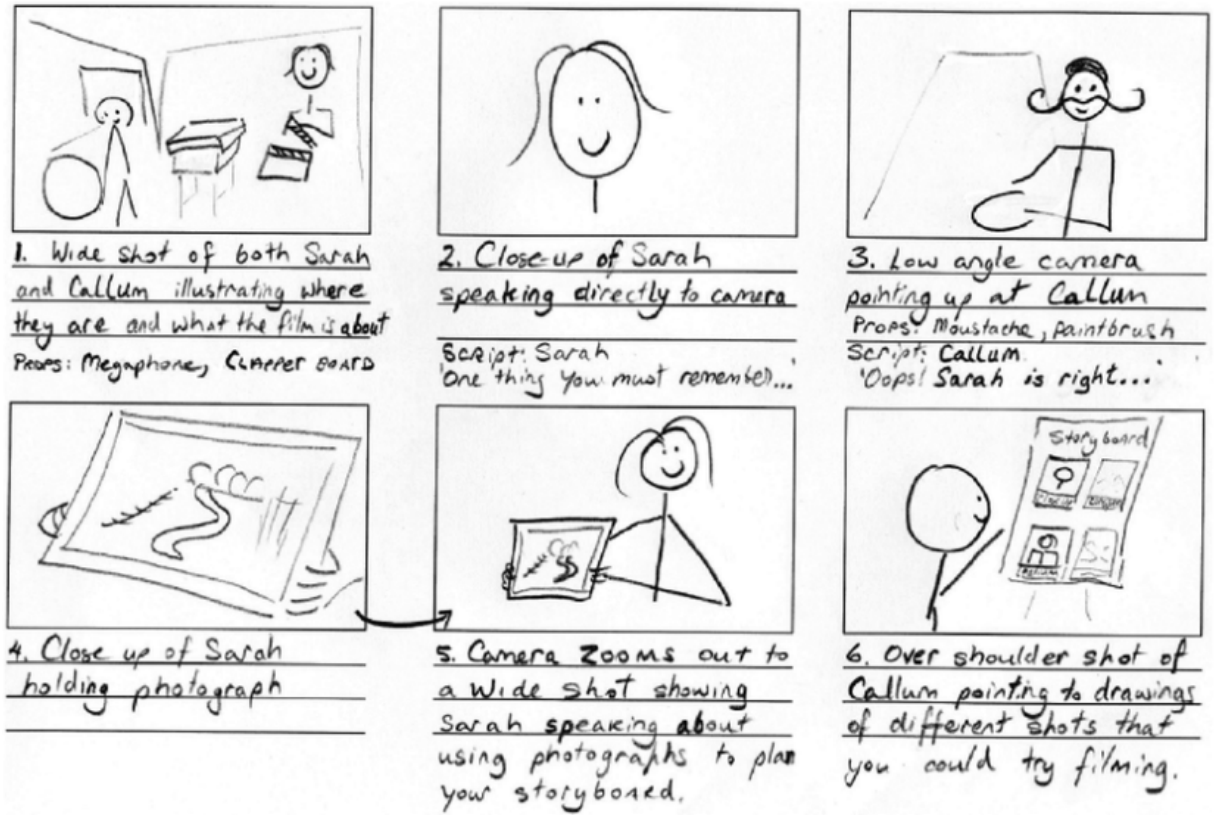
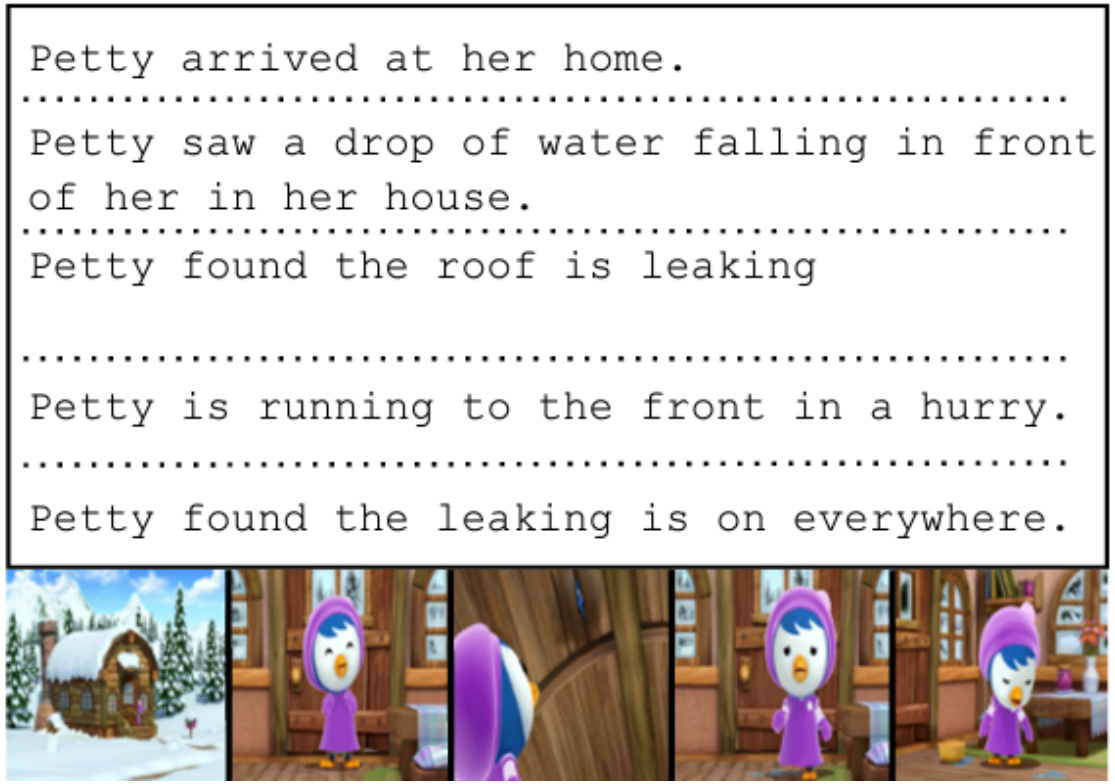
Recent approaches to Story Visualization (Storyboard Generation)

