A. Network Configuration

This section gives the network structure used in StoryGAN. In the following, 'CONV' means the 2D convolutional layer, which is configured by output channel number 'C', kernel size 'K', step size 'S' and padding size 'P'. 'LINEAR' is fully connected layer, with input and output dimensions given in the parenthesis. Note that the Filter Network is contained in the Text2Gist cell, which transforms $\boldsymbol{i}_t$ to a filter. This is introduced in detail in section 3.2.

Table 5: Network Structure used in StoryGAN. * This layer combines the conditional input and the encoded images.

Layer	Story Encoder
1	LINEAR-(128 × T, 128), BN, RELU
Layer	Context Encoder
1	LINEAR-(NOISEDIM + TEXTDIM, 128), BN, RELU
2	GRU-(128, 128)
3	Text2Gist-(128, 128)
Layer	Filter Network
1	LINEAR-(128, 1024), BN, TANH
2	RESHAPE(16, 1, 1, 64)
Layer	Image Generator
1	CONV-(C512, K3, S1, P1), BN, RELU
2	UPSAMPLE-(2,2)
3	CONV-(C256, K3, S1, P1), BN, RELU
4	UPSAMPLE-(2,2)
5	CONV-(C128, K3, S1, P1), BN, RELU
6	UPSAMPLE-(2,2)
7	CONV-(C64, K3, S1, P1), BN, RELU
8	UPSAMPLE-(2,2)
9	CONV-(C3, K3, S1, P1), BN, TANH
Layer	Image Discriminator
1	CONV-(C64, K4, S2, P1), BN, LEAKY RELU
2	CONV-(C128, K4, S2, P1), BN, LEAKY RELU
3	CONV-(C256, K4, S2, P1), BN, LEAKY RELU
4	CONV-(C512, K4, S2, P1),BN, LEAKY RELU
5*	CONV-(C512, K3, S1, P1), BN, LEAKY RELU
6	CONV-(C1, K4, S4, P0), SIGMOID
Layer	Story Discriminator (Image Encoder)
1	CONV-(C64, K4, S2, P1), BN, LEAKY RELU
2	CONV-(C128, K4, S2, P1), BN, LEAKY RELU
3	CONV-(C256, K4, S2, P1), BN, LEAKY RELU
4	CONV-(C512, K4, S2, P1),BN, LEAKY RELU
5	CONV-(C32, K4, S2, P1),BN, CONCAT
6	RESHAPE-(1, 32 × 4 × T)
Layer	Story Discriminator (Text Encoder)
1	LINEAR-(128 × T, 32 × 4 × T), BN

B. More Examples of CLEVR-SV Dataset

Here we perform the test by using the same attributes of the first object. We test if the models can keep the first object consistent through the following generations. Figure 8 compares the results from different models. Figure 9 gives more samples using StoryGAN.

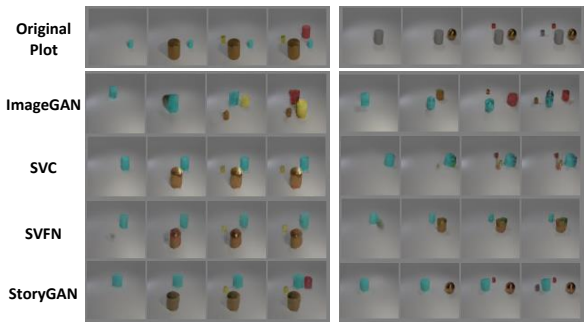

Figure 8: Method comparison on a task where the original story description is changed in the first sentence. Specifically, the first sentence is now "Large, Rubber, Cyan, Cylinder, at (-0.46, -0.36)." Each column corresponds to one layout of the following three objects. The first row is the original image that will be modified. Note that only StoryGAN keeps the story consistency among the compared methods.

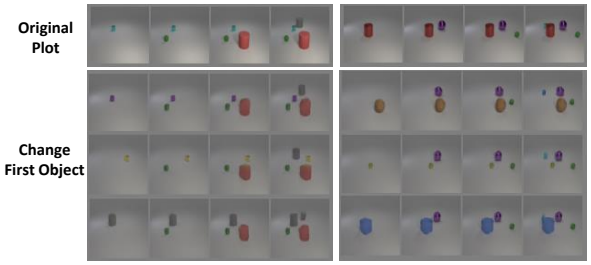

Figure 9: An additional example using the same idea as Figure 8. The top row gives two initial setups. The next three rows correspond to StoryGAN generations with different first sentences. For the left column, the attributes of the first object are:'Small, Metal, Cyan, Cylinder, at (-2.00, 0.02)' (original), 'Small, Metal, Purple, Cylinder, at (-1.15, -0.26)', 'Small, Metal, Yellow, Cylinder, at (0.35, 2.00)' and 'Large, Rubber, Gray, Cylinder, at (-1.77, -0.07)', respectively. For the right column, the attributes of the first object are: 'Large, Metal, Red, Sphere, at (0.52, 0.56)' (original), 'Large, Rubber, Brown, Sphere, at (-1.54, 0.85) ', 'Small, Metal, Yellow, Cylinder, at (-0.85, 2.29) ' and 'Large, Rubber, Blue, Cube, at (0.15, -0.19) ', respectively. Again, we omit the attribute input of the second, third and forth objects to save the space. Note that regardless of the initial description, StoryGAN effectively captures the story consistency.

