Exploring the Landscape: A Survey of Story-to-Image Generator Models on a Novel Dataset

Siddharth Baskar
sbaskar2@wisc.edu

Karan Vikyath Veeranna Rupashree
veerannarupa@wisc.edu

Anudeep Kumar
kumar256@wisc.edu

1. Introduction

The original idea behind the project was to generate images using text prompts related to historical to provide context and generate interest. But the project is now diverted from the original idea due to implementation handicaps faced due to the unavailability of any dataset that could be used to train the model on for an accurate historical depiction. We then decided to train story visualization models on a new dataset titled Sequential Storytelling Image dataset that has just been released and none of the models have been trained on it yet. The decision to proceed with this dataset relied on its similarity to VIST dataset that is one of the more popular dataset for story visualization, in terms of having images in groups of 5 and labelled captions for story description. But, unlike VIST which is generated using random real world images, SSID is created using frames taken out of videos. This in theory should increase coherence in the images generated and should facilitate to better performance. Since, SSID was published during this past year there has not been used in any new studies and none of the existing story to image generating models have been trained on this dataset. So, instead of drawing inspiration from StoryGAN[1], StoryDALLE[3] and, ARLDM[4] to generate historical images, decided to conduct a survey paper geared towards comparing the performance of these models, on this SSID[6] dataset. Our motivations hence for this course project are :

- **Prove that the story Visualisation models are capable of generalisation**. Since these models have been trained on just 1 real world dataset namely VIST, it does not guarantee the generalisation capability of the models with respect to any real world dataset.
- **More coherent Dataset should result in better results.** Compared to VIST being a collection of random Flickr images made into coherent stories, SSID has been made from real movie and Youtube clips hence actually being part of a story.

2. Literature Review

The work done in the field of story visualisation depends heavily on training dataset used, and there are no publicly available implementation that generalises over any unseen character or story context continuation. StoryGAN [1] started the story visualisation task based on GANs. It consisted of context text encode, image generator, separate image and image discriminator. These were successful in keeping image consistency. Then there were various papers which implemented techniques like copy mechanism using attention[2],Word-Level Story visualisation focusing on text inputs,and propose to use structured input and sentence representation to better guide visual story generation. The two most successful and recent works were StoryDALL-E[3] and ARLDM[5]. StoryDALL-E uses story continuation, taking input a source image of the first scene and retro fitting a pre-trained transformer DALL-E. Auto-Regressive Latent Diffusion Model comes really close to real world story telling by exploiting Diffusion Models and VIST[7] dataset and also propose adaptation method to allow itself to generalise to unseen character. For evaluation of ARLDM[5] over varying datasets, they had used Fréchet Inception Distance (FID), which captures the similarity of generated images to real ones better than the Inception Score, as a quantitative metric.

3. Method

We primarily address two tasks for ARLDM: story continuation and story visualization. The goal of story visualization is to combine a number of images to tell a tale with several sentences. With the same objective as story visualization, story continuation is a variation that is based on a source frame, or the first frame. By addressing the problems of limited information and generalization in story visualization, this setting enables models to produce more relevant and cohesive visuals. For the model to be history aware, the CLIP and BLIP layers are used.

The core of it works on Diffusion model.AR-LDM leverages the history captions and images for future frame generation. For a certain story with a length of L, let $C =$

$[c_1, , c_j, , c_L]$ be input captions and $X = [x_1, , x_j, , x_L]$ be the image targets, each caption c_j is corresponding to an image $x_j \in R^{C \times H \times W}$. Existing implementation treats each frame independently but ARLDM removes this by additionally conditioned on history images $x_{<j}$and directly estimating the posterior based on chain rule.

AR-LDM efficiently conducts both forward and reverse diffusion processes within a low-dimensional latent space. This latent space is designed to be perceptually equivalent to the high-dimensional RGB space, effectively eliminating redundant and semantically meaningless pixel information. The model employs perceptual compression stages, involving encoding (E) and decoding (D) processes, to transform real data into latent space and vice versa, ensuring that the reconstructed data closely approximates the original (D(E(x)) ≈ x). During the diffusion process, AR-LDM utilizes latent representations (z = E(x)) instead of pixel values, and the final output can be decoded back into pixel space with D(z). This two-stage perceptual compression strategy selectively removes imperceptible details, enabling the model to achieve competitive generation results with significantly lower training and inference costs.