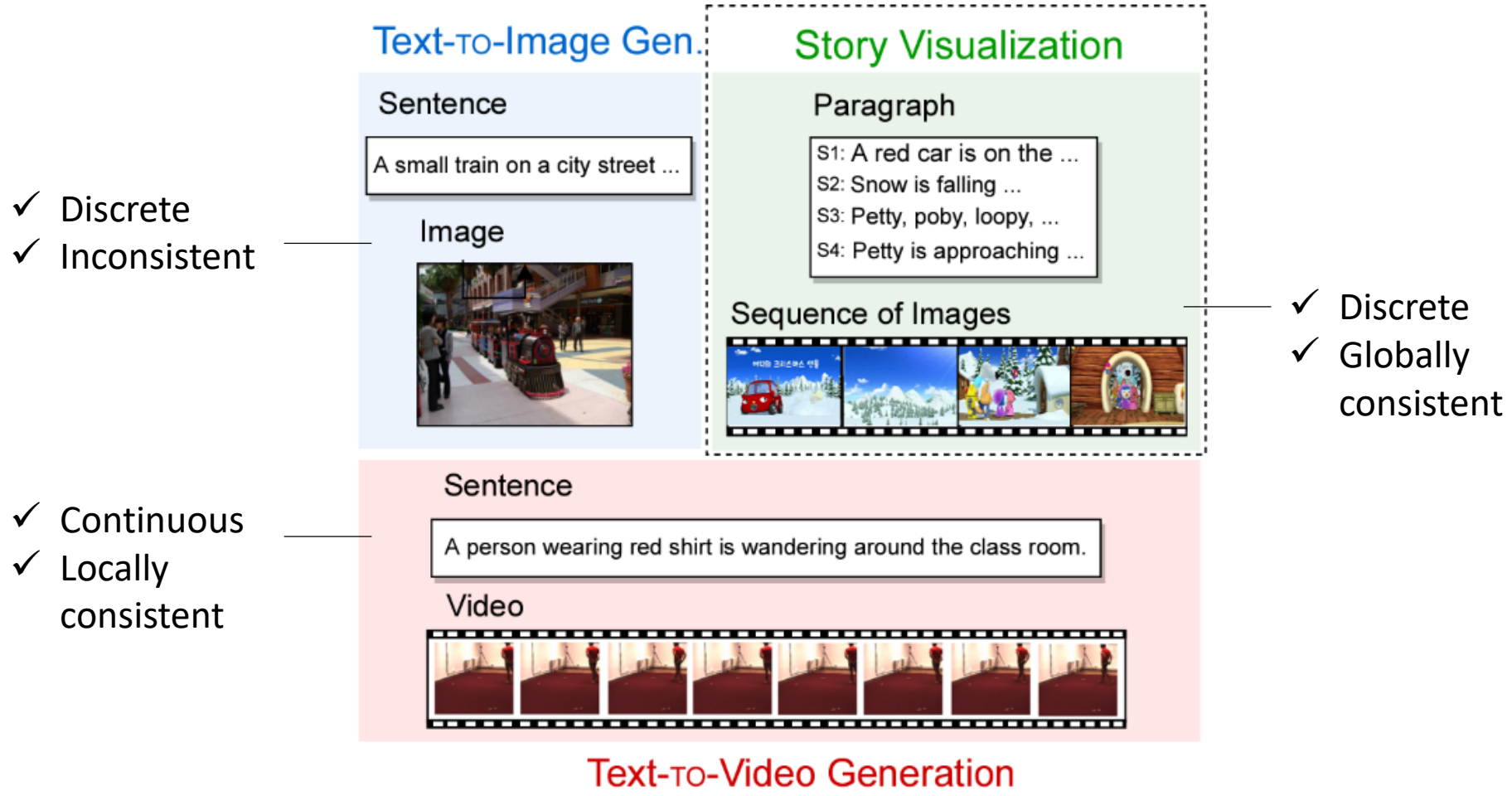




발표: 정채연

Task

Image vs. Visual Story vs. Video



- ✓ Discrete
- ✓ Inconsistent

- ✓ Discrete
- ✓ Globally consistent

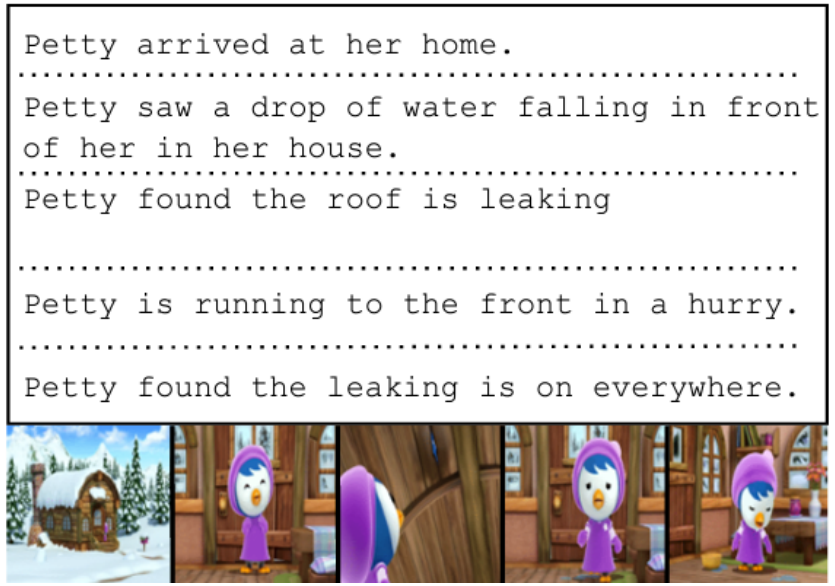
- ✓ Continuous
- ✓ Locally consistent

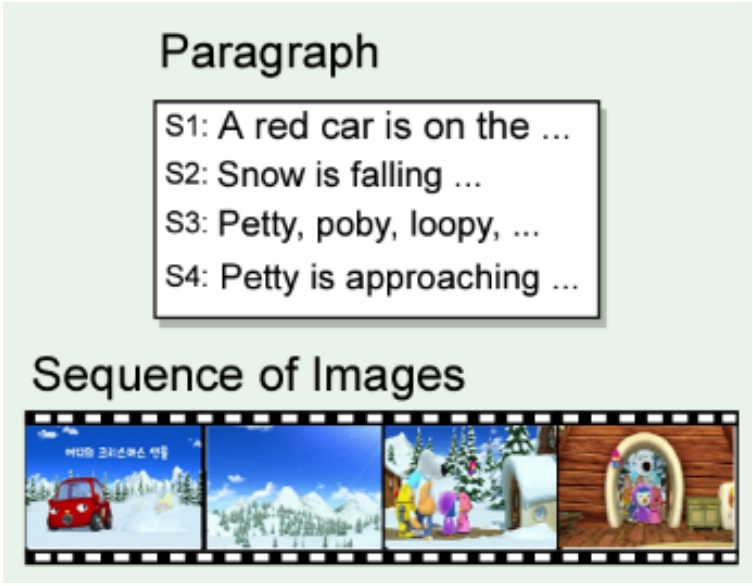
Task

Challenges in Story Visualization

1) Global consistency 유지

2) 현 sentence에 맞게 variation 필요 (e.g., 장면 전환)



