Generating Videos from Stories Using Conditional GAN

Takahiro Kozaki, Fumihiko Sakaue and Jun Sato
Nagoya Institute of Technology, Nagoya 466-8555, Japan
{*kozaki@cv., sakaue@, junsato@*}*nitech.ac.jp*

Keywords: Generative AI, Story, Multiple Sentences, Video, GAN, Captioning.

Abstract: In this paper, we propose a method for generating videos that represent stories described in multiple sentences. While research on generating images and videos from single sentences has been advancing, the generation of videos from long stories written in multiple sentences has not been achieved. In this paper, we use adversarial learning to train pairs of multi-sentence stories and videos to generate videos that replicate the flow of the stories. We also introduce caption loss for generating more contextually aligned videos from stories.

1 INTRODUCTION

In recent years, techniques for generating images from text information have progressed much, making it possible to generate a high-quality image from text information that describes the content of the image (Rombach et al., 2022; Ramesh et al., 2021). Furthermore, research on video generation from text information is also progressing (Ho et al., 2022), and it is becoming possible to generate short videos from text information.

However, generating a long video that represents a long story composed of multiple sentences has not yet been achieved. If long story videos could be automatically generated from long stories, it could be used in a variety of fields such as movie production. Thus, in this paper, we propose a method for generating long videos from multiple sentences in a story. This replicates the human ability to imagine and visualize the scenes described in the text while reading a novel.

For visualizing stories, Li et al. proposed StoryGAN (Li et al., 2019), which takes the entire text of a story as input and generates a sequence of keyframe images that represent the overall flow of the story. The StoryGAN successfully generates keyframe images corresponding to each sentence in the story. However, since only one keyframe image is generated from a single sentence, the resulting image sequence cannot represent the motions of characters and moving scenes described in the sentence.

Thus, in this research, we propose a method for generating long videos from input stories composed of multiple sentences as shown in Fig. 1. In contrast to StoryGAN, which generates a single image from each sentence, our method generates a short video from each sentence and then combines these videos to generate a long video representing the entire story. In our network, we introduce a caption loss that focuses on the appropriateness of regenerated stories from the generated videos, aiming to reproduce long videos from multiple sentences more accurately. Our network is trained and tested by using the Pororo dataset (Li et al., 2019) as in the case of StoryGAN.

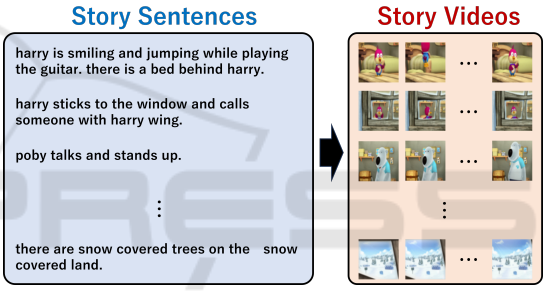


Figure 1: Generating long videos from long story sentences. The proposed method generates long videos with scene changes according to the story sentences. Our method generates multiple short videos from multiple sentences in a story, and then combines these videos to form a long video that represents the entire story.

2 RELATED WORK

In the field of image generation, Generative Adversarial Networks (GANs) have achieved significant success (Goodfellow et al., 2014; Zhu et al., 2017; Brock et al., 2019). Furthermore, the introduction of conditional GANs (cGANs), which allow the generation

Kozaki, T., Sakaue, F. and Sato, J.
Generating Videos from Stories Using Conditional GAN.
DOI: 10.5220/0012405300003660

In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 2: VISAPP*, pages 671-677
ISBN: 978-989-758-679-8; ISSN: 2184-4321