The conditioning network comprises CLIP and BLIP, responsible for the current caption encoding and the previous caption-image encoding, respectively. BLIP undergoes pre-training in vision-language tasks using extensive filtered clean web data. In contrast to CLIP, which straightforwardly concatenates unimodal embeddings, BLIP employs a cross-attention module to intricately integrate visual and language modalities. This capability enables BLIP to anchor the entities generated in historical frames, empowering the generative network to reference past scenes. For StoryGan and StoryDall-E, since the models are old and are not optimized, and the lack of time we were not able to come up with viable results. For StoryDall-E, we were able to pipeline it but the we ran out of memory for it to run even after trying to run it on both CHTC and GCP. And for the StoryGan, the pipelining, optimizing and, training for the particular dataset, we were able to make a lot of progress but it still requires time for it to fully optimize and train as there is a certainity that we would've ran out of memory again and hence we are unable to do it due to lack of time.

4. Experiment and Results

For the experiment we used the SSID dataset. This dataset contains 17365 images with 256x256 resolution. This dataset 3473 groups of 5 images with 17895 ground truths. For our experiment we froze the BLIP CLIP layers and used 2 A100 GPUs for a total of 40GB RAM. Trained on AdamW for 5 epochs. The paper trained for 50 epochs for 2 days and unfrozen BLIP and CLIP layer which took 2 days. Our 5 epochs took approximately 6 hours. We also resized images to 256×256 that saved some VRAM. using the

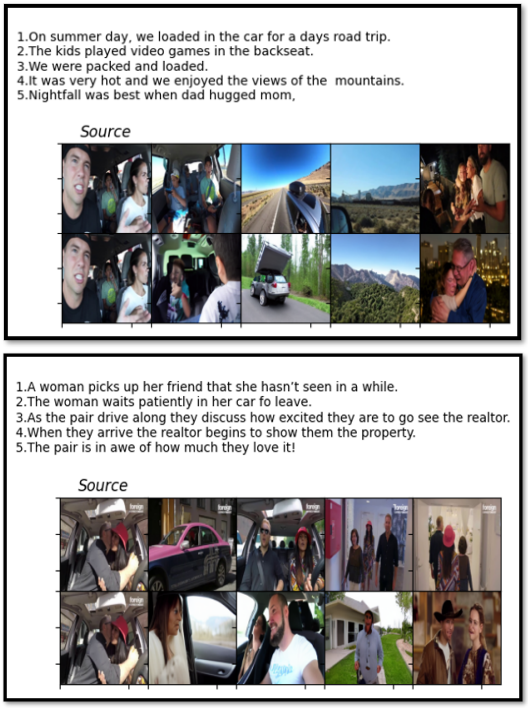

Figure 1. Story Continuation

AdamWoptimizer with an initial learning rate of 10^{-5} and a weight decay of 10^{-4}. During training, we randomly drop the condition phi at a probability of 0.1 for each frame. A cosine scheduler and 8000 steps linear learning rate warm-up are used during training. During inference, we sample images using the DDIM scheduler for 250 inference steps with guidance scale set to 6.0.

The result shown in the figures below have 2 sets of 5 images. The top row of the images are the ground truth which has already been provided to us by the dataset and the second row is the generated images.

The story generated by AR-LDM, for story continuation is depicted in fig 1. For evaluation the authors have used Frechet Inception Distance. It is a popular metric used to evaluate quality and diversity of generated images. It uses a pretrained neural network in our case Inception net for computing feature vectors from original and generated data. These feature vectors represent high level features learned by Nueral Network and used to characterise the distribution of the image. FID is based on mean and covariance of these vectors. Lower the FID better it is.

