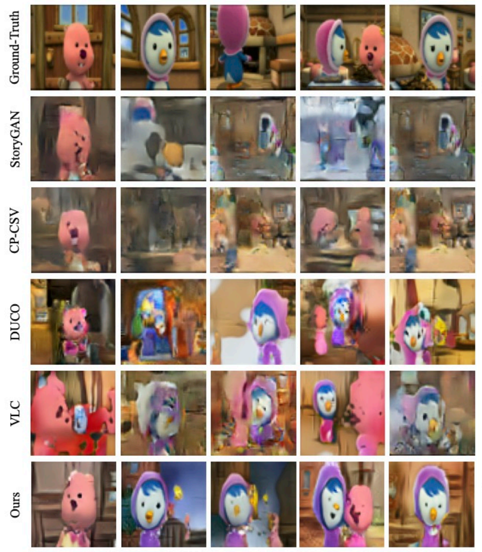
Word-Level Fine-Grained Story Visualization

Bowen Li and Thomas Lukasiewicz

UNIVERSITY OF OXFORD

Introduction



1. Loopy is in the house. Loopy holds skis in her hand. Loopy is looking around the house.

2. Petty gets surprised. Petty raises her hands. Outside the window the sky is blue and clear.

3. Petty rushes to the oven and pulls out the plate.

4. Petty and Loopy are seated beside a table. On the table there are two plates of cookies. Lots of cookies are stacked on the plate. The cookies are all burnt.

5. Petty has a plate of cookies on the table. The cookies are burnt.

Fig. 1. Examples of story visualization on different methods, with the given story sentences and ground-truth story images.

Method

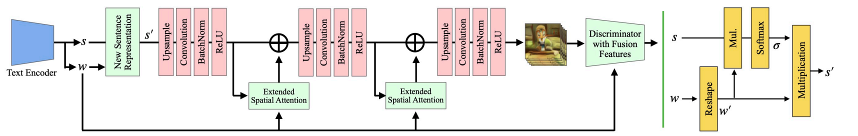
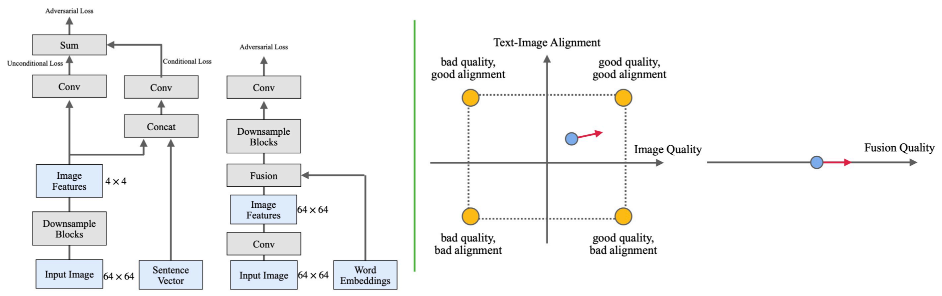


Fig. 2. Examples of story visualization on different methods, with the given story sentences and ground-truth story images.



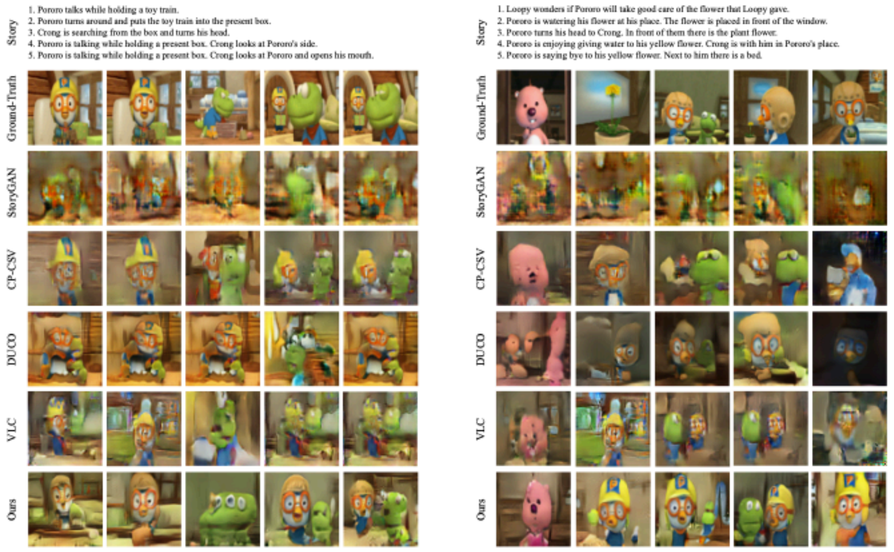
(1) normal discriminator (2) our discriminator (3) two goals (4) one goal

Fig. 3. Left: comparison between the normal discriminator (1) and ours (2). Right: a diagram for two goals of current methods (3) and the goal of our proposed one-way output design (4).

Experiments

Table 1. Quantitative evaluation between different methods on Pororo-SV and Abstract Scenes. For FID and FSD, lower is better; for text-image cosine similarity (Cosine), higher is better.

