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

Anonymous Author(s)

Affiliation

Address

email

Abstract

Benefiting from the impressive diffusion models, conditional generative models have exhibited exceptional capabilities in various generation tasks, for example, image or short video generation based on text description. In this work, we focus on the task of generating a series of coherent images based on a given storyline, denoted as *open-ended visual storytelling*. We make the following three contributions: (i) to fulfill the task of visual storytelling, we introduce two modules into a pre-trained stable diffusion model, and construct an auto-regressive image generator, termed as **StoryGen**, that enables to generate the current frame by conditioning on a text prompt and preceding frame; (ii) to train our proposed model, we collect paired image and text samples by sourcing from various online sources, such as videos, E-books, and establish a data processing pipeline for constructing a diverse dataset, named **StorySalon**, with a far larger vocabulary than existing animation-specific datasets; (iii) we adopt a three-stage curriculum training strategy, that enables style transfer, visual context conditioning, and human feedback alignment, respectively. Quantitative experiments and human evaluation have validated the superiority of our proposed model, in terms of image quality, style consistency, content consistency, and visual-language alignment. We will make the code, model, and dataset publicly available to the research community.

1 Introduction

"Mirror, mirror, here I stand! Who is the fairest in the land?"

— *Grimms' Fairy Tales*

This paper explores an exciting yet challenging task of *visual storytelling*, with the goal of training a model that can effectively capture the relationship between visual elements in images and their corresponding language descriptions, to generate a sequence of images that tell a visual coherent story. The ultimate goal is to generate a sequence of images that tell a coherent story, provided in the form of natural language. The outcome of this task has significant potential for education, providing children with an engaging and interactive way to learn complex visual concepts, and develop imagination, creativity, emotional intelligence, and language skills, as evidenced by the research in psychology [4, 41].

In the recent literature, there has been significant progress in image generation, particularly with the guidance of text, such as stable diffusion [35], DALL-E [33] and Imagen [9]. However, these models are not sufficient for visual storytelling for two reasons: (i) existing models generate images independently without considering the context of previous frames or the overall narrative, resulting in visual inconsistencies and a lack of coherence in the visual story; (ii) generating images by only conditioning on text can lead to ambiguities or require unnecessarily long descriptions, particularly

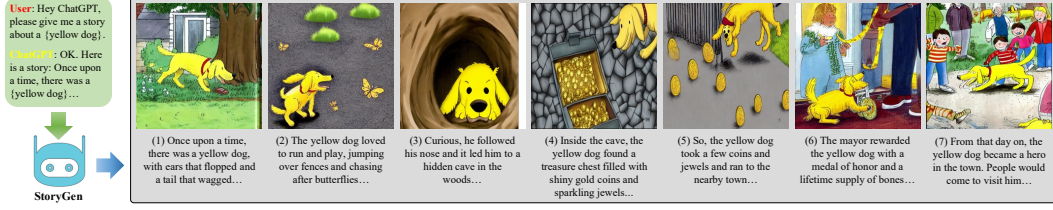


Figure 1: An illustration of open-ended visual storytelling. In practise, one user can prompt a large language model, for example, ChatGPT, to generate a unique and engaging story, which is then fed into our proposed **StoryGen** model, to generate a sequence of images that are not only aligning to the given storyline, but also coherent. We recommend the reader to zoom in and read the story.

when dealing with subtle differences, for example, distinguishing between animals from the same category but different breeds. To address the limitations, we introduce a novel auto-regressive architecture, termed as **StoryGen**, that builds upon pre-trained stable diffusion model, with two extra modules serving for style transfer and visual context conditioning. At inference time, StoryGen takes the preceding frames and text prompts as conditions to synthesize the current frame, *i.e.*, iteratively creating visual sequences that are not only aligning to language description, but also coherent.

In practise, visual storytelling is faced with significant challenges due to the lack of high-quality image-text data. Most existing works are limited to train on a few specific animations, for example, StoryGAN [17], StoryDALL-E [22] and AR-LDM [28], resulting in a small and restricted vocabulary and characters. To overcome this limitation, we have created a dataset called **StorySalon**, that features a rich source of coherent images and stories, primarily comprising children’s storybooks collected from three different sources: videos, E-books, and synthesized sample data from our StoryGen model with human verification. As a result, our dataset includes a diverse vocabulary with different characters, storylines, and artistic styles.