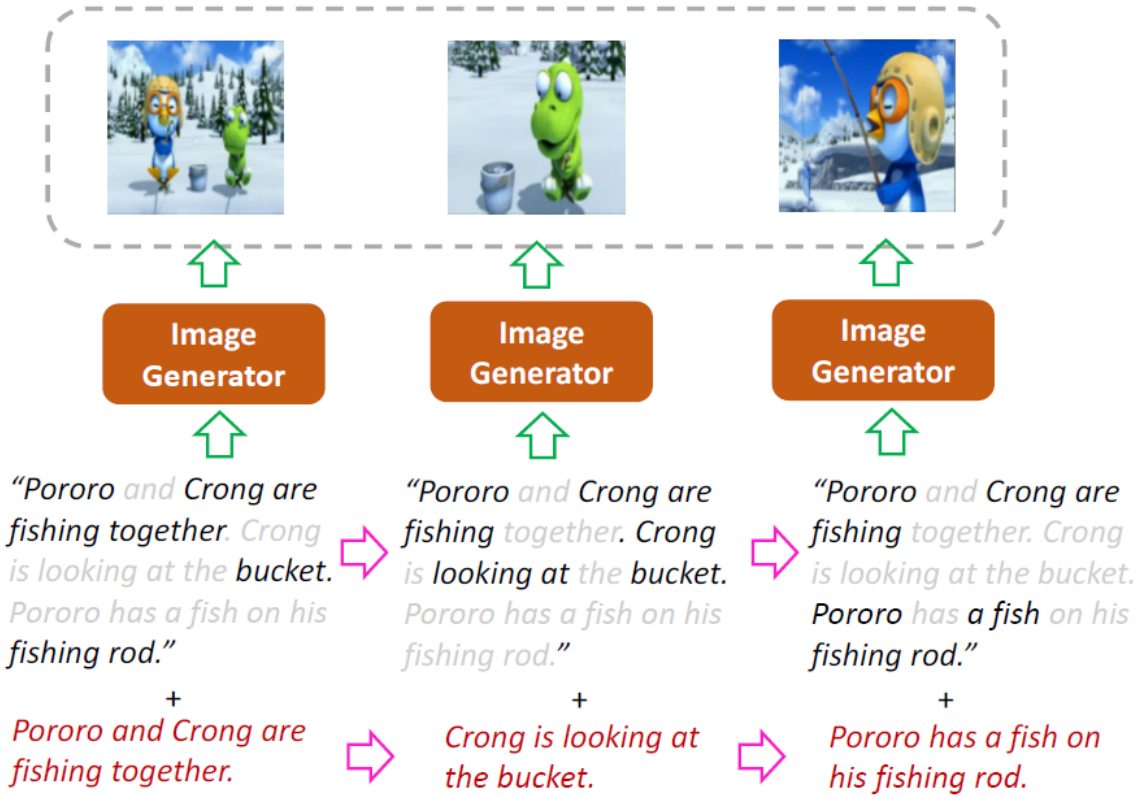


Previous Approaches

1. **StoryGAN**: A Sequential Conditional GAN for Story Visualization, *CVPR 2019*
2. **Improved-StoryGAN** for sequential images visualization, *JVCIR 2020*
3. (**CP-CSV**) Character-Preserving Coherent Story Visualization, *ECCV 2020*
4. (**DUCO-StoryGAN**) Improving Generation and Evaluation of Visual Stories via Semantic Consistency, *NAACL 2021*
5. (**VLC-StoryGAN**) Integrating Visuospatial, Linguistic and Commonsense Structure into Story Visualization, *EMNLP 2021*
6. Generating a Temporally Coherent Visual Story with Multimodal Recurrent Transformers, *ACL ARR 2022*

StoryGAN (CVPR 2019)

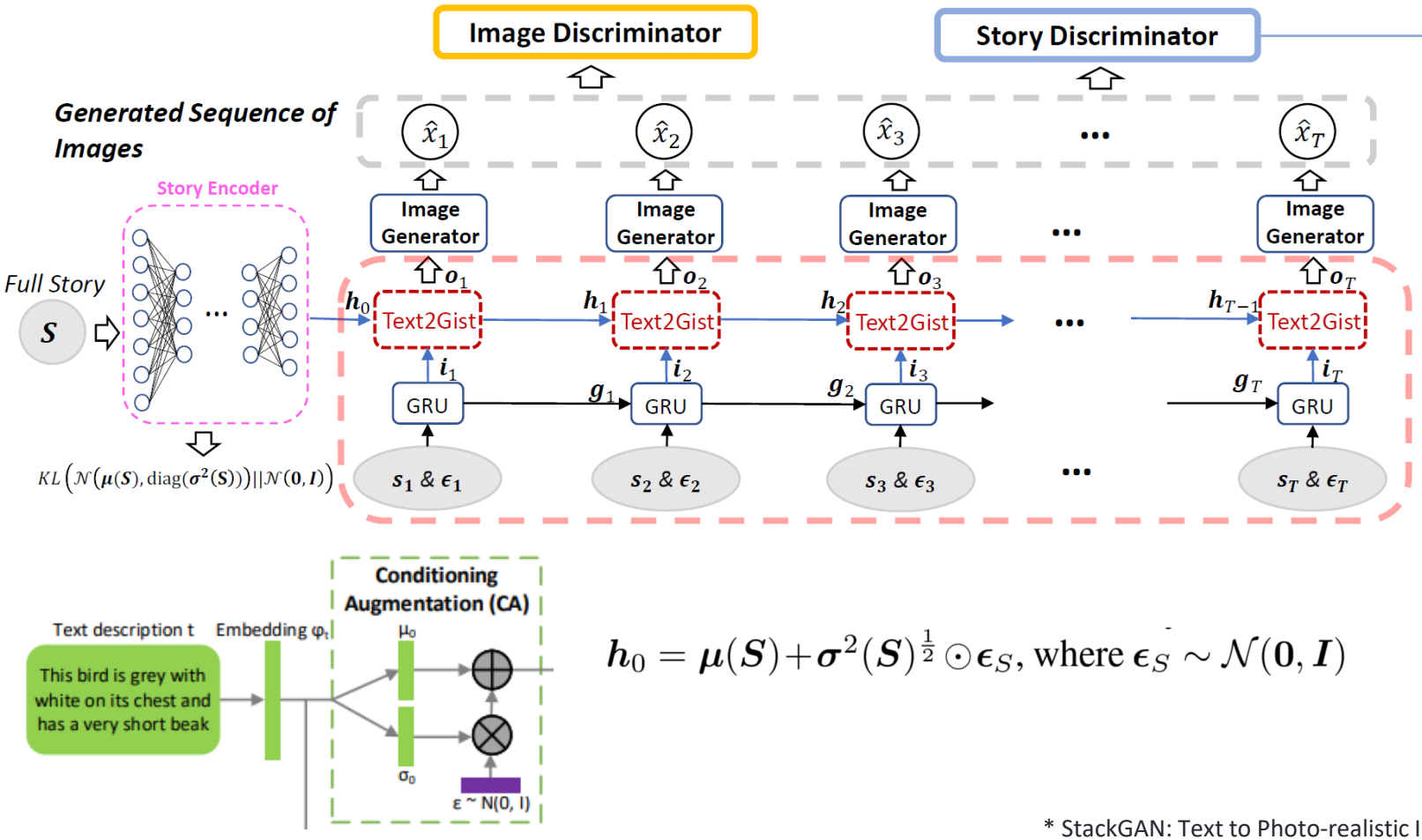
Overview

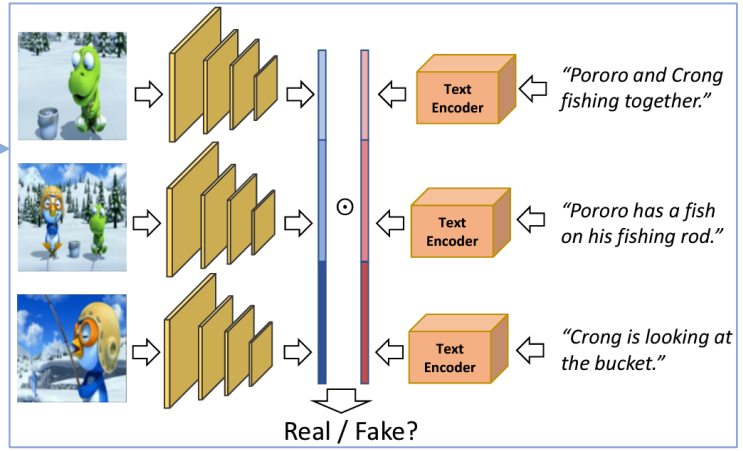


- Story visualization task 처음 제안
- GRU 기반 구조를 통해, full story + current sentence 정보 함께 반영한 image sequence 생성

StoryGAN (CVPR 2019)