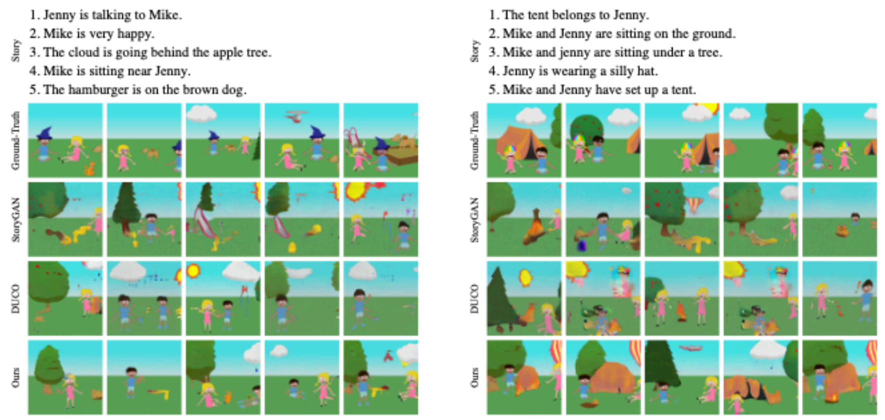
	Pororo-SV dataset			Abstract dataset		
Method	FID↓	FSD↓	Cosine↑	FID↓	FSD↓	Cosine↑
StoryGAN [14]	78.64	94.53	0.22	135.16	55.80	3.59
CP-CSV [24]	67.76	71.51	0.32	-	-	-
DUCO [17]	95.17	171.70	0.08	142.34	49.16	3.95
VLC [16]	94.30	122.07	0.21	-	-	-
Ours	**56.08**	**52.50**	**2.98**	**72.34**	**14.86**	**4.05**



1. Pororo talks while holding a toy train.
2. Pororo turns around and puts the toy train into the present box.
3. Crong is searching from the box and turns his head.
4. Pororo is talking while holding a present box. Crong looks at Pororo's side.
5. Pororo is talking while holding a present box. Crong looks at Pororo and opens his mouth.

1. Loopy wonders if Pororo will take good care of the flower that Loopy gave.
2. Pororo is watering his flower at his place. The flower is placed in front of the window.
3. Pororo turns his head to Crong. In front of them there is the plant flower.
4. Pororo is enjoying giving water to his yellow flower. Crong is with him in Pororo's place.
5. Pororo is saying bye to his yellow flower. Next to him there is a bed.

Fig. 4. Comparison between different methods on the Pororo-SV.



1. Jenny is talking to Mike.
2. Mike is very happy.
3. The cloud is going behind the apple tree.
4. Mike is sitting near Jenny.
5. The hamburger is on the brown dog.

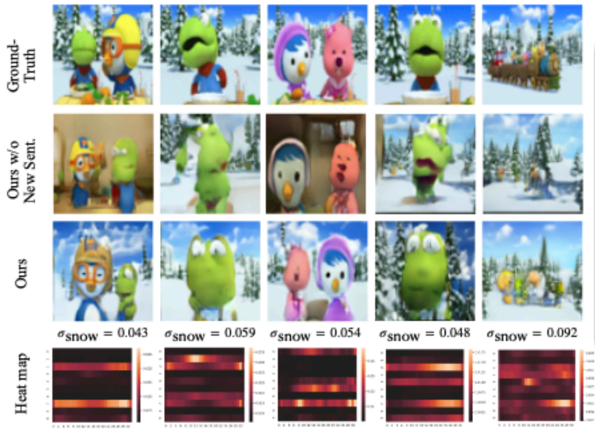
1. The tent belongs to Jenny.
2. Mike and Jenny are sitting on the ground.
3. Mike and jenny are sitting under a tree.
4. Jenny is wearing a silly hat.
5. Mike and Jenny have set up a tent.

Fig. 5. Comparison between different methods on the Abstract Scenes dataset.

Experiments

Table 2. Component Analysis on Pororo-SV. "Ours w/o Sentence" stands for without using the proposed new sentence representation; "Ours w/ Discriminator" stands for using the discriminator in current story visualization methods [14,24,17]; "Ours w/o Extended Spatial Attention" stands for without adopting the proposed attention; "Ours w/ Word-Level Spatial Attention" is with the implementation of word-level spatial attention [26], instead of our proposed extended spatial attention.

Method	FID	FSD	Cosine
Ours w/o New Sentence Representation	68.48	62.85	2.24
Ours w/ Discriminator [14]	62.23	59.33	2.61
Ours w/o Extended Spatial Attention	83.66	78.80	2.26
Ours w/ Word-Level Spatial Attention [26]	58.26	63.39	2.54
Ours w/ Pretrained BERT	52.38	49.69	3.71
Ours w/ FT BERT	50.96	48.81	3.95
Ours w/ BERT Scratch	55.78	51.71	2.80
Ours	56.08	52.50	2.98



1. Crong is eating vegetables. Pororo looks it strange.

2. Crong looks so happy. Crong finished the vegetables.

3. Loopy and petty are so happy to see Crong ate all Crong vegetables.

4. Crong looks happy. Crong nods Crong head.

5. Everyone is in the train. There are snow covered trees.

Fig. 6. Visualization of extended spatial attention and new sentence representation.

Table 3. Effects of our new sentence representation adopted in story visualization on Pororo-SV and text-to-image generation on CUB birds. For FID and FSD, lower is better, for IS, higher is better. "+ New Sent." means using our new representation.

Method	FID	FSD	IS
StoryGAN + New Sent.	72.81	84.06	-
CP-CSV + New Sent.	63.12	64.29	-
DUCO + New Sent.	87.82	131.83	-
VLC + New Sent.	82.19	100.94	-
AttnGAN [26]	23.98	-	4.36
AttnGAN + New Sent.	19.20	-	4.71
DFGAN [25]	14.81	-	5.10
DFGAN + New Sent.	11.98	-	5.16
Ours	56.08	52.50	-