StoryGAN: A Sequential Conditional GAN for Story Visualization

Zhe Gan

4/1/2019

New Task

Image Generation

"A small yellow bird with a black crown and beak."

Image Editing

"The flower has purple petals with a white stamen"

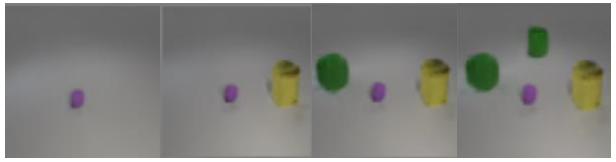
Previous Work

Story Visualization

"Pororo and Crong fishing together. Crong is looking at the bucket. Pororo has a fish on his fishing rod."

Interactive Image Editing

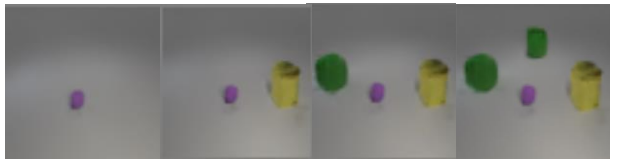
"Small purple rubber sphere, position is 1.4, -0.7.
"Large yellow metallic cylinder, position is 2.1, 2.6…"



Our Contribution

New Datasets

- CLEVR-SV Dataset (interactive image editing)
  - ***Original:*** Image QA task (100K images with object descriptions)
  - ***Ours:*** sequence of similar images with incremental complexity



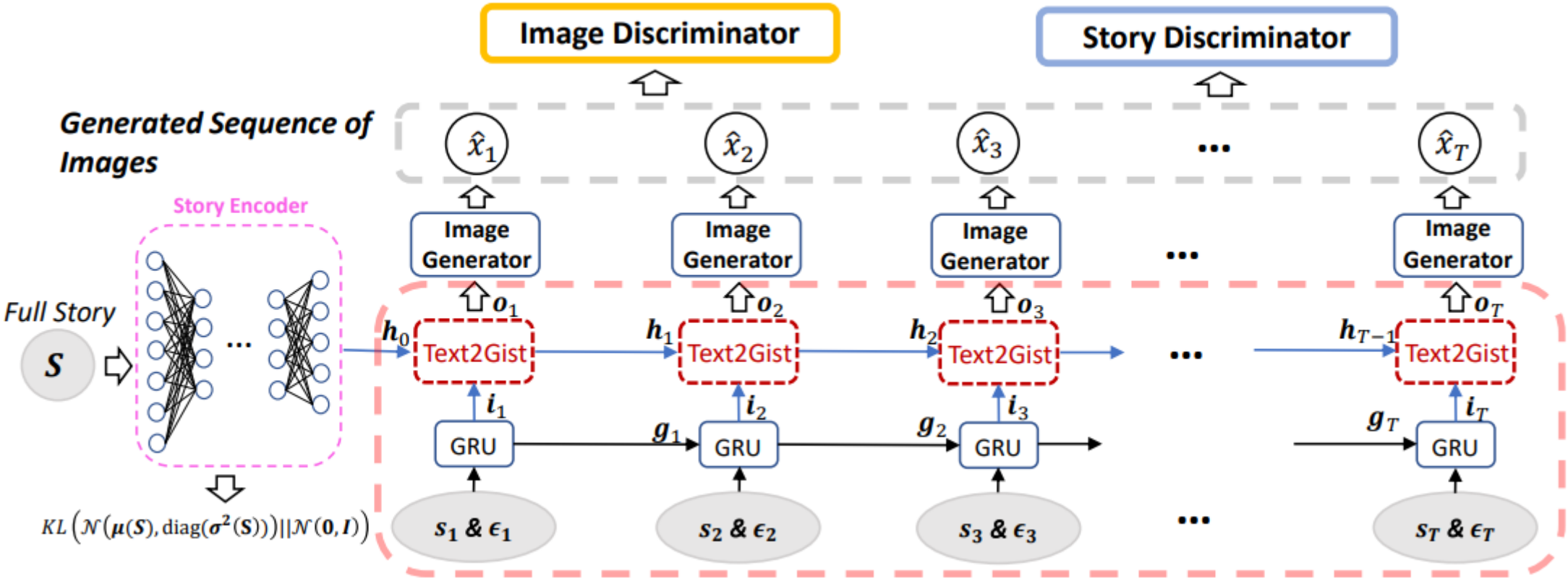
"Small purple rubber sphere, position is 1.4, -0.7.
"Large yellow metallic cylinder, position is 2.1, 2.6…"