Method





$$KL\left(\mathcal{N}(\boldsymbol{\mu}(\boldsymbol{S}), \mathrm{diag}(\boldsymbol{\sigma}^2(\boldsymbol{S}))) || \mathcal{N}(\mathbf{0}, \boldsymbol{I})\right)$$

$$\boldsymbol{h}_0 = \boldsymbol{\mu}(\boldsymbol{S}) + \boldsymbol{\sigma}^2(\boldsymbol{S})^{\frac{1}{2}} \odot \boldsymbol{\epsilon}_S, \text{ where } \boldsymbol{\epsilon}_S \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$$

Text2Gist

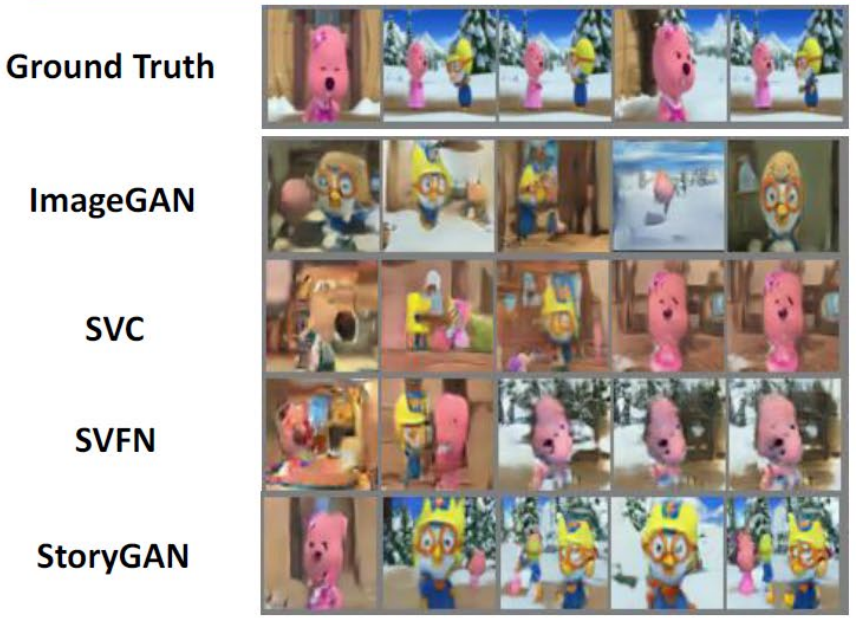
$$\begin{aligned}
\boldsymbol{z}_t &= \sigma_z\left(\boldsymbol{W}_z \boldsymbol{i}_t + \boldsymbol{U}_t \boldsymbol{h}_{t-1} + \boldsymbol{b}_z\right), \\
\boldsymbol{r}_t &= \sigma_r\left(\boldsymbol{W}_r \boldsymbol{i}_t + \boldsymbol{U}_r \boldsymbol{h}_{t-1} + \boldsymbol{b}_r\right), \\
\boldsymbol{h}_t &= (\mathbf{1} - \boldsymbol{z}_t) \odot \boldsymbol{h}_{t-1} \\
&\quad + \boldsymbol{z}_t \odot \sigma_h\left(\boldsymbol{W}_h \boldsymbol{i}_t + \boldsymbol{U}_h(\boldsymbol{r}_t \odot \boldsymbol{h}_{t-1}) + \boldsymbol{b}_h\right), \\
\boldsymbol{o}_t &= \mathrm{Filter}(\boldsymbol{i}_t) * \boldsymbol{h}_t,
\end{aligned}$$

* StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks (ICCV2017)

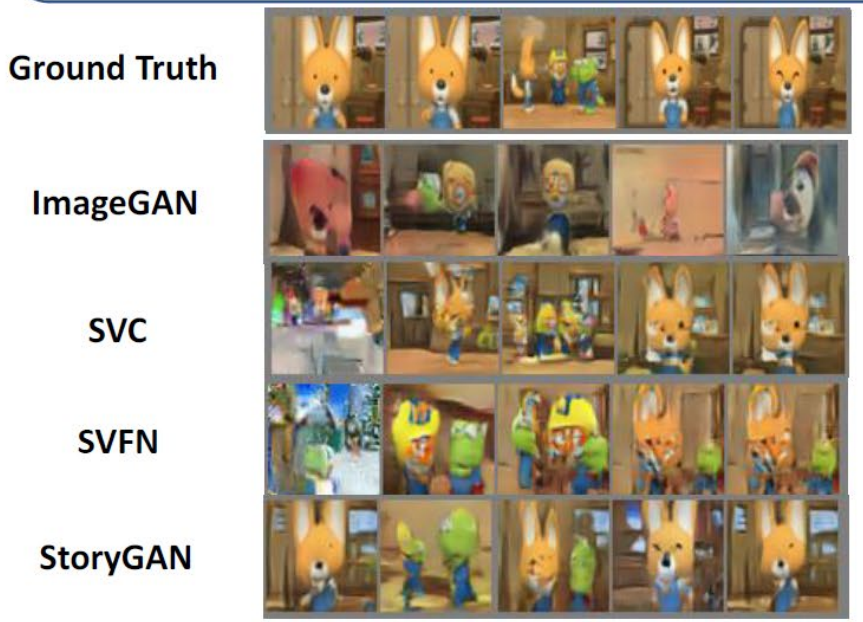
StoryGAN (CVPR 2019)

Experiment Results

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.



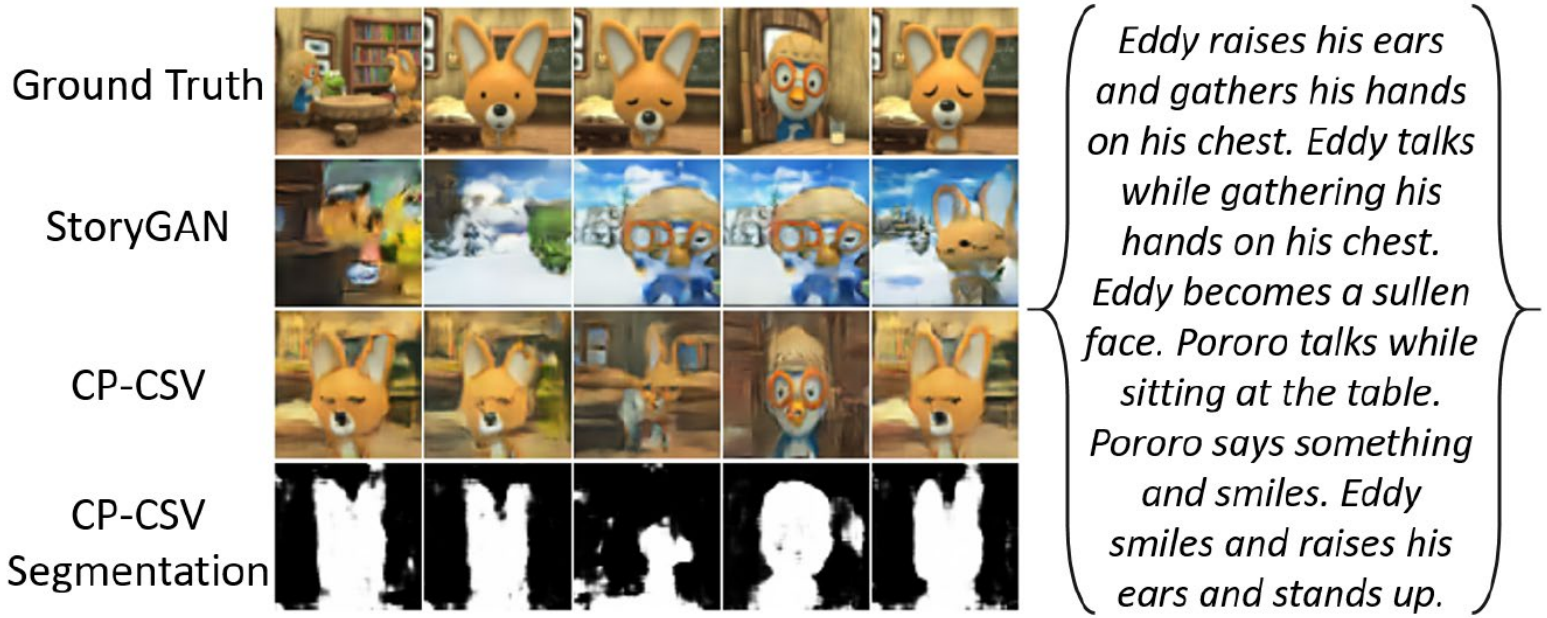
Eddy is shocked at what happened now.
Pororo tells Eddy that Crong was cloned.
Pororo tells Eddy that Crong got into the machine.
Eddy says it is not a problem.
Eddy tells them that Eddy made a machine to reverse the cloning.