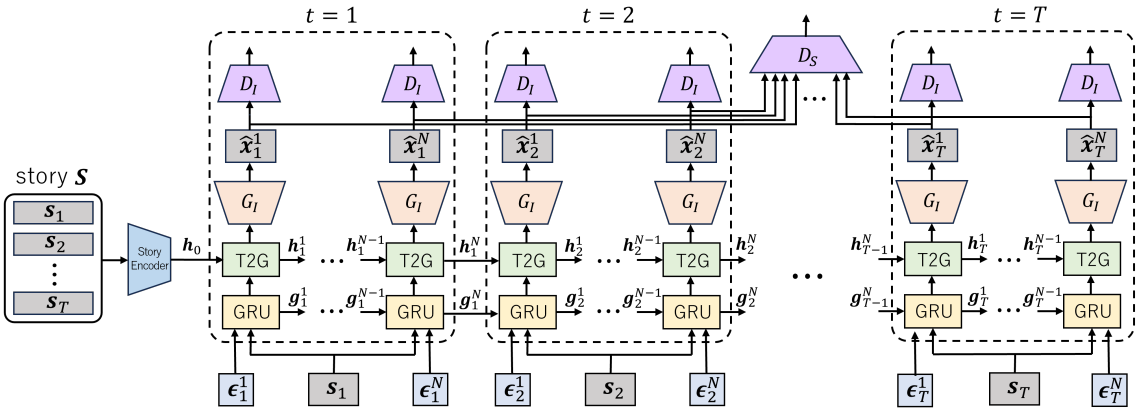


Figure 2: Network structure of the proposed method. Our story video generator consists of GRU, Text2Gist (T2G) and convolutional image generator (G_I). G_I generates N frames of continuous images from each sentence $\boldsymbol{s}_t$. As a result, $T \times N$ story video images are generated from T sentences. The image discriminator D_I compares generated image $\hat{\boldsymbol{x}}_t^n$ with sentence $\boldsymbol{s}_t$, while the story video discriminator D_S compares a set of $T \times N$ images with a set of T sentences. Initial vector $\boldsymbol{h}_0$ for T2G is obtained from input story $\boldsymbol{S}$ by using story encoder proposed in (Li et al., 2019).

of images that adhere to specific conditions, has further improved the capabilities of GANs (Isola et al., 2017). Moreover, recent research is also advancing in generating videos from text (Li et al., 2018). However, these GANs typically receive very short text inputs, often as short as a single sentence, and are not capable of generating videos from longer texts, such as stories composed of multiple sentences.

Research on text-based image generation has made significant progress in recent years (Reed et al., 2016; Ramesh et al., 2021; Rombach et al., 2022; Ramesh et al., 2022). One major trend in image generation from text is to combine well-trained text encoders (Radford et al., 2021) with image decoders (Jonathan Ho, 2020; Dosovitskiy et al., 2021), and generate an image representing multiple pieces of text. There has been significant technological progress in the field of image generation, and nowadays we can generate quite good high-resolution images which represent the content of multiple texts (Ramesh et al., 2021; Rombach et al., 2022).

Another important task in the field of text-based image generation is generating videos from text information (Li et al., 2018; Ho et al., 2022). In this field, research has been conducted to generate short videos representing a single sentence. However, unlike generating a single image, generating a video is a very difficult problem. For this reason, it has not yet been possible to generate a long video representing a long text such as a novel. Thus, generating long videos from long story sentences, such as novels is a very challenging problem.

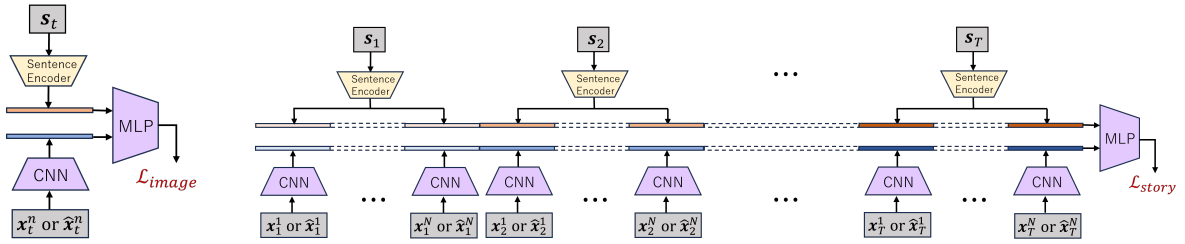
For visualizing long stories in multiple images, Li et al. proposed StoryGAN (Li et al., 2019) which generates multiple keyframe images from long story sentences. In StoryGAN, the generator performs adversarial learning with two discriminators, an image discriminator and a story discriminator, and generates a keyframe image from each sentence in the story. As a result, it generates a set of keyframe images that represent the long story. Furthermore, several improvements have been proposed based on this StoryGAN (Zeng et al., 2019; Li et al., 2020; Maharana et al., 2021). Li et al. (Li et al., 2020) introduced Weighted Activation Degree (WAD) for improving the discriminator in StoryGAN. Maharana et al. (Maharana et al., 2021) introduced a MART-based transformer (Lei et al., 2020) to model the correlation between the text and image in StoryGAN. They also used video captioning to improve the quality of generated images.