We follow a three-stage training procedure: *Firstly*, we insert a LoRA-like architecture into text conditioning module in stable diffusion model, *i.e.*, on top of the image-to-text cross-attention, in order to adapt the pre-trained diffusion model on our collected dataset; *Secondly*, we introduce a visual context module that enables the generation process to condition on preceding image; *Lastly*, we finetune the model with data after human verification, *i.e.*, leveraging the feedback to further align the model with human preference. As a result, the scale and diversity of our collected dataset enable the model to acquire the ability of open-vocabulary visual storytelling, by that we mean, our model can generate new image sequences that are not limited to pre-defined storylines, characters, or scenes. For example, we can prompt a large language model to create unique and engaging stories, then feed into StoryGen for generation, as shown in Figure 1.

To summarise, we make the following contributions in this paper: (i) we propose the task of *open-ended visual storytelling*, that involves generating engaging image sequence that is aligning to a given storyline, for example, written by a large language model; (ii) we develop a novel architecture based on stable diffusion, termed as StoryGen, which can generate image sequences in an auto-regressive manner, taking both preceding image and text prompt of current frame as condition; (iii) we initiate a data collection pipeline and collect a large-scale, diverse datasets of storybooks, from online videos, E-books and synthesized samples, including paired image-text samples of a diverse vocabulary with different characters, storylines, and artistic styles; (iv) we adopt a three-stage curriculum training strategy, that enables style transfer, visual context conditioning, and human feedback alignment, respectively. Experimentally, we conduct both quantitative comparison and human evaluation, showing that the outputs from our proposed model are more preferred, in terms of image quality, style consistency, and image coherence.

2 Related Works

Diffusion Models learn to model a data distribution via iterative denoising and are trained with denoising score matching. Drawing from the principles of Langevin dynamics and physical diffusion processes, diffusion models have undergone refinement through a multitude of works [39, 27, 45]. Notably, DDPM [10] has demonstrated improved performance over other generative models, while DDIM [40] has significantly boosted generation efficiency. In view of their superior generative

capabilities, diffusion models have found extensive utility in various downstream applications besides image generation, such as video generation [46, 5, 12, 9, 38], image manipulation [2, 24, 14, 7], grounded generation [18], 3D texturing [34], and image inpainting [47, 26, 19, 1].

Text-to-image Generation involves the creation of images from textual descriptions. The task has been tackled using various generative models, with Generative Adversarial Networks (GANs) [6] being the first widely-used model. Several GAN-based architectures such as StackGAN [50], StackGAN++ [51], and AttnGAN [48] have achieved notable success in this area. Additionally, pre-trained auto-regressive transformers [43] such as DALL-E [33] have demonstrated the ability to generate high-quality images in response to textual prompts. Recently, diffusion models have emerged as a popular approach to text-to-image generation. New images can be sampled under text guidance from the data distribution learned by diffusion models with iterative denoising process. DALL-E 2 [32] leverages CLIP [29] features to achieve well-aligned text-image latent space, while Imagen [37] relies on large language models like T5 [31] to encode text. Stable Diffusion (or Latent Diffusion) Model [35] performs diffusion process in latent space, and it can generate impressive images after pre-training on a large-scale text-image datasets.

Story Synthesis is first introduced as the task of story visualization (SV) by StoryGAN [17], which presents a GAN-based framework and the inaugural dataset named Pororo, derived from cartoons. Subsequently, some other works also follow the GAN-based framework, such as DUCO-StoryGAN [21] and VLC-StoryGAN [20]. In the case of word-level SV [16], more emphasis is placed on the representation of text, whereas VP-CSV [3] employs VQ-VAE [42] and a transformer-based language model to conserve character appearance and enhance visual quality. StoryDALL-E [22] extends the story visualization task to story continuation with the initial image given and recommends using a pre-trained DALL-E generative model [33] to produce coherent images. AR-LDM [28] introduces an auto-regressive latent diffusion model that can generate highly realistic images, but with only a limited vocabulary. NUWA-XL [49] is a concurrent work that exploits hierarchical diffusion model to synthesize long videos, with the keyframes generated first, followed by frame interpolation.

Existing models are mostly developed on specific scenes, which limits their ability for generating image sequences for diverse stories. In this paper, we target for more ambitious applications, to develop an open-ended visual storytelling model, that can digest stories of arbitrary length and diverse topics, and synthesize a sequence of coherent images in terms of both style and semantic.

