Generating a Temporally Coherent Visual Story with 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. To generate a semantically coherent image sequence, we propose an explicit memory controller which can augment the temporal coherence of images in the multi-modal autoregressive transformer, and call it **S**tory visualization by **M**ultimod**A**l **R**ecurrent **T**ransformers or SMART for short. Our method generates high resolution high quality images, outperforming prior works by a significant margin across multiple evaluation metrics on the PororoSV dataset.

1 Introduction

Story visualization is a challenging task of text-to-image generation. 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 input. The task lies at the intersection of natural language processing and computer vision. It is more challenging than the conventional text-to-image generation task owing to additional objectives such as understanding narrative in the text input, semantic relevance and temporal consistency, *e.g.*, foreground and background consistency, in the generated sequence of images. At the first glance, the story visualization task may seem similar to text-based video synthesis. Nevertheless, story visualization has a unique challenge as the image frames of a story are more disjoint than that of a video. In sum, the task of story visualization shares the difficulty

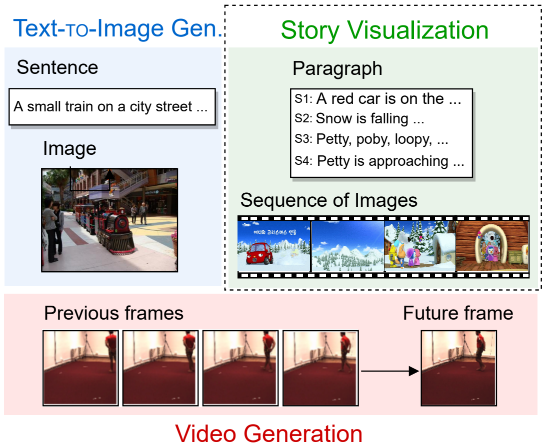

Figure 1: **Comparing visual generation tasks from texts.** Story visualization task aims to generate a sequence of images to describe a given story written in a natural language paragraph and is different from text-to-image or video generation.

of both text-to-image and text-to-video generation tasks as depicted in Fig. 1.

To generate a semantically relevant and temporally consistent sequence of images, we need to utilize both past and current scene narratives extracted from the sentence inputs. The recently proposed copy-transform mechanism (Maharana et al., 2021) based on attention-based semantic alignment (Tao Xu, 2018) has shown some promising results but has a large room for improvement.

Memory-Augmented Recurrent Transformer (MART) (Lei et al., 2020), a recent advancement in video-captioning task, presents an interesting research avenue in story visualization. It is based on a shared gated-memory module, similar to an RNN, which determines the importance of the preservation of historical feature information. The memory module is added between each layer of the recurrent transformer and helps in the generation of more coherent and diverse video captions while maintaining semantic relevance to video events.

Inspired by these insights, we propose to use a dynamic gated-memory module in a multimodal recurrent autoregressive transformer to model the correlation of the generated image with both past and current sentence inputs. The autoregressive transformer is a likelihood-based-model (Chen et al., 2020) and presents several advantages over traditional GAN-based generation modules with respect to mode-collapse, training instabilities, and lack of sample diversity (Adiga et al., 2018). Furthermore, a multimodal self-attention module preserves context over long-range text and image inputs for improved image resolution. With the added gated-memory module, we can expect the multimodal recurrent autoregressive transformer to generate images with a substantially higher degree of semantic relevance and temporal consistency, all on account of the sophisticated utilization of historical information.

We call our proposed model architecture SMART (**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 enhanced image quality and coherency between generations, as shown in Fig. 4.

We summarize our contributions as follows:

- We propose the first model using multimodal self-attention on long-range input of text and image in a recurrent manner for generating a temporally coherent image sequence given a paragraph.
- We explicitly generate sequences of images at a higher resolution with higher quality than ever before on a benchmark dataset.
- We outperform prior works by a large margin on the image quality and temporal coherence between generated images.

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., 2020) to encode video captions, DALL-E presents a generation framework based on joint autoregressive modeling of text and images.

3 Method

SMART generates a semantically relevant and temporally consistent sequence of images corresponding to an input multi-sentence story input. 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 generate an image sequence, 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.