C. Significance Test on Pororo-SV Dataset

We perform pairwise t-test on the human evaluated ranking test. As we can see, StoryGAN is statistically significant over other baseline models.

Table 6: p-value on the human evaluated ranking test.

Method	ImageGAN	SVC	SVFN	StoryGAN
ImageGAN	1.0	5e-13	0.04	1e-40
SVC	5e-13	1.0	1e-8	4e-14
SVFN	0.04	1e-8	1.0	3e-36
StoryGAN	1e-40	4e-14	3e-36	1.0

D. Characters Photo and More Examples of Pororo-SV Dataset

For the classification accuracy compared in Table 2, nine characters are selected: 'Pororo', 'Crong', 'Eddy', 'Poby', 'Loopy', 'Petty', 'Harry', 'Rody' and 'Tongtong'. Profile pictures of them are given in Figure 10.

Figure 10: Main character names and corresponding photos from the dataset.

Then, more generated samples on Pororo-SV dataset are given in Figure 11.

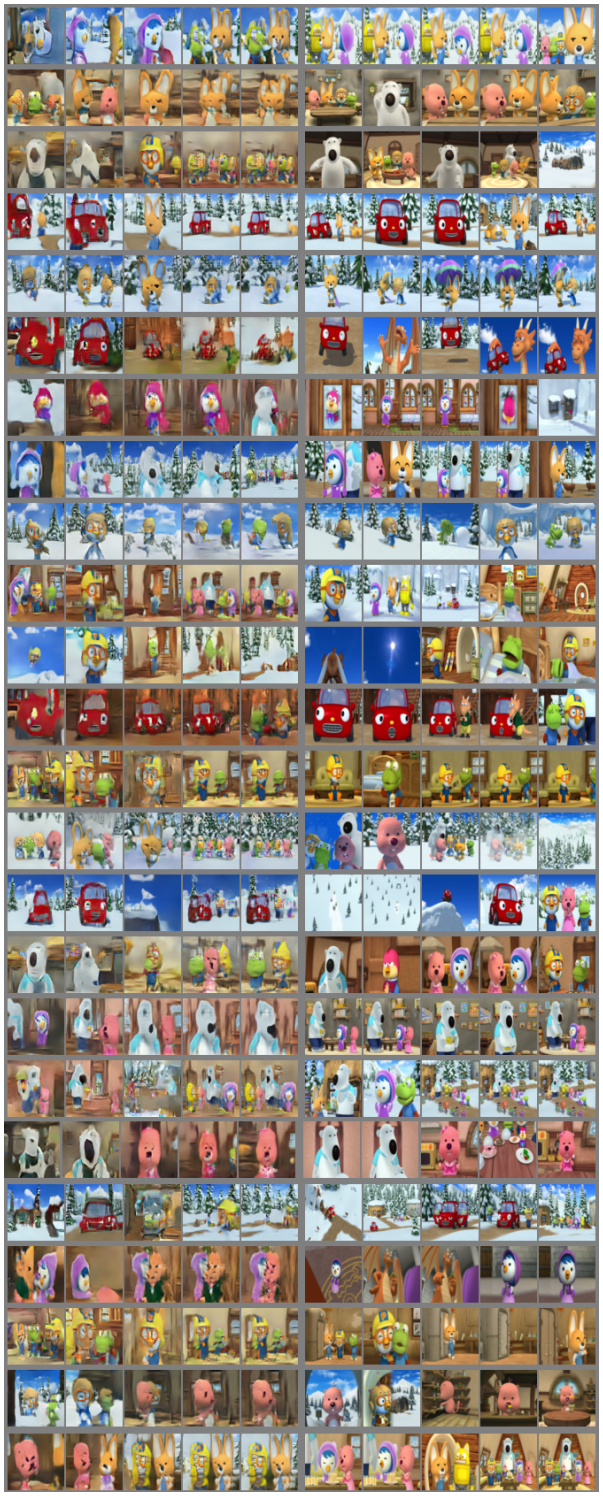

Figure 11: More samples on Pororo-SV test set. For simplicity, we give the ground truth story images instead of the raw story text. The left five columns are generated images. The right five columns are ground truth. Note that there is no need for the generation to exactly match the ground truth. Those samples with similar (but not exactly the same) images are caused by repeated input sentences.