CP-CSV (ECCV 2020)

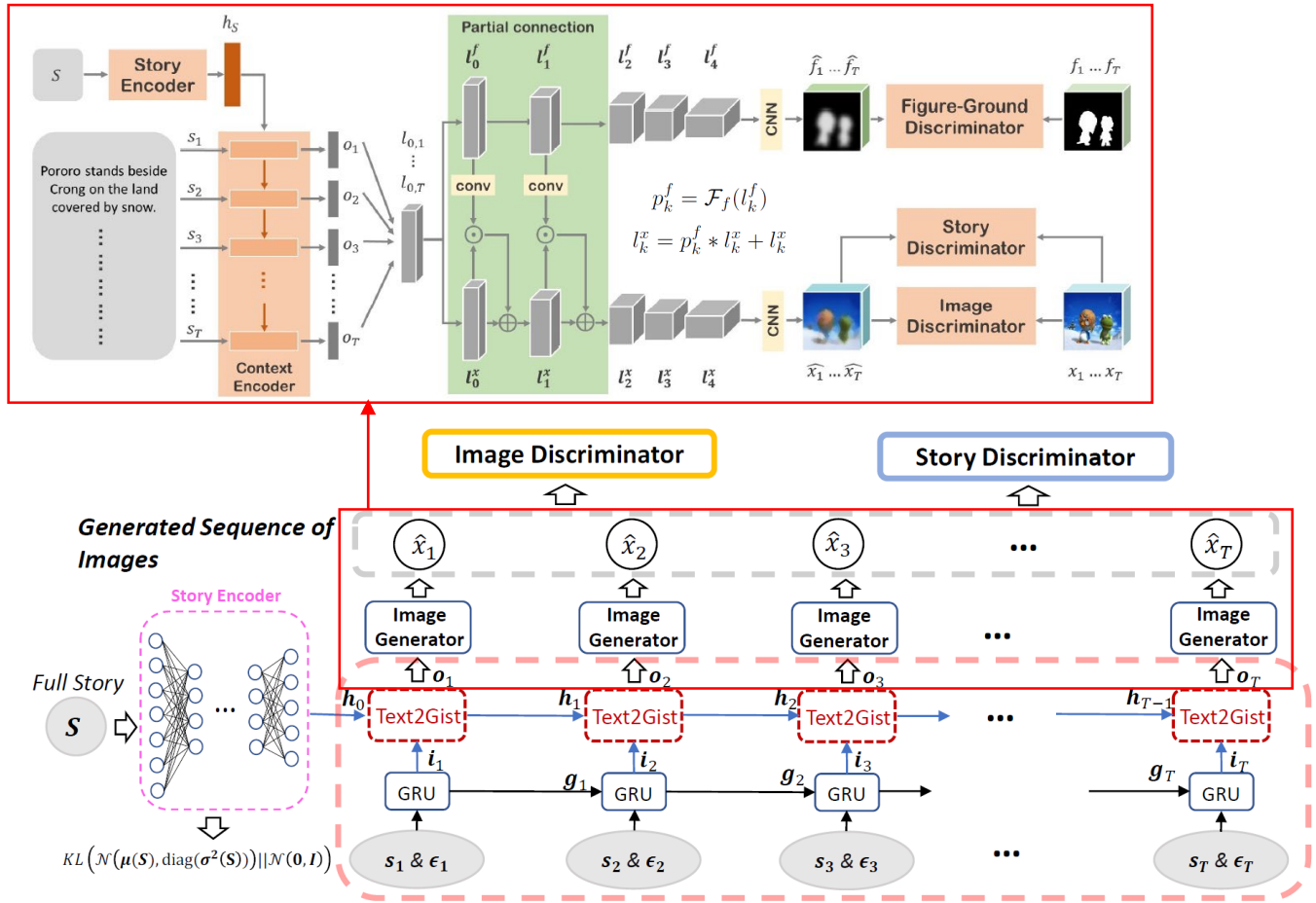
Overview

- StoryGAN + α
- Figure-ground segmentation as auxiliary task
- Evaluation metric Frechet Story Distance (FSD) 제안



CP-CSV (ECCV 2020)

Method



- Frechet Story Distance (FSD) using R(2+1)D network

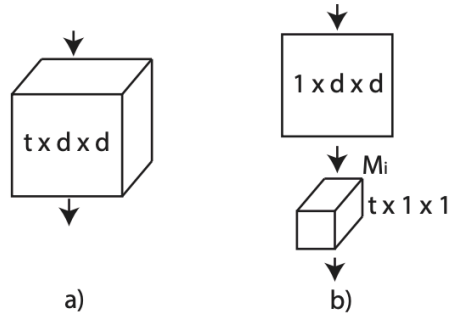


Figure 2. **(2+1)D vs 3D convolution**. The illustration is given for the simplified setting where the input consists of a spatiotemporal volume with a single feature channel. (a) Full 3D convolution is carried out using a filter of size $t \times d \times d$ where t denotes the temporal extent and d is the spatial width and height. (b) A (2+1)D convolutional block splits the computation into a spatial 2D convolution followed by a temporal 1D convolution. We choose the numbers of 2D filters (M_i) so that the number of parameters in our (2+1)D block matches that of the full 3D convolutional block.

DUCO-StoryGAN (NAACL 2021)

