Synthesizing Coherent Story with Auto-Regressive Latent Diffusion Models

Authors: Xichen Pan , Pengda Qin, Yuhong Li, Hui Xue, Wenhu Chen
University of Waterloo, Alibaba Group, Vector Institute

Presented By: Anusha Lokumarambage

Story generation

- Problem
  - Single image generation relevant to a caption
  - Relevance and consistency in series of images

#2. *Wilma, Betty, and Barney are standing in the living room. Wilma is talking.*
#3. *Fred and Wilma are driving in the car.*
#4. *Fred is driving the car while listening to Wilma who is the passenger. Wilma looks angry while speaking to Fred as she has her arms crossed.*
#5. *The man in blue with a bow tie is sitting with his hands on a desk in the room. He is talking and then shakes his head while talking.*

Story Visualization & Continuation

- Synthesize a series of images to describe a story containing multiple sentences
  - Need to identify characters, objects,
  - Consistently follow history during the image generation

Related work

- Single image generation models
  - DALL-E
  - Imagen
  - Stable Diffusion

- Textual Inversion
- DreamBooth
- Re-Imagen

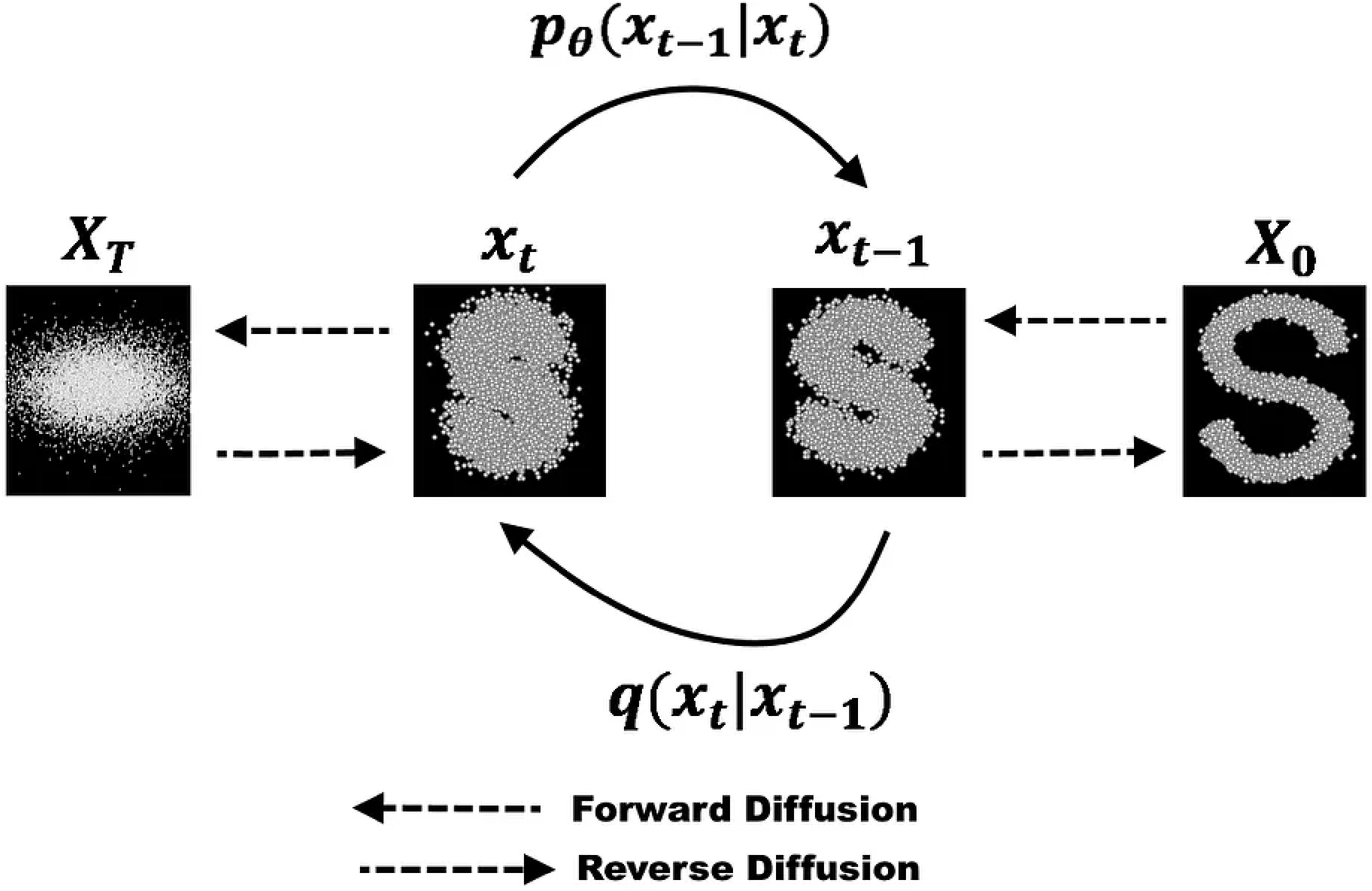
GAN based models for story visualization