3 Method

To be self-contained, we first present a brief overview to diffusion model in Section 3.1; then, we detail our proposed model for storybook generation in Section 3.2, starting from problem formulation, then architecture details, and lastly on training details.

3.1 Preliminaries on Diffusion Models

Diffusion models are a type of generative models that undergo a denoising process, converting input noise into meaningful data samples. Diffusion models comprise a forward diffusion process that incorporates Gaussian noise into an image sample $\mathbf{x}_0$, accomplished via a Markovian process over T steps. If we denote the noisy image at step t as $\mathbf{x}_t$, the transition function $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ connecting $\mathbf{x}_{t-1}$ and $\mathbf{x}_t$ can be expressed as follows:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \quad q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^T q(\mathbf{x}_t|\mathbf{x}_{t-1}) \quad (1)$$

where $\beta_t \in (0, 1)$ is the variance schedule controlling the step size.

Using Gaussian distribution property and reparametrization, if we define $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$, we can write equation 1 as follows:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1 - \bar{\alpha}_t)\mathbf{I}) \quad (2)$$

Diffusion models also comprise a reverse diffusion process that learns to restore the initial image sample from noise. A UNet-based model [36] is utilized in the diffusion model to learn the reverse

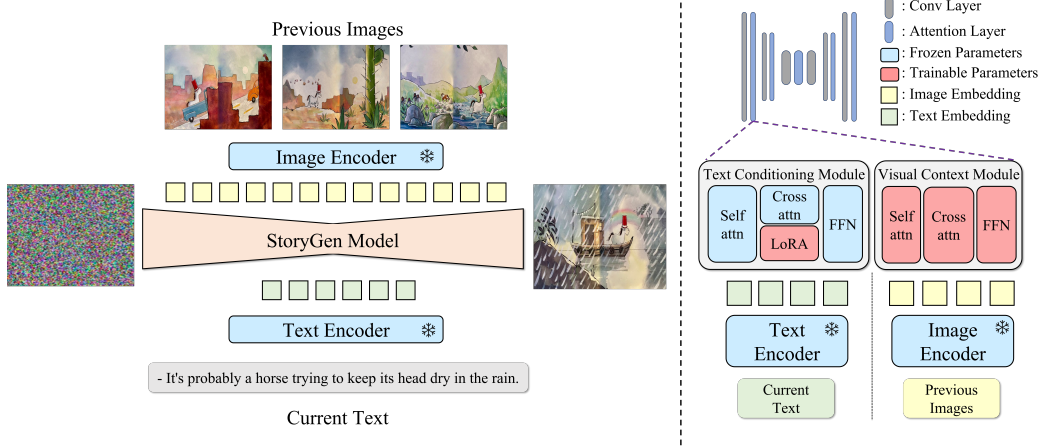


Figure 2: **Architecture Overview.** The **left** figure illustrates the complete procedure of *visual storytelling*. Our StoryGen model utilizes contextual information from previous frame and the text description at current step, to generate an image. The **right** figure displays the structure of our proposed modules, (i) style transfer module that is inserted into the text-conditioning module, with a LoRA-like achitecture; (ii) visual context module that enables the model to also condition on the features from preceding image for generation.

diffusion process p_θ . The process p_θ can be expressed using the following equation.

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \quad p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)) \quad (3)$$

where $\boldsymbol{\mu}_\theta$ is the predicted Gaussian distribution mean value.

As we compute the loss function by taking the mean absolute error of the noise term ϵ_θ into account, we can express the mean value $\boldsymbol{\mu}_\theta$ in terms of the noise term ϵ_θ as follows:

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \epsilon_\theta(\mathbf{x}_t, t) \right) \quad (4)$$

Therefore, the objective can be written as:

$$\mathcal{L}_t = \mathbb{E}_{t \sim [1, T], \mathbf{x}_0, \epsilon_t} \left[\|\epsilon_t - \epsilon_\theta(\mathbf{x}_t, t)\|^2 \right] \quad (5)$$

3.2 StoryGen Model

In this section, we start by defining the problem for visual storytelling, then we introduce the core components in our proposed architecture, namely, Style Transfer Module and Visual Context Module, lastly, we present details for training the model with a curriculum learning regime.