- Pororo-SV Dataset (story visualization)
  - ***Original:*** Video QA task (16K video-description pairs)
  - ***Ours:*** sequence of frames from five consecutive video clips to form a story, paired with sequence of textual descriptions

"The woods are covered with snow. The sky is blue and clear. Pororo went to Loopy's house and saw Crong..."

New Model: StoryGAN

- Challenges
  - How to model *global consistency* across images, compared with image generation
  - How to model *sharp change of scenes* in a story, compared with video generation



New Model: StoryGAN

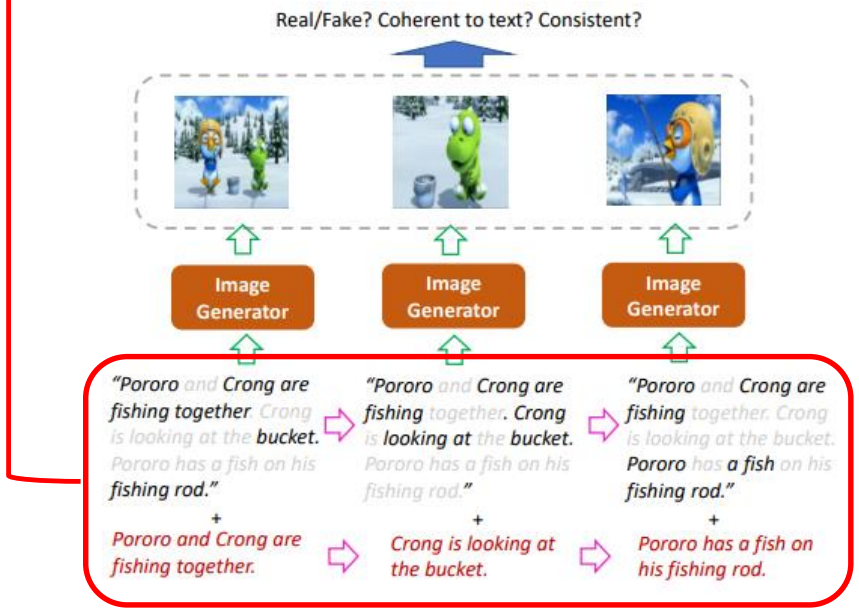
Text2Gist: combining both global and local context information