The FID score generated for the AR-LDM for the SSID dataset is compared with the scores generated with existing other models for different datasets and is shown in table 1.

Fig 2 depicts one of the story generated for story of visualization where we can see that the first picture of the generated picture and the ground truth are the same.

The FID score obtained for AR-LDM for the story con-

Model Name	# parameters	Pororo-SV	Flintstone-SV	VIST-SIS	VIST-DIL	SSID
StoryGanc	-	70.63	90.29	-	-	-
StoryDall-E (Prompt Tuning)	1.3B	61.23	53.71	-	-	-
StoryDall-E	1.3B	24.90	26.5	-	-	-
MEGA-StoryDall-E	2.8B	23.48	24.48	20.98	24.61	-
Ar-LDM	1.5B	17.40	19.28	16.95	17.03	52.39

Table 1. AR-LDM(Story Visualization)

Model Name	FID Score
StoryGan	158.06
CP-CSV	149.28
DUCO-StoryGan	96.51
VLC-StoryGan	84.96
VP-CSV	65.51
Word-Level-SV	56.08
AR-LDM(VIST)	16.59
AR-LDM(SSID)	52.39

Table 2. AR-LDM(Story Visualization)

Figure 2. Story Visualization

tinuation model is compared with existing models on different dataset as is shown in table 2.

5. Conclusion

It can be seen from table 1 that AR-LDM on SSID dataset, performs way better than many other models trained on different datasets and we believe that it could perform even better if we had higher memory resources since we had to freeze the CLIP and BLIP layers which are essential for history awareness. It is also observed that story visualization performed better than story continuation which could be because of freezing the history aware layers. The interesting thing to observe is that in story visualization we were able to obtain better FID score than other networks although they are on different datasets. We believe if trained for higher epochs we can get more coherent results on SSID. With these results we can conclude that our motivation to train existing Story Visualisation technique on a more coherent dataset successfully shows, *not only that story visualisation model can be generalised on any real world dataset but also that a more coherent dataset created from continuous videos or movies can lead to better results even with less training.*

The code can be found in Github Link. The model save files were 8GB hence are not uploaded.

All the 696 story results for both visualisation and continuation can be found here GDrive Link

References

[1] Li, Y., Gan, Z., Shen, Y., Liu, J., Cheng, Y., Wu, Y., Carin, L., Carlson, D. and Gao, J., 2019. Storygan: A sequential conditional gan for story visualization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6329-6338).

[2] Li, B., 2022, October. Word-level fine-grained story visualization. In European Conference on Computer Vision (pp. 347-362). Cham: Springer Nature Switzerland.

[3] Maharana, A., Hannan, D. and Bansal, M., 2022, October. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. In Euro-

Anudeep Kumar	Karan Vikyath Veeranna Rupashree	Siddharth Baskar
ARLDM Implementation Data Prep and Pipeline Report writing	StoryDALLE Implementation Data Prep and Pipeline Made PPT	StoryGAN Implementation Data Prep and Pipeline Report Writing

Table 3. Contributions

pean Conference on Computer Vision (pp. 70-87). Cham: Springer Nature Switzerland.

[4] Gong, Y., Pang, Y., Cun, X., Xia, M., Chen, H., Wang, L., Zhang, Y., Wang, X., Shan, Y. and Yang, Y., 2023. TaleCrafter: Interactive Story Visualization with Multiple Characters. arXiv preprint arXiv:2305.18247. 1

[5] Pan, Xichen, Pengda Qin, Yuhong Li, Hui Xue, and Wenhu Chen. "Synthesizing coherent story with auto-regressive latent diffusion models." arXiv preprint arXiv:2211.10950 (2022). 1

[6] Malakan, Zainy M., Saeed Anwar, Ghulam Mubashar Hassan, and Ajmal Mian. "Sequential Vision to Language as Story: A Storytelling Dataset and Benchmarking." IEEE Access (2023). 1

[7] Huang, Ting-Hao, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick et al. "Visual storytelling." In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies, pp. 1233-1239. 2016.

1

1

1