3.2.1 Problem Formulation

In visual storytelling, our goal is to generate a sequence of coherent and consistent images that correspond to a given story in the form of natural language. To achieve this, we propose an auto-regressive image generation model, called **StoryGen**, that generates the current frame $\mathcal{I}_k$ by conditioning on both current text description $\mathcal{T}_k$ and the previous frame $\mathcal{I}_{k-1}$, as illustrated in Figure 2. The model is formulated as follows:

$$\begin{aligned} \{\hat{\mathcal{I}}_1, \hat{\mathcal{I}}_2, \dots, \hat{\mathcal{I}}_L\} &= \Phi_{\text{StoryGen}}(\{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_L\}; \Theta), \\ &= \psi_{\text{SDM}}(\hat{\mathcal{I}}_1 | \mathcal{T}_1; \theta) \dots \psi_{\text{SDM}}(\hat{\mathcal{I}}_k | \hat{\mathcal{I}}_{k-1}, \mathcal{T}_k; \theta) \dots \psi_{\text{SDM}}(\hat{\mathcal{I}}_L | \hat{\mathcal{I}}_{L-1}, \mathcal{T}_L; \theta) \end{aligned}$$

Here, $\{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_L\}$ represents the given story in text sentences, $\{\hat{\mathcal{I}}_1, \hat{\mathcal{I}}_2, \dots, \hat{\mathcal{I}}_L\} \in \mathbb{R}^{H \times W \times 3}$ denotes the generated storybook, H, W refer to the width and height, respectively. $\psi_{\text{SDM}}(\cdot)$ refers

to a semi-frozen stable diffusion model (SDM), with a small number of newly-added trainable parameters (θ). It takes randomly sampled gaussian noise, text description and preceding image as input, and generate coherent image sequence that align with the story’s narrative. In the following sections, we present the architecture detail for one-step generation conditioned on text and image.

3.2.2 Architecture Details

Generally speaking, our model is built upon the foundation of a pre-trained stable diffusion model (SDM), that has been pre-trained on large number of paired image-caption samples, to gradually transform the noisy latent into an image. To tackle the problem of open-ended storytelling, we introduce two computational modules, namely, Style Transfer Module, and Visual Context Module, that enables the model to condition on not only text descriptions, but also the preceding RGB image, as shown in Figure 2. Formally, we can express the generation procedure as:

$$\mathcal{I}_k = \psi_{\text{SDM}}(\hat{\mathcal{I}}_k | \hat{\mathcal{I}}_{k-1}, \mathcal{T}_k) = \psi_{\text{SDM}}(\mathbf{x}, \phi_{\text{text}}(\mathcal{T}_k), \phi_{\text{vis}}(\hat{\mathcal{I}}_{k-1}))$$

where $\mathbf{x}$, $\phi_{\text{text}}(\cdot)$ and $\phi_{\text{vis}}(\cdot)$ denote the noisy latent, encoded text description and preceding image.

Style Transfer Module. To steer a pre-trained stable diffusion model towards the style of children’s storybooks, we propose to insert a lightweight, LoRA-like [13] architecture into the text conditioning module, effectively acting as style transfer. This can be expressed as:

$$Q = W^Q \cdot \mathbf{x} + \Delta Q, K = W^K \cdot \mathcal{C}_k^{\text{text}} + \Delta K, V = W^V \cdot \mathcal{C}_k^{\text{text}} + \Delta V, \text{ where } \mathcal{C}_k^{\text{text}} = \phi_{\text{text}}(\mathcal{T}_k)$$

W^Q, W^K and W^V denote the projection matrices, adopted from the text conditioning module in pre-trained stable diffusion model. $\phi_{\text{text}}(\cdot)$ and $\mathcal{C}_k^{\text{text}}$ refer to the pre-trained CLIP text encoder and extracted text embedding respectively. $\Delta Q, \Delta K$ and ΔV are calculated by a learnable projection of $\mathbf{x}, \mathcal{C}_k^{\text{text}}$ and $\mathcal{C}_k^{\text{text}}$, respectively, resembling LoRA operations.