$$z_t = \sigma_z\left(W_z i_t + U_t h_{t-1} + b_z\right), \quad (4)$$

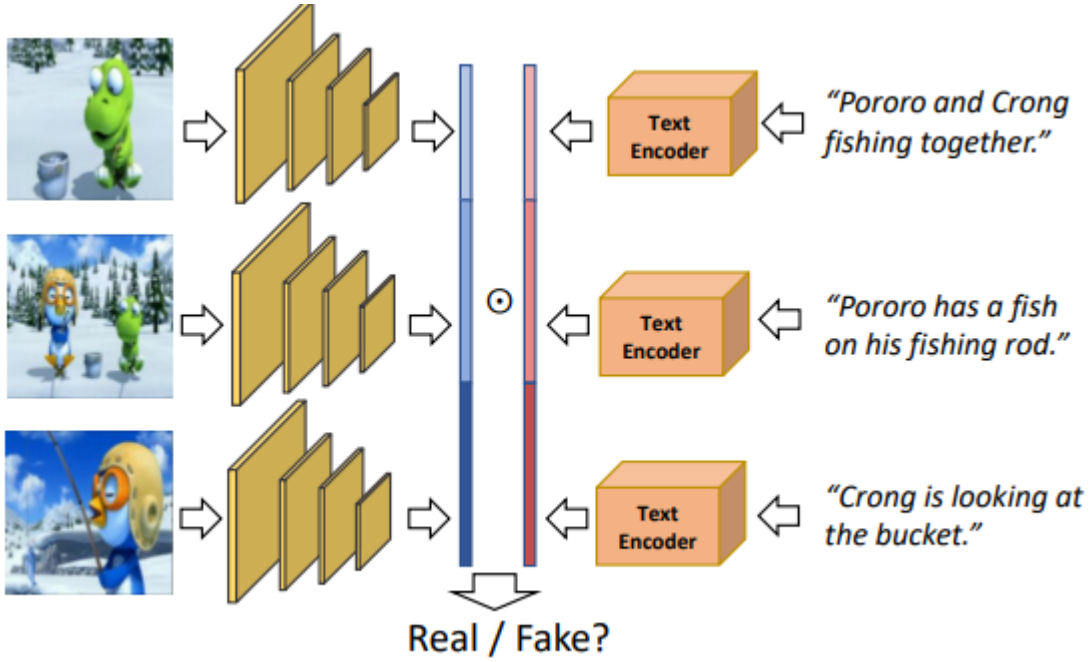
$$r_t = \sigma_r\left(W_r i_t + U_r h_{t-1} + b_r\right), \quad (5)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \sigma_h\left(W_h i_t + U_h(r_t \odot h_{t-1}) + b_h\right), \quad (6)$$

