training data. Thus, to properly utilize the contextual information generated by pseudo-text, we use the late activation strategy (Shah et al., 2019) by producing pseudo-text at later stages of training. In our case, we apply it at half of total training time, which is determined by empirically as shown in Fig. 8

6 Conclusion

We propose C-SMART as bi-directional generative framework for sequential target data (image/text) generation given sequential source data (text/image) for solving the task of story visualization in unified form.

Overall, our C-SMART significantly outperform prior works by large margin on the various evaluation metrics.

Extending our model to out-of-distribution dataset or in zero-shot setup would be an interesting future research avenue.

References

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

Xi Chen, Nikhil Mishra, Mostafa Rohaninejad, and Pieter Abbeel. 2018. Pixelsnail: An improved autoregressive generative model. In *International Conference on Machine Learning*, pages 864–872. PMLR.

Eldan Cohen and Christopher Beck. 2019. Empirical analysis of beam search performance degradation in

neural sequence models. In *International Conference on Machine Learning*, pages 1290–1299. PMLR.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. 2021. Cogview: Mastering text-to-image generation via transformers. *arXiv preprint arXiv:2105.13290*.

Lianli Gao, Daiyuan Chen, Jingkuan Song, Xing Xu, Dongxiang Zhang, and Heng Tao Shen. 2019. Perceptual pyramid adversarial networks for text-to-image synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8312–8319.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Yupan Huang, Hongwei Xue, Bei Liu, and Yutong Lu. 2021. Unifying multimodal transformer for bi-directional image and text generation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1138–1147.

Daphne Ippolito, Reno Kriz, Maria Kustikova, João Sedoc, and Chris Callison-Burch. 2019. Comparison of diverse decoding methods from conditional language models. *arXiv preprint arXiv:1906.06362*.

Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. 2017. Deepstory: Video story qa by deep embedded memory networks. *arXiv preprint arXiv:1707.00836*.

Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara Berg, and Mohit Bansal. 2020a. MART: Memory-augmented recurrent transformer for coherent video paragraph captioning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2603–2614, Online. Association for Computational Linguistics.

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

Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. 2019. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6329–6338.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Adyasha Maharana and Mohit Bansal. 2021. Integrating visuospatial, linguistic and commonsense structure into story visualization. *arXiv preprint arXiv:2110.10834*.

Adyasha Maharana, Darryl Hannan, and Mohit Bansal. 2021. Improving generation and evaluation of visual stories via semantic consistency. *arXiv preprint arXiv:2105.10026*.

Hyojin Park, Youngjoon Yoo, and Nojun Kwak. 2018. Mc-gan: Multi-conditional generative adversarial network for image synthesis. *arXiv preprint arXiv:1805.01123*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. *CoRR*, abs/1804.08771.

Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. 2019. Learn, imagine and create: Text-to-image generation from prior knowledge. *Advances in Neural Information Processing Systems*, 32:887–897.

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. *arXiv preprint arXiv:2103.00020*.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069. PMLR.

Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. 2019. Cycle-consistency for robust visual question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6642–6651.

Shikhar Sharma, Layla El Asri, Hannes Schulz, and Jeremie Zumer. 2017. Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *ArXiv*, abs/1706.09799.

Yun-Zhu Song, Zhi Rui Tam, Hung-Jen Chen, Huiao-Han Lu, and Hong-Han Shuai. 2020. Character-preserving coherent story visualization. In *European Conference on Computer Vision*, pages 18–33. Springer.

Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Yi-Fu Wu, Jaesik Yoon, and Sungjin Ahn. 2021. Generative video transformer: Can objects be the words? In *International Conference on Machine Learning*, pages 11307–11318. PMLR.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915.

Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2018. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962.

Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. 2018. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8748.

Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. 2019. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5810.

A Reproducing Prior Works

We reproduced the performance of prior works using author's implementations. [5]

[5] StoryGAN: https://github.com/yitong91/StoryGAN, CP-CSV: https://github.com/basiclab/CPCStoryVisualization-Pytorch, DuCo-StoryGAN: https://github.com/adymaharana/StoryViz