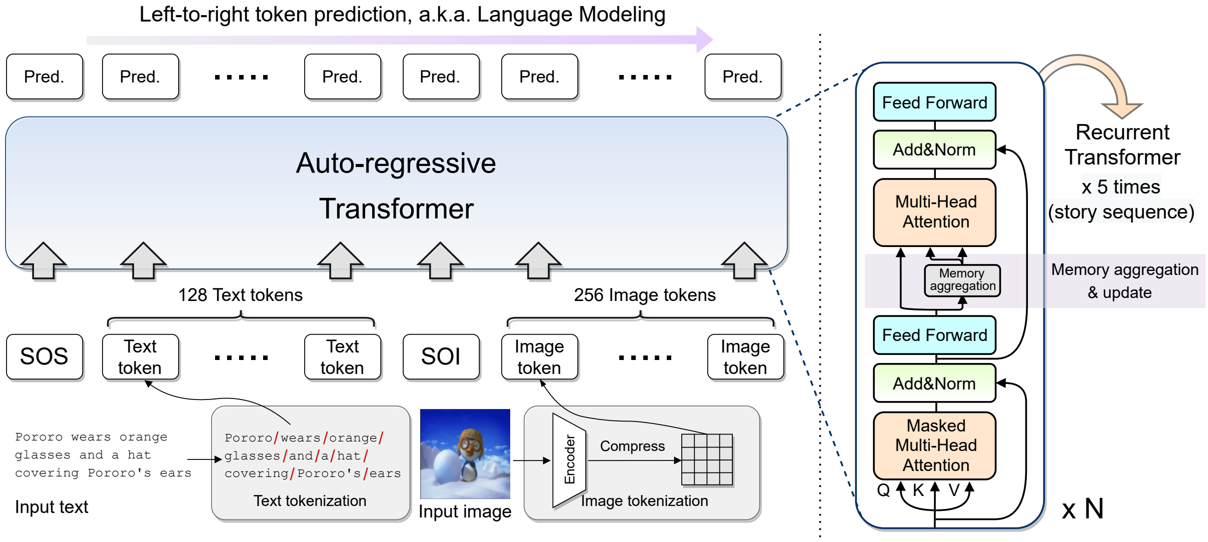

Figure 2: **Proposed Multimodal Recurrent Transformer for generating an image sequence given a multi-sentence paragraph. (Left)**: Illustration of the single text-to-image generation process. With auto-regressive transformer architecture, the training procedure is conducted using left-to-right token 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.

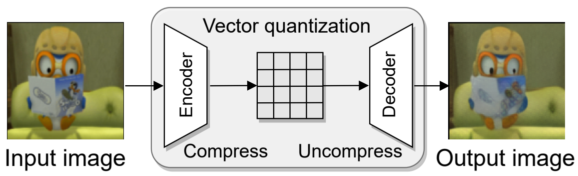

Figure 3: **Image tokenization using VQ-VAE.**

3.1 Image Tokenization

Image tokens are generated at the compression stage of training. Real images usually consist of millions of sub-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. 3, this framework is the autoencoder structure that learns a discretized latent encoding for input data x in the training procedure.

3.2 Generating an Image Seq. from Texts

An agent 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 autoregressive 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.

4 Experiments

Dataset. We use PororoSV 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. Following the task formulated in StoryGAN (Li et al., 2019), we use 13,000 training pairs and 2,334 testing pairs.

Methods	FID↓	FSD↓
StoryGAN (Li et al., 2019)	75.65	80.39
CPC-SV (Song et al., 2020)	68.75	79.86
DuCo (Maharana et al., 2021)	64.94	96.32
SMART (Ours)	**44.78**	**27.29**
SMART w/o Recurrent	50.62	31.43
SMART w/o Character cls.	48.25	30.21

Table 1: **Quantitative comparison.** ↓ indicates 'lower the better'.

Figure 4: **Examples of generated sequence of images using various methods.** Ours generates semantically and visually plausible image sequences.

Metrics. Evaluation method of generated image sequence needs to focus on the generated image quality and coherency between generated images. Following (Song et al., 2020), we use FID (Fréchet Inception Distance) and FSD (Fréchet Story Distance) as quantitative metrics to evaluate the methods. Please refer to (Song et al., 2020) for the details about them.

Implementation details. We use a recurrent GPT-based paragraph-to-image sequence generator having a memory layer for story visualization. In the first stage of training, we train a discrete variational autoencoder with only PororoSV dataset, which compresses each input image into 16×16 grid of image tokens having 8192 possible values for each element. Then, we use a simple text tokenizer[1] 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).

