
Supplementary: Intelligent Grimm - Open-ended Visual Storytelling via Latent Diffusion Models

Anonymous Author(s)

Affiliation

Address

email

1 Dataset Statistics

In Section 1.1, we will present additional details of our **StorySalon** dataset. Subsequently, we will demonstrate the statistical observations about the vocabulary of our dataset in Section 1.2.

1.1 Dataset Overview

As described in the main text, our dataset is consisted of three components: the video component, the E-book component, and the synthetic sample component. Our statistical analysis within this section will exclusively concentrate on the video and E-book components, as we will enrich the dataset with novel synthetic samples progressively. The basic version of our dataset (without synthetic data) comprises a total of 2,184 storybooks and 34,860 text-image pairs. Specifically, the video component consists of 1,286 storybooks and 21,778 text-image pairs, whereas the E-book component comprises 898 storybooks and 13,082 text-image pairs. We divide the dataset into train and test sets following a 9 : 1 ratio. Both the video and E-book components are randomly split into train and test sets according to this proportion.

1.2 Vocabulary Statistics

We conduct an analysis on the number of distinct entities featured in the storybooks and text-image pairs. To further check these categories with the most occurrence, we have divided them into more specific sub-categories. For instance, within the *Human* category, we have further classified individuals as *Boy*, *Girl*, *Man*, and *Woman*. This statistic is still about the video and E-book components.

The distribution of the entities are illustrated in Figure 1. Our findings reveal a total of 178 unique categories of main entities. Notably, in our 2,184 storybooks, the five categories with the highest frequency among storybooks include *Human* (928), *Dog* (170), *Cat* (130), *Monkey* (103), and *Bear* (81). Similarly, in our 34,860 text-image pairs, the five categories with the highest occurrence among text-image pairs consist of *Human* (20,571), *Cat* (2,203), *Dog* (2,058), *Bear* (1,162), and *Monster* (1,039). Other categories with a high frequency include *Mouse*, *Pig*, *Rabbit* and so on.

2 Further Experiments

In Section 2.1, we will provide further implementation details about our experiment. Subsequently, in Section 2.2, we will present additional ablation results pertaining to the performance of single-frame model and model with augmented human feedback. Finally, we will present more visualization results obtained from our model in Section 2.3 to exhibit our model’s excellent visual quality.

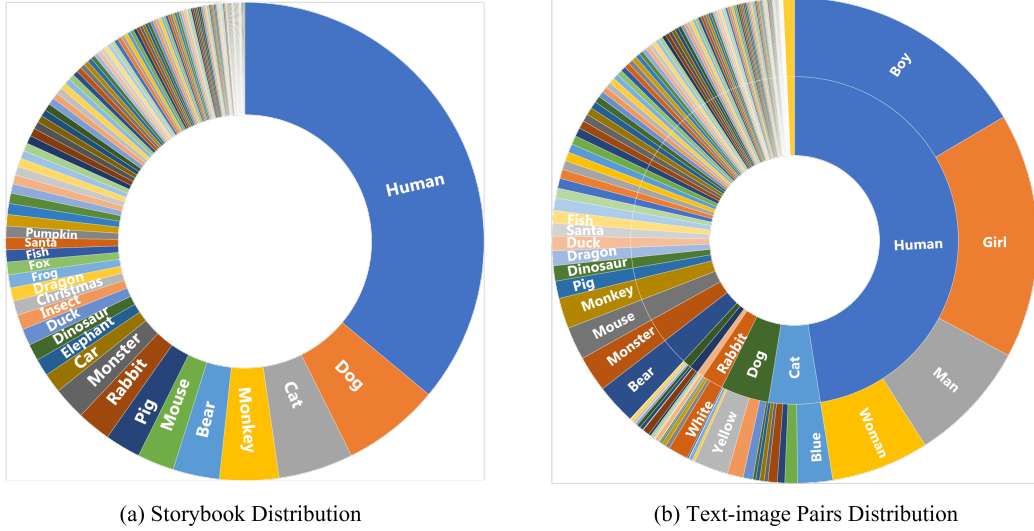


Figure 1: Storybooks and text-image pairs distribution of our StorySalon dataset by main character.

2.1 Implementation Details

Since the detailed training settings have already been presented in the main body of the paper, we will focus on the configuration of our quantitative experiments in this section. The configuration of FID experiments, human evaluation experiments and ablation studies will be introduced respectively in the following paragraphs, with a main focus on how the data is selected.