Visual Context Module. In order to generate visually coherent images, we insert a visual context module after the text conditioning, specifically, it is a transformer decoder comprising a self-attention layer, a cross-attention layer, and a feed-forward network, where the cross-attention layer employs a casual attention mechanism by using the noisy latent as query and the visual features of the previous frame as key and value, which can be formally denoted as:

$$Q = W^Q \cdot \mathbf{x}, K = W^K \cdot \mathcal{C}_k^{\text{vis}}, V = W^V \cdot \mathcal{C}_k^{\text{vis}}, \text{ where } \mathcal{C}_k^{\text{vis}} = \phi_{\text{vis}}(\mathcal{I}_{k-1})$$

W^Q, W^K and W^V refer to three learnable projection matrices, $\phi_{\text{vis}}(\cdot)$ denotes the visual feature extracted by a pre-trained CLIP visual encoder. It is worth noting that the visual context module can also extend to multiple condition frames by concatenating their CLIP features as visual contexts.

Training Objective. At training stage, we randomly sample a triplet each time, *i.e.*, $\{\mathcal{I}_k, \mathcal{I}_{k-1}, \mathcal{T}_k\}$, and the objective defined in Equation 5 can now be transformed into:

$$\mathcal{L}_t = \mathbb{E}_{t \sim [1, T], \mathbf{x}_0, \epsilon_t, \mathcal{C}_k^{\text{vis}}, \mathcal{C}_k^{\text{text}}} \left[\|\epsilon_t - \epsilon_{\theta}(\mathbf{x}_t, t, \mathcal{C}_k^{\text{vis}}, \mathcal{C}_k^{\text{text}})\|^2 \right] \quad (6)$$

and as we adopt classifier-free guidance [11] in inference, the predicted noise can be expressed as:

$$\bar{\epsilon}_{\theta}(\mathbf{x}_t, t, \mathcal{C}_k^{\text{vis}}, \mathcal{C}_k^{\text{text}}) = (w + 1)\epsilon_{\theta}(\mathbf{x}_t, t, \mathcal{C}_k^{\text{vis}}, \mathcal{C}_k^{\text{text}}) - w\epsilon_{\theta}(\mathbf{x}_t, t) \quad (7)$$

where w is the guidance scale.

3.2.3 Curriculum Learning

In this section, we describe the three-stage training strategy, that includes single-frame pre-training, multiple-frame fine-tuning, and alignment with human feedback. This curriculum learning approach enables the model to gradually learn from simple to complex tasks, ultimately improving its ability to generate high-quality images that align with the given story narratives. Details for our proposed curriculum learning are presented below.

Single-frame Pre-training. We start by training the style transfer module, which has been inserted into the text conditioning module of a pre-trained stable diffusion model, in a single-frame manner. At this pre-training stage, we do not introduce the visual context module, and freeze all other parameters except for the LoRA-like plug-in. This training approach allows us to quickly adjust to the desired

181 visual style and character appearance in storybooks, while also maintaining the generation ability of
182 the pre-trained stable diffusion model.

183 **Multiple-frame Fine-tuning.** Here, we fine-tune the visual context module while freezing other
184 parameters of the model. Till this point, this allows the generation procedure to utilize information
185 from either the text description or the preceding frame. To avoid over-fitting to the text descriptions,
186 we adopt a technique inspired by BERT training [15], randomly dropping some words in the text with
187 certain probability. The entire visual context module is fine-tuned during this stage.

188 **Fine-tuning with Human Feedback.** After multiple-frame fine-tuning, the model has developed
189 basic storybook generation capabilities, to avoid from generating, potentially scary, toxic or biased
190 content, we also propose to align the model with human preference. Specifically, we prompt ChatGPT
191 to generate approximately 200 stories and use our model to synthesize images. After manually
192 filtering around 100 high-quality storybooks from this corpus, we add them to the training set for
193 further fine-tuning. As future work, we aim to add more books into this human feedback step.

194 **Inference.** With the three-stage training regime, we can streamline the entire inference process
195 into a unified generation framework. As shown in the Figure 1, at inference time, we can prompt
196 the ChatGPT to generate engaging, yet educational storylines, and synthesize the first image using a
197 single-frame approach with only style transfer module involved; the previously synthesized frames,
198 along with story description at current step, are treated as condition to generate image sequence in an
199 auto-regressive manner. Experimentally, our proposed StoryGen is shown to generate images that
200 align with the storyline, as well as maintaining consistency with previously generated frames.