- StoryGAN
  - Context encoder
  - Image generator
  - Sperate image and story discriminator – aim to preserve image consistency
- DUCO-StoryGAN
  - Uses copy-transform and dual learning by using features from previously generate images through attention mechanism to improve consistency to improve story visualization.

GAN based models for story visualization

- VLC-StoryGAN
- WordLevel SV
  - Focus on text inputs, use structured input and sentence representation to better guide visual story generation
- Story-DALL-E
  - Uses pre-trained transformers – DALL-E and achieves better results than GAN based models

Diffusion Models



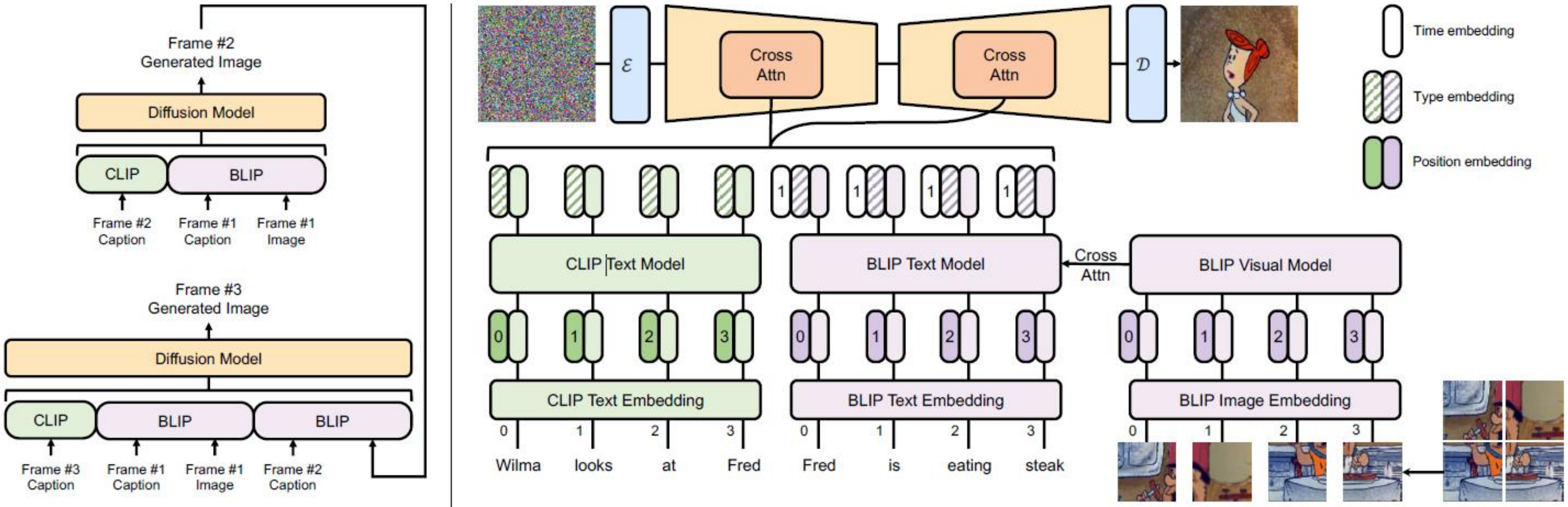
Auto Regressive Latent Diffusion Model

Requirement

- Model to be aware of history descriptions and scenes
- Ex :
  - “A red metallic cylinder cube is at the center.
  - Then add a green rubber cube at the right”

AR-LDM

- Get rids of the assumption of conditional independence between each frame in existing models



(a) Auto-regression process.

(b) Model Architecture.

History-Aware Conditioning Network

- Encode the history caption-image pairs into multimodal condition to guide denoising process
  - Consists of
    - CLIP – Concatenates unimodal embeddings
    - BLIP – pre-trained using vision-language understanding and generation tasks, uses cross-attention mechanism to integrate visual and language modalities

Experiments

Uses PororoSV, Flinstone and VIST

- Contains stories in 5 consecutive frames and captions.

First frame is fed as source frame, rest of the 4 frames has to be generated using captions with reference to the source frame.

Results - Quantitative

- Uses SoTA FID score
  - metric used to evaluate the quality and diversity of generated images
  - FID measures the similarity between the distribution of real images and the distribution of generated images. It calculates the Fréchet distance between two multivariate Gaussian distributions fitted to the real and generated image features, respectively. The lower the FID score, the better the generated images are considered to be.