However, in these methods, only a single keyframe image is generated from each sentence, and it is not possible to generate a video from each sentence. Thus, generating long videos from long story sentences is still a challenging problem.

3 GENERATING VIDEOS FROM STORIES

3.1 Generating Videos from Multiple Sentences

In this research, we aim to generate long videos from stories composed of multiple sentences. Building upon StoryGAN (Li et al., 2019), we propose a method for generating contextually aligned videos



(a) D_I (b) D_S

Figure 3: Image discriminator (D_I) and story video discriminator (D_S). Story video discriminator identifies whether generated entire video aligns with story sentences by comparing video vector obtained from $T \times N$ images $[\hat{\boldsymbol{x}}_1^1, \cdots, \hat{\boldsymbol{x}}_T^N]$ through CNN and story vector obtained from T sentences $[\boldsymbol{s}_1, \cdots, \boldsymbol{s}_T]$ by using sentence encoder (Cer et al., 2018). Image discriminator evaluates each image and sentence in the same way.

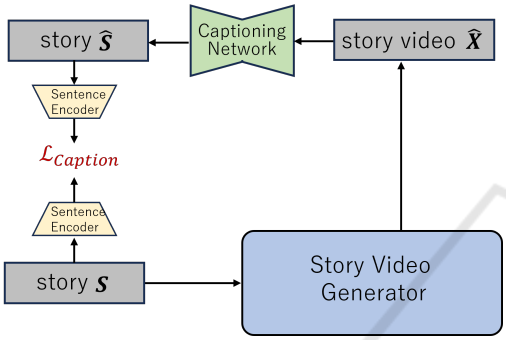


Figure 4: Input story sentences $\boldsymbol{S}$ and sentences $\hat{\boldsymbol{S}}$ obtained by captioning the generated story video $\hat{\boldsymbol{X}}$ are converted into sentence vectors by sentence encoder (Cer et al., 2018), and are used for computing caption loss $\mathcal{L}_{caption}$.

from stories with multiple sentences, where instead of keyframes, we generate videos for each sentence. In the following, we will refer to each individual sentence as a "sentence" and the collection of sentences as a "story."

In this study, we utilize data where sentences are paired with videos for GAN training. By inputting multiple sentences from this dataset into the generator, we train it to generate videos corresponding to each input sentence. By concatenating all the videos generated from these sentences, we create a long video representing the flow of the entire story.

The basic network structure of the proposed method is shown in Figure 2. First, the story encoder converts an input story, $\boldsymbol{S} = [\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_T]$, consisting of T sentences into a low-dimensional initial vector $\boldsymbol{h}_0$. This is achieved by using the story encoder proposed in StoryGAN (Li et al., 2019). Subsequently, the generator generates T sequences of N frames of videos $\hat{\boldsymbol{x}}_t = [\hat{\boldsymbol{x}}_t^1, \hat{\boldsymbol{x}}_t^2, \cdots, \hat{\boldsymbol{x}}_t^N]$ $(t = 1, \cdots, T)$ from this initial vector $\boldsymbol{h}_0$, each sentence $\boldsymbol{s}_t$ and noise ε. Then, the generated T videos are concatenated in sequence to form the final video $\hat{\boldsymbol{X}} = [\hat{\boldsymbol{x}}_1, \hat{\boldsymbol{x}}_2, \cdots, \hat{\boldsymbol{x}}_T]$.

This generator was built by extending the generator proposed in StoryGAN (Li et al., 2019). In StoryGAN, the generator consists of a convolutional image generator, GRU and Text2Gist, and generates a single image $\hat{\boldsymbol{x}}_t$ from a single sentence $\boldsymbol{s}_t$. In this research, we also use a convolutional image generator, GRU and Text2Gist, but we generate N frames of continuous images $[\hat{\boldsymbol{x}}^1, \cdots, \hat{\boldsymbol{x}}^N]$ from each sentence $\boldsymbol{s}_t$. This is achieved by iteratively using a pair of GRU and Text2Gist (T2G) N times inputting the same sentence $\boldsymbol{s}_t$ as shown in Figure 2. As a result, we generate a total of $T \times N$ images from T sentences.