Fréchet Inception Distance (FID) Experiments. As mentioned in Section 5.2, in the FID experiments, we compare the FID scores [1] between our model and other existing ones, including Stable Diffusion Model (SDM) [4] and **Prompt-SDM**, which conditions on an additional cartoon-style-directed prompt "A cartoon style image". We employ ChatGPT to acquire 100 distinct storylines and employ the three models to generate the corresponding images. The FID score is computed between the distribution of the generated images and the StorySalon test set. For the images generated by SDM and Prompt-SDM, we select the images based on the metric proposed in PickScore [2]. PickScore is a CLIP-based [3] scoring function that has been trained on a dataset consisting exclusively of text-to-image generated images. It has demonstrated exceptional predictive capability in identifying the generated images that align closely with human preferences. Therefore, PickScore is highly appropriate for selecting the results generated by SDM. Each chosen image is selected from a pool of 20 candidates.

Human Evaluation Experiments. As mentioned in Section 5.2, in the human evaluation experiments, we conduct two types of experiments to evaluate the quality of our generated storybooks. We prompt ChatGPT to produce multiple storylines and utilize our StoryGen model along with the two variants of stable diffusion to generate corresponding sequences of images.

In the *first* experiment, we randomly selected an equal number of samples from StorySalon dataset, our StoryGen results, and the generated results of SDM and Prompt-SDM. Each time we randomly sample a visual storybook from these four categories of storybooks and participants are then invited to rate the sample on a score ranging from 1 to 5, taking into account text-image alignment, style consistency, content consistency, and image quality. Higher scores indicate better samples. Similar to the FID experiments mentioned earlier, we employ PickScore to filter the generated results of SDM and Prompt-SDM.

In the *second* experiment, each time we randomly sample a storyline and its three corresponding visual storybooks generated by StoryGen, SDM and Prompt-SDM. Participants are invited to rank and select their preferred generated result among these three different image sequences of the same storyline, based on their personal preference level. To mitigate bias, participants are unaware of the type of storybooks they are evaluating during these two human evaluation experiments. In both experiments, we have invited approximately 30 participants in total.

Model	#Samples	FID ↓
SDM	0	115.43
Prompt-SDM	0	101.23
StoryGen-Single	34,860	73.76
StoryGen	34,860	66.60
StoryGen-HF	35,472 ^{↑612}	66.41
StoryGen-HF (augmented)	38,842 ^{↑3,370}	65.97

Table 1: **Ablation study.** #Samples indicates the number of image-text pair has been used for training.

Ablation Studies. As mentioned in Section 5.2, in the ablation studies, we conduct a comparison of the Fréchet Inception Distance (FID) scores among different models: our StoryGen model with human feedback (**StoryGen-HF**), our StoryGen model without human feedback (**StoryGen**), Prompt-SDM, and SDM. The FID score is calculated by comparing their generation results on the test set of StorySalon with the ground truth present in the test set. The test set is utilized as a group of two image-text pairs and we generate the current frame conditioned on the current text and preceding image. Consequently, we employ PickScore to filter the generation results of all these models.

2.2 Impact of Further Human Feedback and Context Module

In order to demonstrate the efficiency of the context module and fine-tuning with human feedback, we conduct additional ablation studies. We utilize ChatGPT to generate about 700 storylines and synthesize corresponding storybooks. Then we employ PickScore to automatically filter and select around 4,000 high-quality text-image pairs, and fine-tune the model on our StorySalon dataset augmented with these new samples. This refined model, which incorporates the augmented human feedback, is denoted as **StoryGen-HF (augmented)**, distinguishing it from StoryGen-HF mentioned in Section 5.2.

To investigate the impact of the context module on image quality, we also examine the performance of the single-frame model without the context module, marked as **StoryGen-Single**. StoryGen-Single solely incorporates the text module in conjunction with the fine-tuned style transfer module, without integrating the context module to capture context information from preceding frames. Consequently, this represents the model in its immediate state following single-frame pre-training as mentioned in Section 3.2.3. We investigate the performance of StoryGen-HF (augmented) and StoryGen-Single on the test set of StorySalon.

In general, we compare the performance of the single-frame model without the context module against StoryGen with augmented human feedback, using the Fréchet Inception Distance (FID) metric as our evaluation criterion. The results are presented in Table 1.

The findings presented in Table 1 demonstrate that the inclusion of the context module not only enhances the consistency of the context, but also improves the quality of the generated images. Moreover, the process of fine-tuning the model with augmented human feedback leads to further improvements in model performance. Consequently, the establishment of an automated pipeline for incorporating human feedback enables better alignment between the model and human preferences, ultimately elevating the model’s performance to a higher standard.