Models	# of Params	PororoSV	FlintstonesSV	VIST-SIS	VIST-DII
StoryGANc [21]	-	74.63	90.29	-	-
StoryDALL·E (prompt tuning) [21]	1.3B	61.23	53.71	-	-
StoryDALL·E [21]	1.3B	25.90	26.49	-	-
MEGA-StoryDALL·E [21]	2.8B	23.48	23.58	20.98*	24.61*
AR-LDM (Ours)	1.5B	**17.40**	**19.28**	**16.95**	**17.03**

Table 1. Story continuation FID scores (lower is better) of AR-LDM and several previous models. * denotes experimental results reproduced by us, where we trained MEGA-StoryDALL·E for 50 epochs using the same training strategies as AR-LDM.

Models	FID
StoryGAN [17]	158.06
CP-CSV [33]	149.29
DUCO-StoryGAN [20]	96.51
VLC-StoryGAN [19]	84.96
VP-CSV [2]	65.51
Word-Level SV [15]	56.08
AR-LDM (Ours)	**16.59**

Table 2. Story visualization FID score results on PororoSV. We use the results reported by [2] and [15].

Results-Human Evaluation

Dataset	Criterion	Win (%)	Tie (%)	Lose (%)
PororoSV	Visual Quality	41.8	17.4	40.8
	Relevance	18.0	28.6	53.4
	Consistency	3.8	3.2	93.0
FlintstonesSV	Visual Quality	42.2	20.0	37.8
	Relevance	24.6	26.4	49.0
	Consistency	2.6	13.2	84.2
VIST-SIS	Visual Quality	14.6	20.6	64.8
	Relevance	19.2	48.6	32.2
	Consistency	3.0	46.2	50.8

Table 5. Human evaluation results of story continuation task on PororoSV, FlintstonesSV, and VIST-SIS datasets. The comparison is between visual stories synthesized by AR-LDM and ground truth reference ones.

diffusion using same 3-5 images cannot obtain satisfying generation result, because it confuses other characters with `<char>` and fails to generate them. Additional cases can be found in Appendix E.

Limitations

observe that 49.2% of generated stories on VIST are as consistent as ground truth.

PororoSV and FlintstonesSV datasets whose frames are sampled from videos, few synthesized visual stories are as consistent as ground truth references

Consistency is short and there is room for improvement

Thank you