On the other hand, the discriminator receives generated videos $\hat{\boldsymbol{x}}_t$ $(t = 1, \cdots, T)$ or ground truth videos $\boldsymbol{x}_t$ $(t = 1, \cdots, T)$ corresponding to the sentences $\boldsymbol{s}_t$ $(t = 1, \cdots, T)$ in the story $\boldsymbol{S}$, and evaluate the authenticity of the generated videos. We use two types of discriminator, that is image discriminator and story video discriminator. The image discriminator identifies whether each frame image aligns with the corresponding sentence, while the story video discriminator identifies whether the entire video aligns with the story. The image discriminator D_I compares an image vector obtained from a generated image $\hat{\boldsymbol{x}}_t^n$ through CNN and a sentence vector obtained from a sentence $\boldsymbol{s}_t$ by using a sentence encoder (Cer et al., 2018), while the story video discriminator D_S compares a video vector obtained from $T \times N$ images through CNN and a story vector obtained from T sentences by using the sentence encoder as shown in Figure 3. The comparison of the image vector and the sentence vector or the video vector and the story vector is conducted by MLP with two layers. These two types of discriminators enhance the accuracy of both the overall consistency of the video and the alignment with the original sentences, improving the overall quality of the generated story videos.

The story video generator G is trained by using the image discriminator D_I and the story video discriminator D_S as follows:

$$G^* = \arg\min_G \max_{D_I, D_S} \alpha \mathcal{L}_{image} + \beta \mathcal{L}_{story} + \mathcal{L}_{KL} \quad (1)$$

Figure 5: Some of the characters in Pororo dataset.

where, $\mathcal{L}_{image}$ represents the loss associated with the image discriminator D_I, while $\mathcal{L}_{story}$ represents the loss associated with the story video discriminator D_S as follows:

$$\begin{aligned}\mathcal{L}_{image} &= \sum_{t=1}^{T}\sum_{n=1}^{N}\log D_I(\boldsymbol{x}_t^n, \boldsymbol{s}_t, \boldsymbol{h}_0) \\ &+ \log(1 - D_I(G(\varepsilon_t^n, \boldsymbol{s}_t), \boldsymbol{s}_t, \boldsymbol{h}_0)) \quad (2)\\ \mathcal{L}_{story} &= \log D_S(\boldsymbol{X}, \boldsymbol{S}) \\ &+ \log(1 - D_S([[G(\varepsilon_t^n, \boldsymbol{s}_t)]_{n=1}^{N}]_{t=1}^{T}, \boldsymbol{S})) \quad (3)\end{aligned}$$

The story video discriminator D_S discriminates the ground truth story video $\boldsymbol{X}$ and the story video $\hat{\boldsymbol{X}} = [[G(\varepsilon_t^n, \boldsymbol{s}_t)]_{n=1}^{N}]_{t=1}^{T}$ generated by the generator G from the noise ε_t^n and the story $\boldsymbol{s}_t$.

Following StoryGAN (Li et al., 2019), we also use the following KL divergence $\mathcal{L}_{KL}$ between the standard Gaussian distribution and the learned distribution, which helps prevent mode collapse in the generation of the initial vector.

$$\mathcal{L}_{KL} = KL(\mathcal{N}(\mu(\boldsymbol{S}), diag(\sigma^2(\boldsymbol{S})))\|\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})) \quad (4)$$

where, $\mu(\boldsymbol{S})$ and $\sigma^2(\boldsymbol{S})$ are the mean and the variance of initial vectors derived from input story $\boldsymbol{S}$.

3.2 Improving Accuracy by Caption Loss

Next, we will consider improving the accuracy of video generation using video captioning. It is not easy to quantitatively evaluate the appropriateness of videos generated from text. One way to perform such difficult evaluations is to input the generated video into a well-trained captioning network to generate sentences that represent the video and see how close the generated sentences are to the original input sentences.