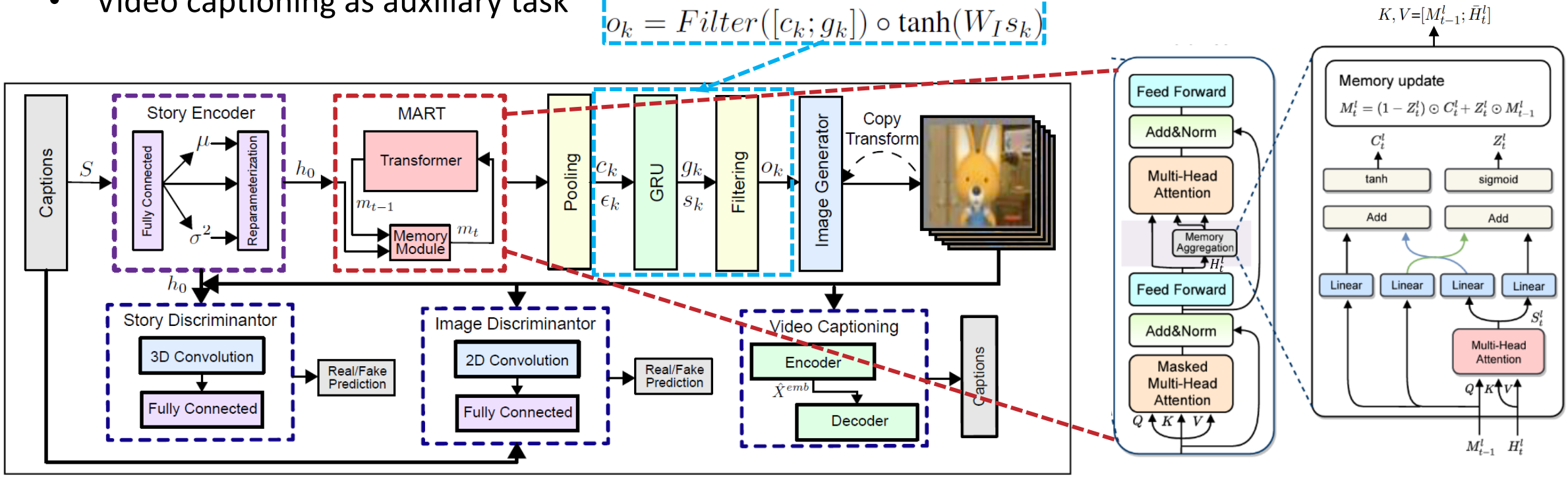
Method

- StoryGAN + α
- Video captioning as auxiliary task

$$[m_{k1}, ...m_{kL}], h_k = \text{MART}([w_{k1}, ...w_{kL}], h_{k-1})$$

$$c_k = \sum_{i=1}^{L} \alpha_{ki} m_{ki};\ \alpha_{ki} = \frac{exp(m_{ki}^T u)}{\sum exp(m_{ki}^T u)}$$

$$g_k,\ q_k = \text{GRU}(s_k, \epsilon_k, q_{k-1})$$
$$o_k = Filter([c_k; g_k]) \circ \tanh(W_I s_k)$$



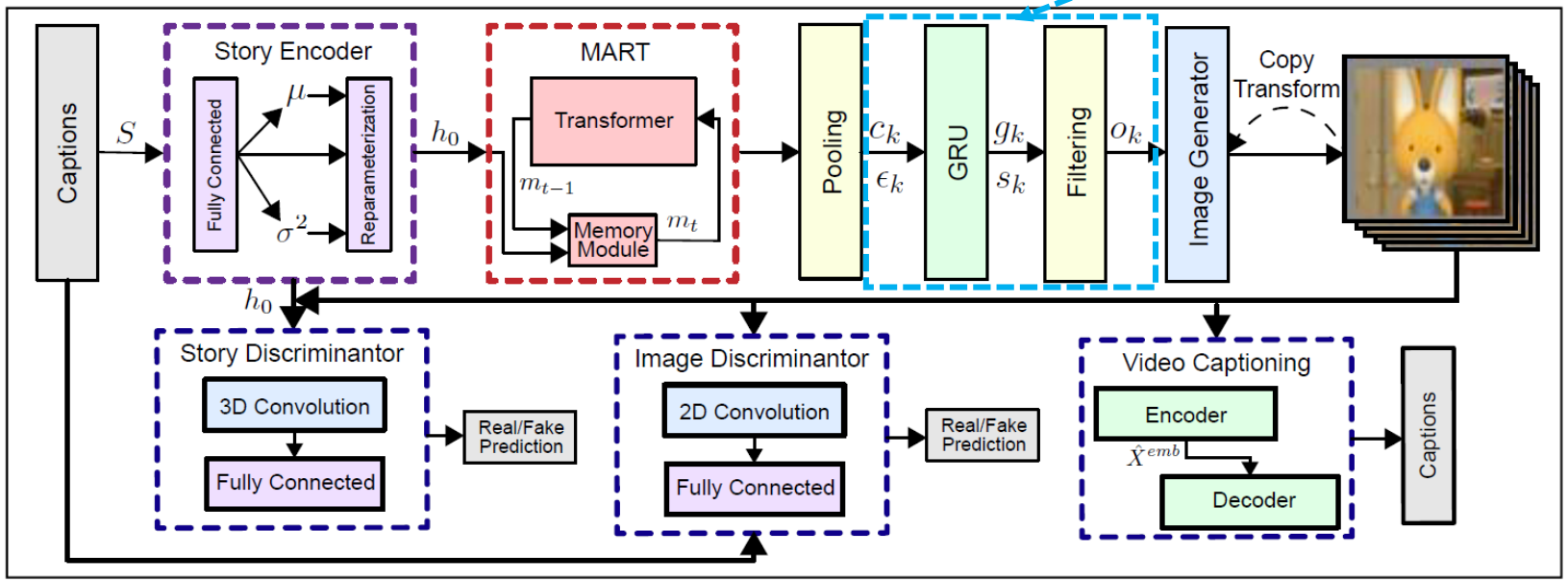
MART: Memory-Augmented Recurrent Transformer for Coherent Video Paragraph Captioning (ACL 2020)

DUCO-StoryGAN (NAACL 2021)

Method

- StoryGAN + α
- Video captioning as auxiliary task

$$g_k,\ q_k = \mathrm{GRU}(s_k, \epsilon_k, q_{k-1})$$
$$o_k = Filter([c_k; g_k]) \circ \tanh(W_I s_k)$$



Copy Transform

$$c_{jk} = \sum_{i=0}^{L} \beta_{ji} m'_{ik};\quad \beta_{jik} = \frac{\exp(h_j^T m'_{ik})}{\sum_{i=0}^{L} \exp(h_j^T m'_{ik})}$$

- h: sub-region in previous image feature
- m: current word vector
- β_ji: the weight assigned by the model to the i-th word when generating the j-th sub-region of the image
- c_jk: weighted sum of all word representations

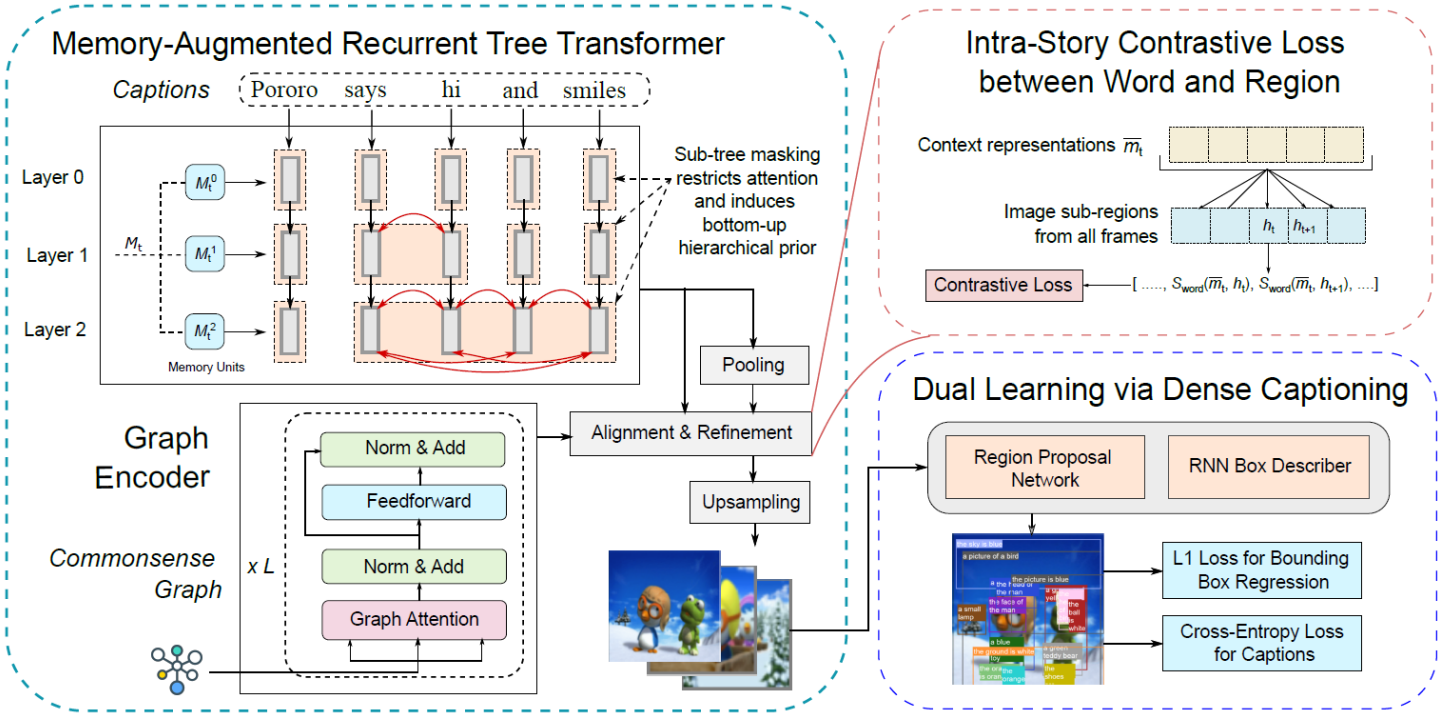
AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks (CVPR 2018)