201 4 Dataset Preparation

202 For training visual storytelling model, we collect a dataset called **StorySalon**, that contains approxi-
203 mately 2K storybooks and more than 30K well-aligned text-image pairs. This dataset is comprised of
204 storybooks with potentially aligned text and image pairs, sourced from three different sources: video
205 data, E-book data, and additional data from human feedback.

206 4.1 Image-text Data from Videos & E-books

207 Here, we elaborate the procedure of extracting paired image-text samples from YouTube videos and
208 E-books (pdf and corresponding audios available).

209 **Visual Frame Extraction.** To begin with, we download a significant amount of videos and subtitles
210 from YouTube, by querying keywords related to children story, for example, *storytime*. We then
211 extract the keyframes from the videos, along with the corresponding subtitles and their timestamps.
212 To remove duplicate frames, we extract ViT features for each frame using pre-trained DINO [23], for
213 the image groups with high similarity score, we only keep one of them. Next, we use YOLOv7 [44]
214 to segment and remove person frames and headshots, as they often correspond to the story-teller and
215 are unrelated to the content of the storybook. Finally, we manually screen out frames that are entirely
216 white or black. Similarly, we also acquire a number of electronic storybooks from the Internet,
217 and extract images from E-book, except for those with extraneous information, for example, the
218 authorship page. We acquire the corresponding text description with Whisper [30] from the audio
219 file. For E-books that do not have corresponding audio files, we use OCR algorithms, to directly
220 recognize the text on each page.

221 **Visual-Language Alignment.** Here, for each image, we acquire two types of text description,
222 namely, story-level description, and visual description. This is based on our observation that there
223 actually exists semantic gap between story narrative and descriptive text, for example, the same image
224 can be well described as *"The cat is isolated by others, sitting alone in front of a village."* in story, or
225 *"A black cat sits in front of a number of houses."* as visual description, therefore, directly finetuning
226 stable diffusion model with story narrative maybe detrimental to its pre-trained text-image alignment.

227 In practise, to get story-level paired image-text samples, we align the subtitles with visual frames by
228 using Dynamic Time Warping (DTW) algorithm [25]. To get visual descriptions, we use ChatCap-
229 tioner [52] to generate captions for each image, as shown in Figure 3. s At training time, this allows
230 us to substitute the original story with more accurate and descriptive captions.



Figure 3: **Dataset Pipeline Overview.** The **left** figure provides an overview of the complete dataset collection pipeline. Unstructured metadata sourced from the Internet undergoes a series of steps including frame extraction, visual-language alignment and image inpainting, resulting in properly aligned image-text pairs. The **right** figure displays examples of video data, E-book data, and synthetic samples. The accompanying texts represent their corresponding textual content, respectively.

231 **Visual Frame Post-processing.** In practice, we discovered that books in the frames can potentially
 232 interfere with our image generation model by having story texts printed on them. To address this, we
 233 use an OCR detector to identify the text regions in the frames and an image inpainting model to fill in
 234 the text and headshot regions. This process results in more precise image-text pairs that can be fed
 235 into the diffusion model.

236 4.2 Additional Data from Human Feedback

237 As outlined in Section 3.2.3, we use the model (trained after two stages) to generate a set of new
 238 storybooks, and incorporate human feedback into the fine-tuning process. Following a rigorous
 239 manual review, we carefully select the best pieces, and add them into the training dataset. This
 240 allows us to continually improve the quality of our model and ensure that it produces engaging, yet
 241 educational storybooks, that align with human’s preference.

242 4.3 Discussion

243 Given the diversity of our data sources and data types, the StorySalon dataset exhibits a significantly
 244 broader range of visual styles and character appearances over other animation-specific datasets.
 245 Moreover, our dataset surpasses others in terms of vocabulary coverage by a substantial margin.
 246 Notably, our texts seamlessly integrate narrative storylines and descriptive visual prompts, ensuring
 247 the preservation of text-image alignment while adapting to the art style of storybooks.

248 5 Experiment

249 In this section, we start by describing our experimental settings, then compare with other models
 250 from three different perspectives: style, quality and coherence with quantitative and subjective human
 251 evaluation. Additionally, we present the results of our ablation experiments to prove the effectiveness
 252 of our proposed training regime.