$$o_t = \text{Filter}(i_t) * h_t, \quad (7)$$

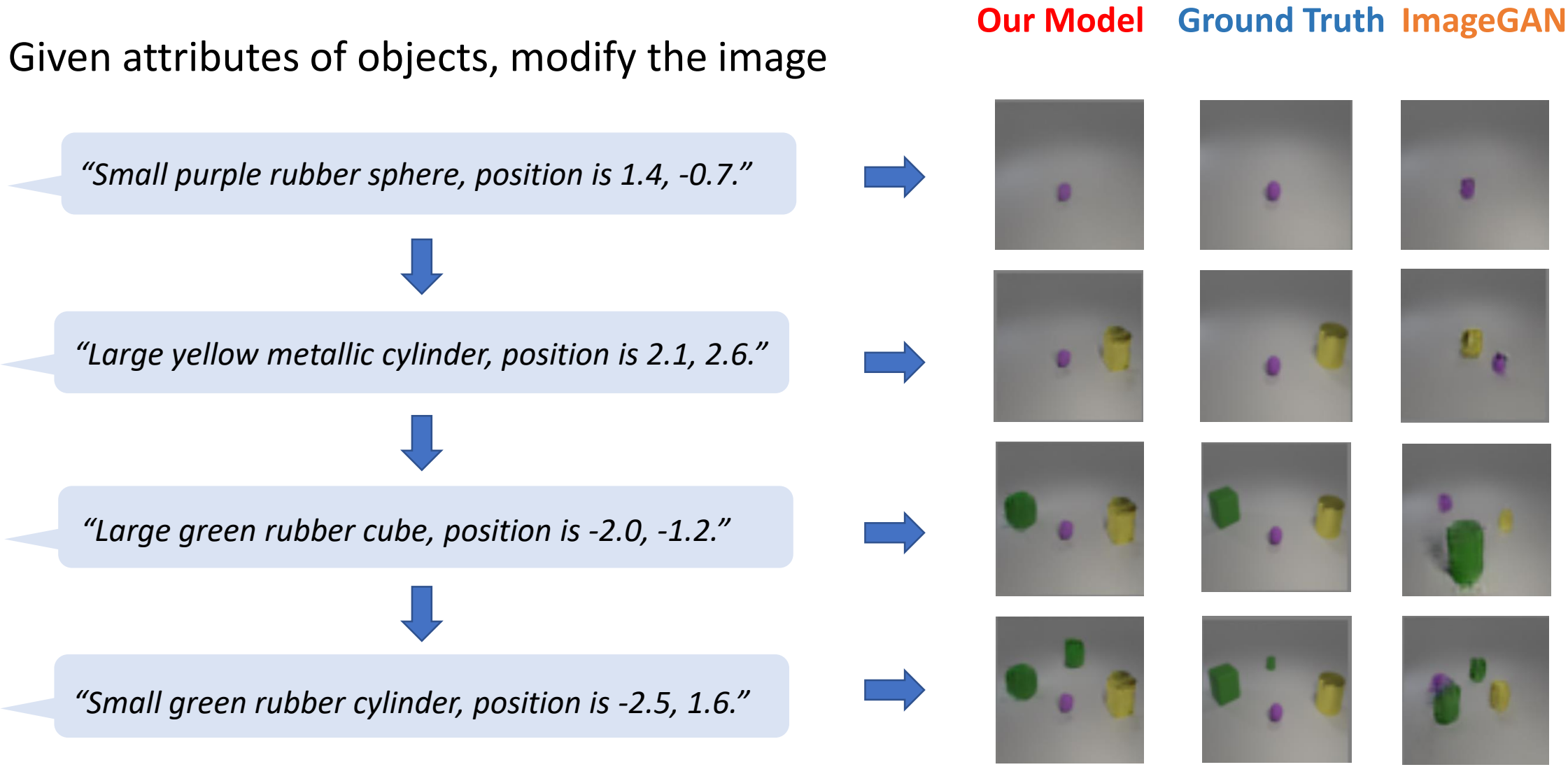


Story Discriminator



StoryGAN on CLEVR-SV Dataset

- Given attributes of objects, modify the image



StoryGAN on Cartoon Dataset

Loopy laughs but tends to be angry.
Pororo is singing and dancing and loopy is angry.
Loopy says stop to Pororo. Pororo stops.
Loopy asks reason to pororo. pororo is startled.
Pororo is making an excuse to loopy.

Input Story: *c1 and c2 are standing in the snow. c1 tells a story to c3. c3 wants to joint c1 and c2. c1 continuous to talk. c1 looks down. They suddenly noticed that there is something lying on the snow.*

StoryGAN: Experimental Results

	ImageGAN [26]	SVC	SVFN	StoryGAN
SSIM	0.596	0.641	0.654	0.672

Table 1: SSIM comparison on CLEVR-SV dataset.

	Upper Bound	ImageGAN [26]	SVC	SVFN	StoryGAN
Acc.	0.86	0.23	0.21	0.24	0.27

Table 2: Character classification accuracy (exact match ratio) comparison on Pororo-SV dataset. The upper bound is the classifier accuracy on the real images associated with the stories.

Table 3: Results of pairwise human evaluation. The ± denotes standard error on the metrics.

	StoryGAN vs ImageGAN		
Choice (%)	StoryGAN	ImageGAN	Tie
Visual Quality	74.17 $\pm$1.38	18.60 $\pm$1.38	7.23
Consistence	79.15 $\pm$1.27	15.28 $\pm$1.27	5.57
Relevance	78.08 $\pm$1.34	17.65 $\pm$1.34	4.27

Table 4: Results of ranking-based human evaluation. The ± denotes standard error on the metrics.

Method	ImageGAN	SVC	SVFN	StoryGAN
Rank	2.91 $\pm$0.05	2.42 $\pm$0.04	2.77 $\pm$0.04	1.94 $\pm$0.05

Concurrent Work

Keep Drawing It: Iterative Language-based Image Generation and Editing

Drawer: what is there ? **Teller:** large apple tree the far right with half of it off the picture. **Drawer:** done and

Teller: cloud in the left corner with just slight part of the left hanging off. **Drawer:** okay and

Teller: girl with arms up under the tree with part of her feet hanging off picture facing right. **Drawer:** okay and

Teller: guy slightly left of middle with hair above horizon both hands in the air facing right. **Drawer:** done and

Teller: table far left under horizon line with left side out of picture , cat under the leg of that. **Drawer:** done

Sequential Attention GAN for Interactive Image Editing via Dialogue