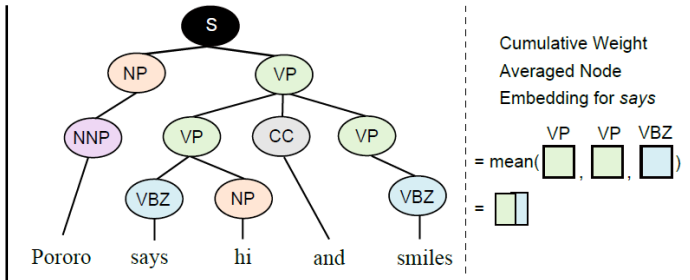
VLC-StoryGAN (EMNLP 2021)

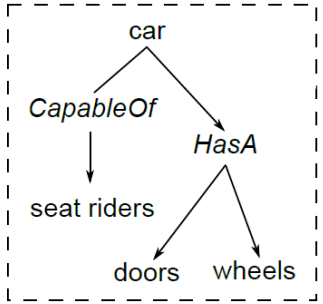
Method

- StoryGAN + α
- Dense captioning as auxiliary task

$$\mathcal{L}_{word} = -log\frac{exp(S_{word}(x_k, s_k))}{\sum_{m=1}^{T} exp(S_{word}(x_m, s_k))}$$







Caption: The red car is offering Petty and Loopy a ride. The red car opens its doors for friends.

VLC-StoryGAN (EMNLP 2021)

Experiment Results

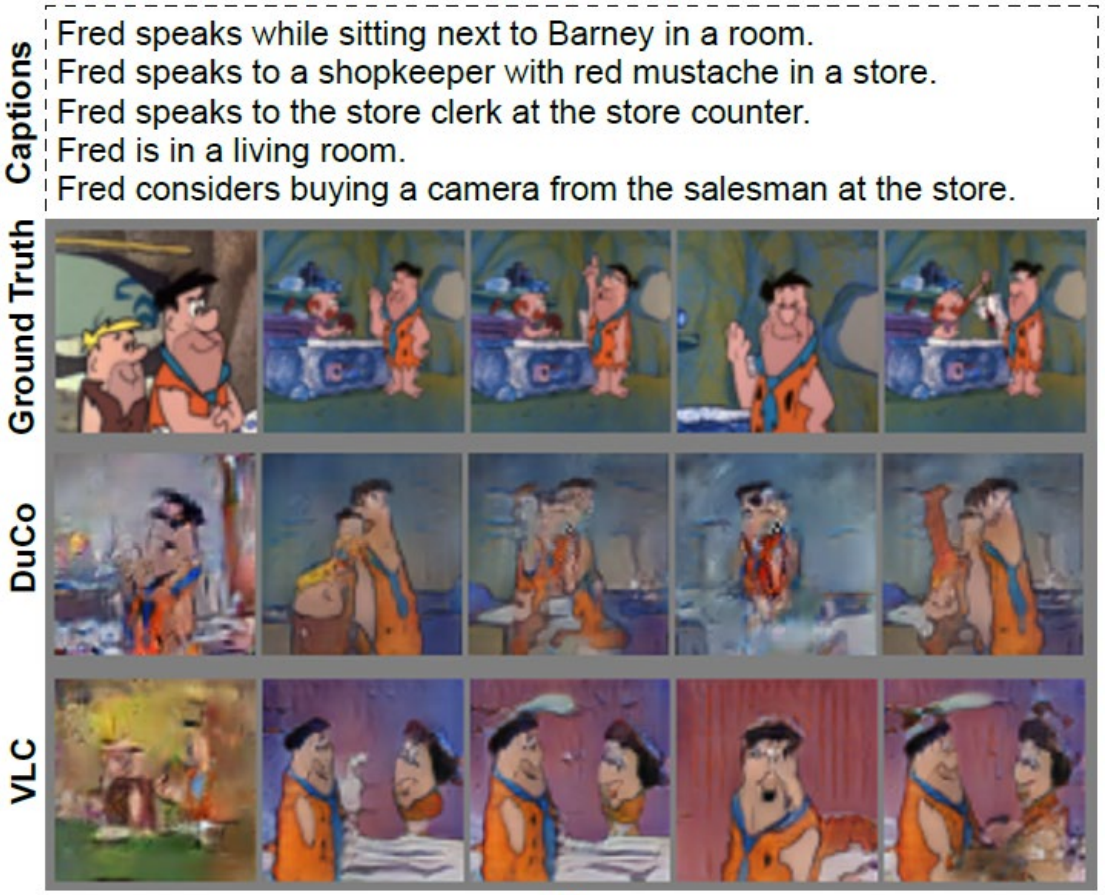
Model	Char. F1	BLEU2/3	R-Precision	Frame Acc.	Top-1 Acc.	Top-2 Acc.
StoryGAN (Li et al., 2019b)	18.59	3.24 / 1.22	1.51 ± 0.15	9.34	23.14	42.27
StoryGAN + Transformer	19.29	3.29 / 1.23	1.49 ± 0.07	9.58	23.31	42.29
CP-CSV (Song et al., 2020)	21.78	3.25 / 1.22	1.76 ± 0.04	10.03	22.23	41.86
DUCO-STORYGAN	**38.01**	**3.68 / 1.34**	$\mathbf{3.56 \pm 0.04}$	**13.97**	**23.72**	**42.48**

Model	Char. F1	BLEU2/3	R-Precision	Frame Acc.	Top-1 Acc.	Top-2 Acc.
StoryGAN (Li et al., 2019b)	41.11	3.86 / 1.72	3.40 ± 0.01	21.90	22.42	45.40
StoryGAN + Transformer	42.45	3.92 / 1.73	4.03 ± 0.17	22.14	23.79	46.15
StoryGAN + MART	47.03	4.15 / 1.81	5.11 ± 0.12	22.25	24.48	46.42
+ Story Captioning	47.23	4.78 / 1.87	6.32 ± 0.08	22.30	24.53	47.41
+ Copy Transform	48.27	4.51 / 1.92	6.10 ± 0.07	22.71	25.62	47.39