2.3 Qualitative Results

We provide more visualization samples in this section. As shown in Figure 2, Figure 3, and Figure 4, the results obtained from SDM and Prompt-SDM show a deficiency in maintaining both style and character consistency. StoryGen-Single demonstrates the ability to generate images with similar style, but falls short in preserving the visual attributes of the characters. On the other hand, StoryGen, StoryGen-HF, and StoryGen-HF (augmented) demonstrate a progressive improvement in both consistency and quality, showcasing images that display enhanced coherence and fidelity.

3 Broader Impacts

The potential impact of our visual storytelling model on society lies in several aspects. First, our proposed StoryGen model is a diffusion-based generative model, that will suffer from some common problems of generative models such as data bias: If the training data contains biased or discriminatory content, the model may inadvertently generate narratives and images that reinforce harmful stereotypes or prejudices, leading to negative social implications and reinforcing inequalities.

On the other hand, our storytelling model also has some positive impacts on the industry of creation and education: The widespread application of our visual storytelling model has the potential of inspiring creators and artists to create a large number of visual storybooks rich in basic knowledge, which will have a profound impact on children’s early education, as demonstrated by related work in psychology.

4 Limitations

Training a large-scale foundation generative model requires a large amount of data and computing resources. Although we have established a comprehensive data collection and human feedback augmentation pipeline, and trained StoryGen on the basis of the pre-trained SDM model, its scalability is still limited. In the future, we hope to collect and generate data for training in a more automatic and efficient manner.

In addition, our research on generating coherent visual stories is preliminary, it occasionally lacks robustness in preserving character consistency for objects that are not commonly seen during training. Addressing these challenges is an important focus for our future work, and with the improvement of computing resources, finetuning the generative foundation model, for example, stable diffusion model, may be one of the feasible solutions.

References

- [1] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 2017. 2
- [2] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *arXiv preprint arXiv:2305.01569*, 2023. 2
- [3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 2021. 2
- [4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2



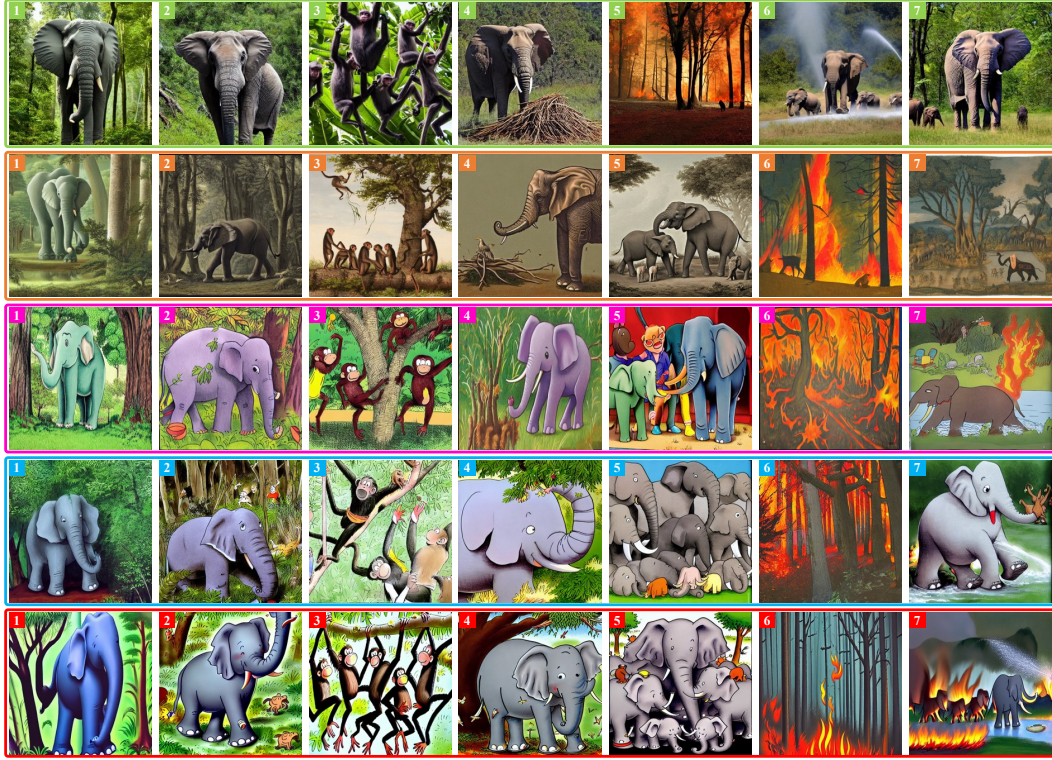
A story of a {white cat}; (1) Once upon a time, in a far-off land, there was a beautiful white cat with big, bright eyes that shone like stars in the night sky. (2) The cat was so pure and spotless that she earned the admiration and love of everyone who saw her. Children would chase after her to pet her soft fur and stroke her tail, while adults would stop in their tracks to admire her beauty. (3) But despite her popularity, the white cat was a quiet and timid creature who kept to herself most of the time. She would spend her days lazing in the sun or exploring the fields and gardens, staying away from trouble and danger. (4) One day, while wandering through the woods, the white cat stumbled upon a magical creature - a unicorn with a golden horn and a shimmering coat. (5) The cat was amazed by the unicorn's beauty and grace and decided to follow her. The unicorn led the cat through a mystical forest with giant mushrooms and glowing fireflies. (6) Eventually, they reached a crystal-clear lake where the cat saw her own reflection in the water and realized just how special she was too. She felt a sense of pride and belonging that she had never experienced before. (7) From that day forward, the white cat was more confident and outgoing. She made new friends and explored new places, certain that her uniqueness was something to be celebrated.