Thus, in this research, we perform captioning on the video $\hat{\boldsymbol{X}}$ generated by the story video generator using a trained video captioning network (Vladimir Iashin, 2020), and compute the caption loss $\mathcal{L}_{caption}$ by comparing the generated story sentences $\hat{\boldsymbol{S}}$ with the original input story sentences $\boldsymbol{S}$ as shown in Fig. 4. The loss obtained in this way is a measure of how well the generated video represents the input sentence. Therefore, by adding this loss to the network training, we can further improve the accuracy of the generator.

Table 1: Cosine similarity between input sentences and sentences obtained from generated videos.

	proposed	w/o caption loss
cos similarity (↑)	0.211	0.154

The similarity between the input story sentences $\boldsymbol{S}$ and the sentences $\hat{\boldsymbol{S}}$ obtained by captioning the generated story video is evaluated by using the cosine similarity between sentence vectors, $\boldsymbol{v}$ and $\hat{\boldsymbol{v}}$, derived from $\boldsymbol{S}$ and $\hat{\boldsymbol{S}}$ by using the sentence encoder (Cer et al., 2018). Therefore, the caption loss $\mathcal{L}_{caption}$ is defined as follows:

$$\mathcal{L}_{caption} = 1 - \cos(\boldsymbol{v}, \hat{\boldsymbol{v}}) \quad (5)$$

By adding this $\mathcal{L}_{caption}$ to the network training as follows, the video generation accuracy of the generator can be further improved:

$$G^* = \arg\min_{G}\max_{D_I, D_S} \alpha\mathcal{L}_{image} + \beta\mathcal{L}_{story} + \gamma\mathcal{L}_{caption} + \mathcal{L}_{KL} \quad (6)$$

4 DATASET

Since the proposed method generates videos from stories consisting of multiple sentences, a dataset consisting of paired stories and videos is required. Thus, in this research, we construct a dataset based on the Pororo dataset (Kim et al., 2017), which consists of short video clips and their description texts.

The Pororo dataset is an anime video dataset with Pororo as the main character of the stories. Fig. 5 shows some of the characters in the story. The anime video is divided into short video clips every few seconds, and each short video clip is annotated with a sentence. Therefore, the Pororo dataset consists of pairs of short video clip and its description text. The Pororo dataset consists of a series of episodes, so the stories are connected in continuous data.

The story GAN and other related research use Pororo-SV dataset (Li et al., 2019) which is created by extracting a single image randomly from each short video clip, so that the dataset consists of pairs of an image and a sentence. On the other hand, in this research, we need to train our network to generate N images from each sentence. Thus, we extracted N consecutive images from each video clip at equal intervals and created pairs of N images and a sentence. Since the stories are connected in continuous data, we cut out T consecutive pairs of an image sequence and a sentence as one set of story data. Therefore, one set of data is composed of $T \times N$ images and T sentences. In this research, we set $T = 5$ and $N = 10$, and generated 2995 sets of data. Each story sentence in this

s_1 : harry is smiling and jumping while playing the guitar. there is a bed behind harry.
s_2 : harry sticks to the window and calls someone with harry wing.
s_3 : poby talks and stands up.
s_4 : poby comes to the window.
s_5 : there are snow covered trees on the snow covered land.

(a) story 1

s_1 : eddy decides to fix the cloning machine now.
s_2 : eddy opens the door and checks it.
s_3 : eddy puts the sandwich into the machine.
s_4 : eddy gets the two sandwiches.
s_5 : eddy eats one of them.

(b) story 2

s_1 : pororo thinks that eddy is doing something fun.
s_2 : crong agrees and they go to eddy.
s_3 : pororo and crong are happy because they succeeded at something.
s_4 : eddy is angry at pororo. eddy asks what pororo is doing.
s_5 : pororo answers to eddy that pororo is having fun.

(c) story 3

Figure 6: Generated video images. Proposed method can generate a series of short videos from multiple sentences, while StoryGAN can only generate a single keyframe image from each sentence and cannot visualize the motion of the characters and scenes described in each sentence.

dataset was encoded into a fixed-length sentence vector using the Universal Sentence Encoder (Cer et al., 2018).