VLC-StoryGAN (EMNLP 2021)

Experiment Results

(a)



(b)

VLC-StoryGAN (EMNLP 2021)

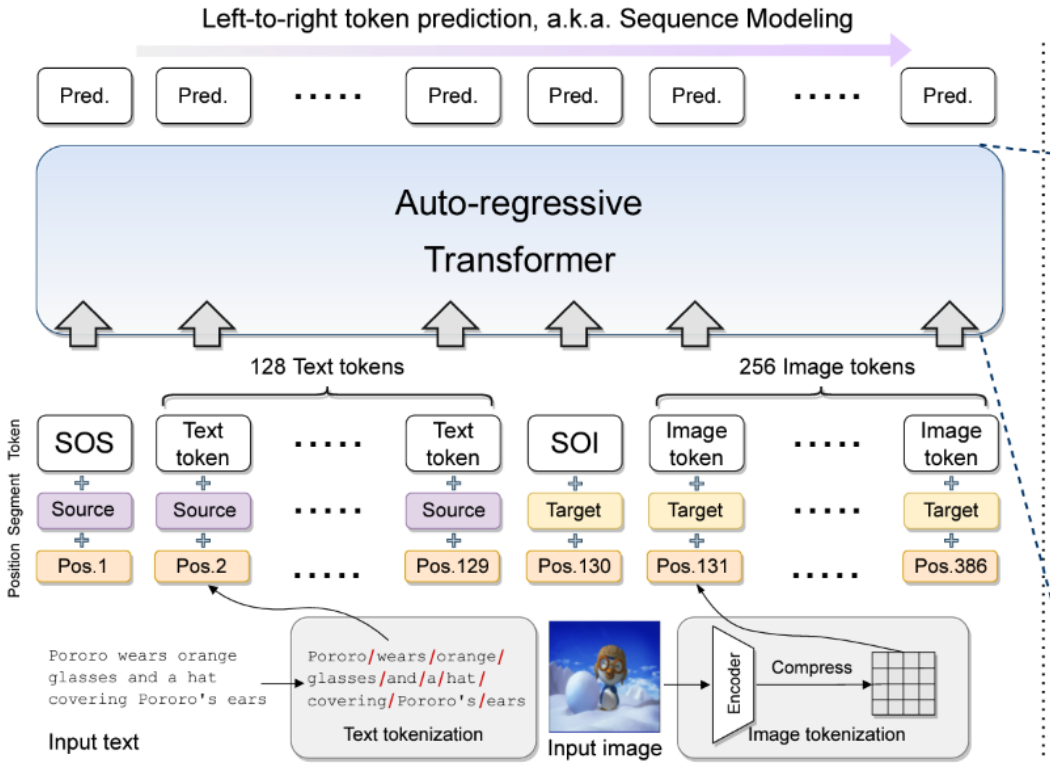
Experiment Results

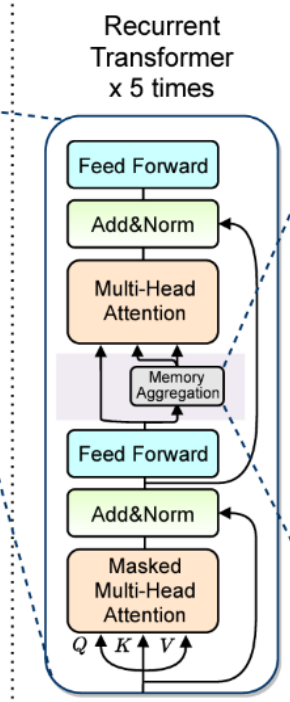
Generated GT

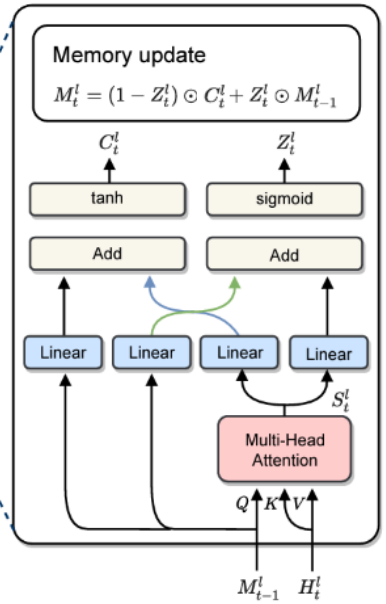
Generating a Temporally Coherent Visual Story with Multimodal Recurrent Transformers (ACL ARR 2022)

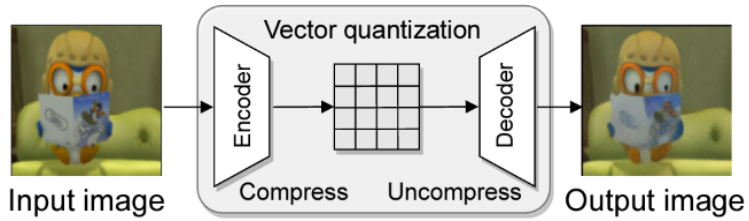
Method

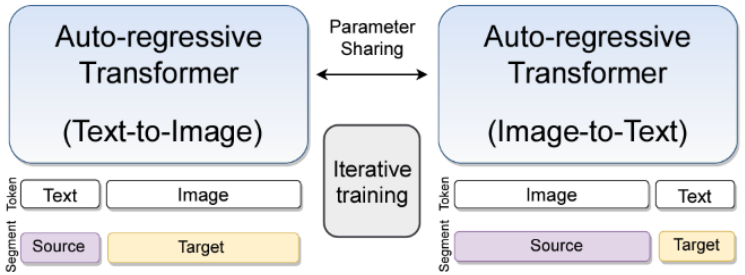
- **StoryGAN 구조 탈피!** → Auto-regressive Transformer + Memory module + VQGAN











$$\mathcal{L}_{txt2img} = \sum_{k=1}^{n} -\log P(z_k|t_1, ..., t_m, z_1, ..., z_{k-1})$$

$$\mathcal{L}_{img2txt} = \sum_{k=1}^{n} -\log P(t_k|z_1, ..., z_n, t_1, ..., t_{k-1})$$

Generating a Temporally Coherent Visual Story with Multimodal Recurrent Transformers (ACL ARR 2022)

Experiment Results

Petty arrived at her home.

Petty saw a drop of water falling in front of her in her house.

Petty found the roof is leaking

Petty is running to the front in a hurry.

Petty found the leaking is on everywhere.

Generating a Temporally Coherent Visual Story with Multimodal Recurrent Transformers (ACL ARR 2022)

Experiment Results

Methods	FID↓	FSD↓	Char. F1↑	Frame Acc.↑	BLEU2/3↑
DuCo (Maharana et al., 2021)	91.96	171.36	36.13	13.03	3.39 / 1.40
Baseline (Transformer-based model)	66.51	40.34	48.38	18.38	4.34 / 1.77
+ Memory-Augmented Recurrent	65.89	36.81	57.53	27.65	4.90 / 2.01
+ Nucleus Sampling	56.04	33.27	**59.20**	**28.69**	5.18 / 2.18
+ Bi-directional	52.20	31.43	57.18	26.81	5.23 / 2.27
+ Cyclic Pseudo-Text (**C-SMART**)	**50.24**	**30.40**	58.11	28.06	**5.30 / 2.34**