Figure 2: Qualitative Comparison. The images in green, orange, purple, blue, and red boxes are generated by SDM, Prompt-SDM, StoryGen-Single, StoryGen, and StoryGen-HF respectively. The results of our proposed models have exhibited superior style and content consistency, text-image alignment, and image quality.



A story of a {black bear}: (1) Once upon a time, in a dense forest lived a mighty black bear. He was the king of the forest and all the other creatures feared him. The black bear was strong and powerful, with sharp claws and a fierce roar. No one dared to come too close to him. (2) However, the black bear was lonely. He longed for a best friend. One day, he met a small bird who was chirping on a tree branch. The bird was not afraid of the black bear and landed on his paw. (3) The black bear was surprised by the little bird's courage and they quickly became friends. The bird would sit on the bear's head as they explored the forest together. The bear loved listening to the bird's sweet melodies. (4) Years passed and one day, the bird became sick and couldn't sing anymore. (5) The black bear was devastated. He knew he had to do something to help his best friend. So, he searched the forest for medicinal herbs and brought them to the bird's nest. (6) With the bear's care, the bird eventually got better and his beautiful melodies returned. The two friends continued to explore the forest, and the bear realized that true strength comes from the heart, not just from physical power. (7) From that day on, the bear saw all the creatures of the forest as friends rather than potential threats. And despite being the biggest and strongest of them all, the black bear remained humble and kind-hearted.

Figure 3: **Qualitative Comparison.** The images in green, orange, purple, blue, and red boxes are generated by SDM, Prompt-SDM, StoryGen-Single, StoryGen, and StoryGen-HF respectively. The results of our proposed models have exhibited superior style and content consistency, text-image alignment, and image quality.



A story of an {elephant}: (1) Once upon a time, in a lush green forest, there lived a giant elephant. He was the biggest and strongest of all the animals in the forest. His thick grey skin was tough and his trunk could stretch up to ten feet long! (2) The elephant loved to roam around the forest, trumpeting loudly and shaking trees with his powerful tusks. He was a gentle giant who never hurt any of the smaller creatures living in the forest. (3) One day, a group of monkeys came to the elephant and asked him to help them reach a bunch of ripe bananas hanging high on a tree. The elephant, being kind and helpful, reached for the bananas with his long trunk and handed them over to the monkeys. (4) The next day, a group of birds came to the elephant, asking for his help. They were building their nests, but were having trouble finding enough twigs to complete them. The elephant happily used his trunk to gather twigs and helped the birds build their nests. (5) As the days passed, more animals came to the elephant for help. He never turned them away and always helped them out with a kind gesture. (6) One day, the forest caught fire. All the animals were scared and didn't know what to do. The elephant, being the largest and strongest of them all, took charge. He gathered all the other animals and led them to a nearby lake. (7) The elephant used his trunk to suck up water and sprayed it over the fire to extinguish it. All the animals cheered and thanked the elephant for saving their lives.

Figure 4: Qualitative Comparison. The images in green, orange, purple, blue, and red boxes are generated by SDM, Prompt-SDM, StoryGen-Single, StoryGen, and StoryGen-HF respectively. The results of our proposed models have exhibited superior style and content consistency, text-image alignment, and image quality.