5 Results

5.1 Quantitative Analysis

In Table 1, we summarize the performance comparison to prior works and ablated components on PororoSV (Li et al., 2019) dataset. In both metrics (*i.e.*, **FID** and **FSD**) used for evaluating image quality and temporal coherency, SMART outperforms prior existing works by a large margin. Particularly, SMART shows a significant gain of **FSD**, which measures the temporal coherency in the story, over existing works.

Furthermore, to assess the contribution of recurrent architecture and character classification loss, we performed an ablation experiment with different configurations as shown in Table 1. Removing the recurrent framework from the model degrades quite a bit of performance, indicating that it is needed for both local (*e.g.*, **FID**) and global (*e.g.*, **FSD**) understanding of the story. Removing the character classification loss also hurts the model performance as shown in Table 1 The reason is that because the dataset domain is quite simple, the object information could guide for improving the performance. Thus, the generative model in which the component of character classification loss has been removed has deteriorated.

5.2 Qualitative Analysis

We empirically investigate the advantage of recurrent memory and summarize the results in Fig. 4. 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). We further compare our method to prior arts qualitatively in Appendix for the space sake.

6 Conclusion

We propose a novel architecture based on multimodal recurrent transformer for solving the task of story visualization. Extending our model to out-of-distribution datasets or in zero-shot setup would be an interesting future research avenue.

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

References

Sudarshan Adiga, Mohamed Adel Attia, Wei-Ting Chang, and Ravi Tandon. 2018. On the tradeoff between mode collapse and sample quality in generative adversarial networks. In *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1184–1188. IEEE.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. 2020. Generative pretraining from pixels. In *International Conference on Machine Learning*, pages 1691–1703. PMLR.

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.

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 L Berg, and Mohit Bansal. 2020. 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.

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*.

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*.

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.

Qiuyuan Huang Han Zhang Zhe Gan Xiaolei Huang Xiaodong He Tao Xu, Pengchuan Zhang. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks.

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.

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.

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 Additional Qualitative Results

We present additional qualitative results in the following figures.

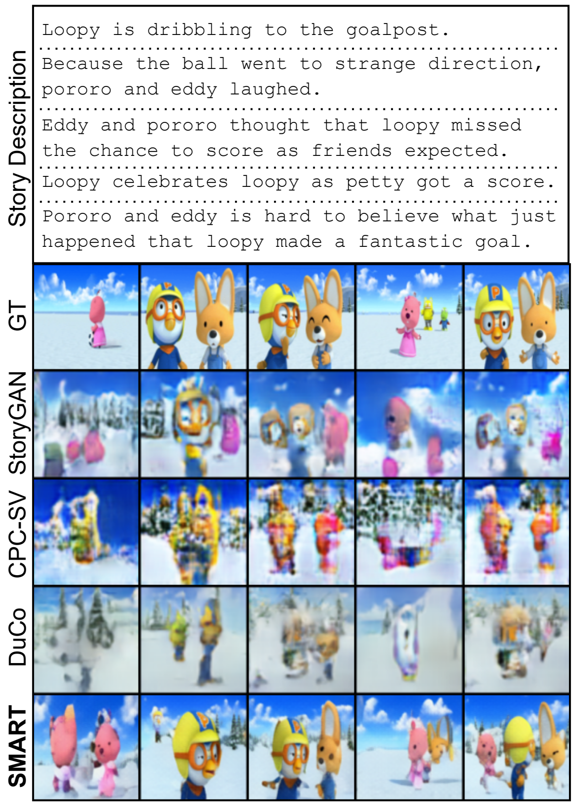

Figure 5: **Comparative Qualitative Results to Prior Arts.** GT refers to ground-truth. We compare our method (SMART) to prior arts including StoryGAN, CPC-SV and DuCo. Ours generates a semantically more plausible and temporally more coherent image sequence than the prior arts. Note that our SMART generates 128×128 whereas other methods generate 64×64, thus the clarity of the images is an additional benefit of our method.

From now, we skip the sentences.