5 EXPERIMENTS

We next show the results of some experiments on generating story videos from story sentences. We trained our network for 120 epochs using 2500 sets of training data described in section 4. The trained network was tested by using 495 sets of test data.

We first show video images generated from test story sentences. Fig. 6 (a), (b) and (c) show three different story sentences and video images generated from our method. For comparison, we also show the results from StoryGAN (Li et al., 2019) and our method trained without using caption loss. As shown in this figure, the proposed method can generate a series of short videos from multiple sentences, while StoryGAN can only generate a single keyframe image from each sentence and cannot visualize the motion of the characters and scenes described in the sentence. We can also see that the characters mentioned in the sentences are properly generated in the videos obtained from the proposed method. Comparing methods with and without the caption loss, the proposed method using the caption loss can generate videos with larger changes, indicating that the generated videos have richer expressions.

To evaluate the accuracy of the generated videos quantitatively, we captioned the generated video by using the captioning network (Vladimir Iashin, 2020) trained by using the Pororo dataset, and computed the cosine similarity with the input sentence. Table 1 shows the cosine similarity between the input sentences and the sentences obtained from the generated videos. For comparison, we also show the cosine similarity when we do not use the caption loss in our method. As shown in Table 1, the caption loss in our method can improve the quality of the generated videos.

6 CONCLUSION

In this paper, we proposed a method for generating videos that represent stories described in multiple sentences.

The existing methods of text-to-video can only generate short videos from texts, and generating long story videos from long story sentences is a challenging problem. We in this paper extended the StoryGAN and showed that we can generate story videos from story sentences. Furthermore, we showed that by using the caption loss, we can further improve the accuracy of generated story videos.

Our method replicated the human ability to imagine and visualize the scenes described in the sentence while reading a novel. On the other hand, a large amount of experience is required to imagine rich scenes from story sentences, so using the foundation model may improve its ability to perform this task.

REFERENCES

Brock, A., Donahue, J., and Simonyan, K. (2019). Large scale gan training for high fidelity natural image synthesis. In *Proc. ICLR*.

Cer, D., Yang, Y., yi Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder. In *arXiv:1803.11175*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc, ICLR*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Proc. Advances in neural information processing systems*, pages 2672–2680.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. (2022). Video diffusion models. In *arXiv:2204.03458*.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134.

Jonathan Ho, Ajay Jain, P. A. (2020). Denoising diffusion probabilistic models. In *Proc. Conference on Neural Information Processing Systems*.

Kim, K., Heo, M., Choi, S., and Zhang, B. (2017). Deepstory: video story qa by deep embedded memory networks. In *Proc. of International Joint Conference on Artificial Intelligence*, page 2016–2022.

Lei, J., Wang, L., Shen, Y., Yu, D., Berg, T. L., and Bansal, M. (2020). Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning. In *arXiv:2005.05402*.

Li, C., Kong, L., and Zhou, Z. (2020). Improved-storygan for sequential images visualization. *Journal of Visual Communication and Image Representation*, 73(102956).

Li, Y., Gan, Z., Shen, Y., Liu, J., Cheng, Y., Wu, Y., Carin, L., Carlson, D., and Gao, J. (2019). Storygan: A sequential conditional gan for story visualization. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 6329–6338.

Li, Y., Min, M. R., Shen, D., Carlson, D., and Carin, L. (2018). Video generation from text. In *Proc. AAAI Conference on Artificial Intelligence*.

Maharana, A., Hannan, D., and Bansal, M. (2021). Improving generation and evaluation of visual stories via semantic consistency. In *Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 2427–2442.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proc. ICML*.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. In *arXiv:2204.06125*.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. (2021). Zero-shot text-to-image generation. In *arXiv:2102.12092*.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H. (2016). Generative adversarial text to image synthesis. In *Proc. ICML*.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, page 10684–10695.

Vladimir Iashin, E. R. (2020). A better use of audio-visual cues: Dense video captioning with bi-modal transformer. In *Proc. British Machine Vision Conference*.

Zeng, G., Li, Z., and Zhang, Y. (2019). Pororogan: An improved story visualization model on pororo-sv dataset. In *Proc. International Conference on Computer Science and Artificial Intelligence*, page 155–159.

Zhu, J., Park, T., Isola, P., and Efros, A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. International Conference on Computer Vision*, pages 2223–2232.