253 5.1 Training Settings

254 Our model is based on publicly released stable diffusion checkpoints, with a learning rate of 1×10^{-5}
 255 and a batch size of 512. We begin with a single-frame pre-training stage, which involves 10,000
 256 iterations on 8 NVIDIA RTX3090. In the multiple-frame fine-tuning stage, we fine-tune the model for
 257 40,000 iterations using a single condition image. To improve the robustness of the training procedure,
 258 we apply a 10% ~ 30% words dropout with a probability to the texts in the current frames. We also
 259 fine-tune the model with human feedback on the training set with 100 manually generated storybooks
 260 in addition for 5,000 iterations. During inference, we utilize DDIM sampling and classifier-free
 261 guidance with a weight of 6.0.

Model	FID ↓	Alignment ↑	Style ↑	Content ↑	Quality ↑	Preference	Model	FID ↓
GT	-	4.22	4.68	4.34	4.32	-	with HF	66.41
StoryGen	120.01	4.02	3.82	3.67	3.53	70.42%	without HF	66.60
Prompt-SDM	167.21	3.45	2.25	2.64	3.21	16.11%	Prompt-SDM	101.23
SDM	184.01	3.28	2.56	2.57	3.49	13.47%	SDM	115.43

Table 1: **Comparison result of human evaluation and FID.** GT stands for the ground truth from the training set. SDM denotes Stable Diffusion and Prompt-SDM denotes SDM with cartoon-style-directed prompts.

Table 2: Ablation study on human feedback.

5.2 Quantitative Results

To evaluate the quality of our generated image sequence, we adopt the widely-used Fréchet Inception Distance (FID) score [8]. However, as there is no standardized metric for evaluating the consistency of images, we include human evaluation for comparison.

Fréchet Inception Distance (FID). We present a comparison of the FID scores between our model and other existing ones, including SDM and Propmt-SDM, which conditions on an additional cartoon-style-directed prompt "A cartoon style image". Specifically, we calculate the FID scores between the distribution of the generated results from these models and the distribution of the our proposed **StorySalon** testset. As shown in Table 1, we evaluate the generated image sequences from 100 storylines, obtained by prompting ChatGPT. Our StoryGen model outperforms the original stable diffusion models (SDM) and Prompt-SDM by a large margin, demonstrating the effectiveness of our model in generating high-quality coherent images.

Human Evaluation. We conduct two types of human evaluation experiments to assess the quality of our generated storybooks. In the *first* experiment, we randomly select an equal number of groundtruth storybooks, the results of our StoryGen and the generation results of SDM and Propmt SDM. Participants are then invited to rate these four categories of storybooks 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. In the *second* experiment, we prompt ChatGPT to produce a number of storylines and use our StoryGen along with the two variations of stable diffusion to generate corresponding image sequences. Participants are asked to choose the preferred results of each storyline. 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.

Table 1 presents the results of our human evaluation. As can be seen, our model has shown significant performance improvement in its overall score compared to stable diffusion models, especially in terms of consistency and alignment, indicating that it can generate images that are highly consistent with the given text prompts and visual contexts, thus better exploiting the contextual information.

Ablation Studies. Our study compares the performance of our model with and without fine-tuning using human feedback. We evaluate the FID score of between our generated image sequence and the test set of our StorySalon dataset. The results demonstrate that fine-tuning with human feedback slightly improves performance, we conjecture that this could be attributed to the fact that the number of human-verified samples are limited, due to a result of resource limitation. In future work, we intend to address this quantity constrain by continually augmenting the dataset with new samples and closely monitoring the progressive advancements.