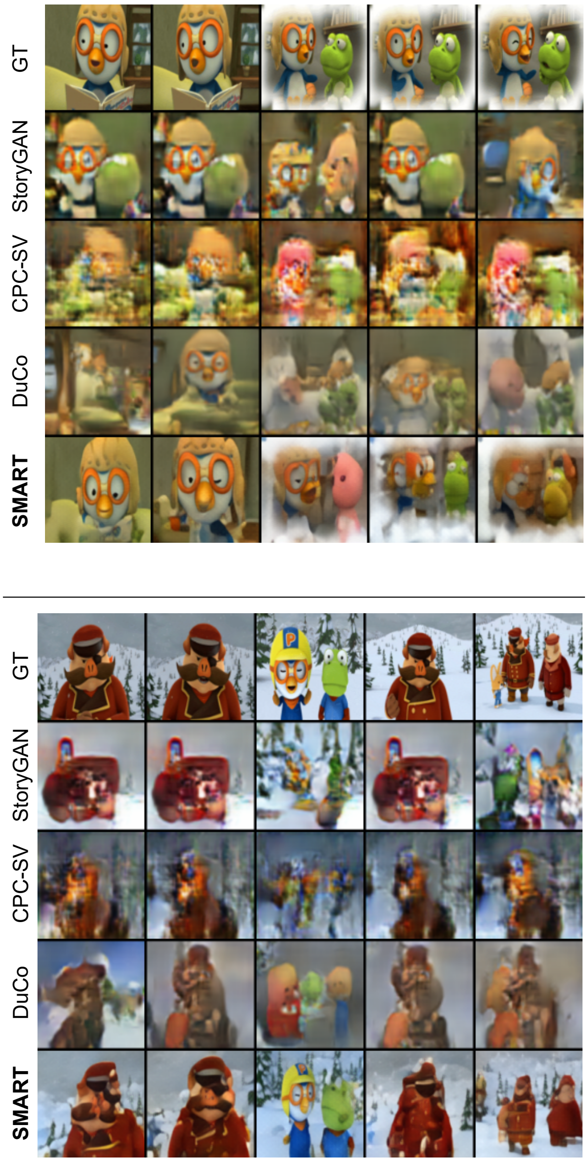

Figure 6: **More Comparative Qualitative Results to Prior Arts.**

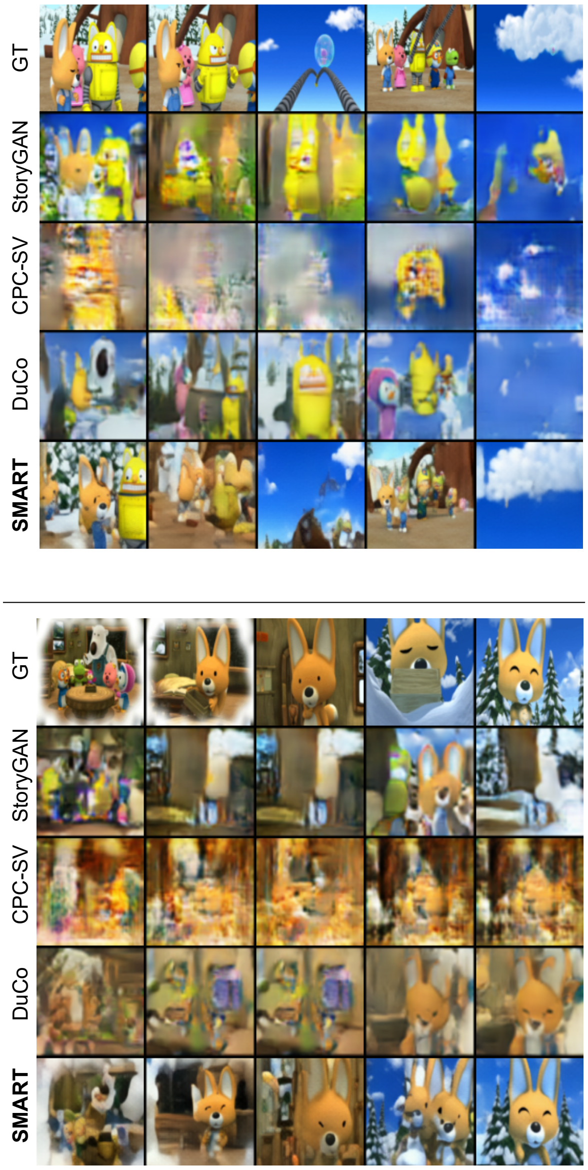

Figure 7: **More Comparative Qualitative Results to Prior Arts.**

Figure 8: **More Comparative Qualitative Results to Prior Arts.**