5.3 Qualitative Results

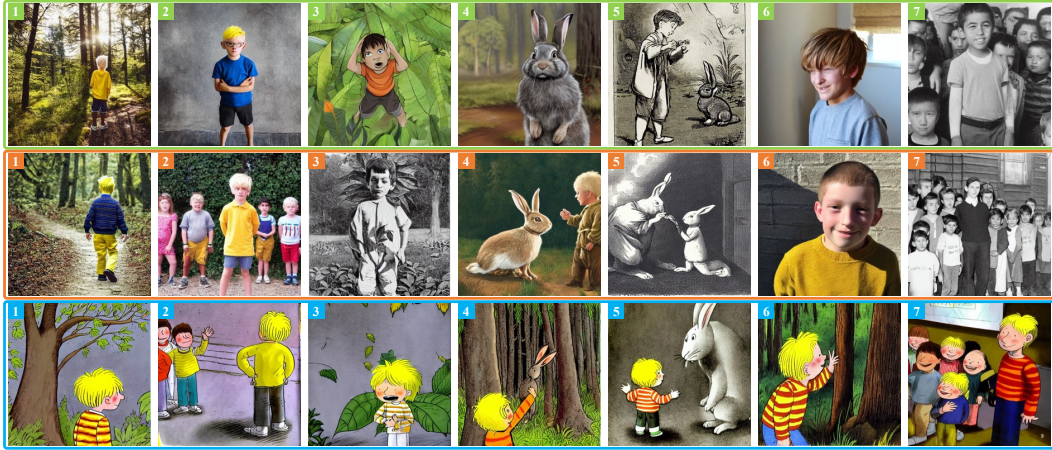
In Figure 4, we present the visualization results, showing that our model can generate storybooks with a broad vocabulary, while maintaining coherence and consistency throughout the narrative. The generated images successfully maintain the consistency of the artistic style and character appearance, whereas the results from SDM and Prompt-SDM fail to do so. Moreover, style of the generated results from SDM’s are also incongruent with the requirements of visual storytelling for childrens.

6 Conclusion

In this paper, we consider the exciting yet challenging task known as *open-ended visual storytelling*, which involves generating a sequence of consistent images that tell a coherent visual story based



(a) A story of a {white dog}: (1) Once upon a time, in a small village, there lived a white dog. It had pure white fur that sparkled in the sunlight. (2) The white dog was loved by all the villagers. Children would play with it, and adults would often walk it around the village. (3) One day, while the white dog was taking a walk, it heard a cry for help. It followed the sound and found a young boy who had fallen into a deep pit. The white dog quickly sprang into action and started barking loudly to get the attention of the villagers. (4) Within minutes, a group of villagers gathered around the pit and lowered a rope to rescue the boy. (5) From that day onwards, the white dog was regarded as a hero in the village. (6) Eventually, the white dog became old. But even in its old age, the white dog would still wag its tail whenever it saw a child or heard someone call its name. (7) The white dog passed away peacefully, surrounded by the love and affection of the entire village.



(b) A story of a {boy with yellow hair}: (1) Once upon a time, there was a boy with yellow hair as bright as the sun. (2) The boy loved his yellow hair, even though other children teased him for looking different. (3) One day, as the boy was walking through the forest, he stumbled upon a magical vase. When he touched the vase, he suddenly shrunk down to the size of a bug. (4) Scared and confused, the boy tried to find his way back to his normal size. He wandered through the forest, until he came across a wise old rabbit. (5) The rabbit taught him a special spell to grow back to his normal size. The boy was grateful for the rabbit's kindness. (6) As the boy regained his size, he realized that his yellow hair had grown even brighter because of the magical vase. He ran back to his village, excited to show off his hair. (7) But when the children saw his hair, they no longer wanted to tease him. They were amazed by the boy's bold and beautiful yellow hair. From that day forward he had grown to understand that being different is what makes us special, and he embraced his uniqueness with pride.

Figure 4: **Qualitative Comparison with other baselines.** The images in green, orange and blue boxes are generated **SDM**, **Prompt-SDM** and **StoryGen** respectively. Our results have superior style and content consistency, text-image alignment, and image quality.

on the given storyline. Our proposed **StoryGen** architecture can take input from the preceding frame along with the text prompt to generate the current frame in an auto-regressive manner. A three-stage curriculum training strategy has been introduced for effective training and alignment with human preference. Due to the limitations of dataset in previous works, we have also collected a large-scale, diverse dataset named **StorySalon** that includes paired image-text samples sourced from storybook data from videos, e-books and synthesized samples. The StorySalon dataset possesses a diverse vocabulary of storyline, character appearances and artistic styles. While comparing with stable diffusion models with quantitative experiment and human evaluation, our proposed model substantially outperform existing models, from the perspective of image quality, style consistency, content consistency, and image-language alignment.

References

- [1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 3
- [2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. 3
- [3] Hong Chen, Rujun Han, Te-Lin Wu, Hideki Nakayama, and Nanyun Peng. Character-centric story visualization via visual planning and token alignment. *arXiv preprint arXiv:2210.08465*, 2022. 3
- [4] K. Dickinson David, A. Griffith Julie, Golinkoff Roberta, Michnick, and Hirsh-Pasek Kathy. How reading books fosters language development around the world. *Child Development Research*, 2012. 1
- [5] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. *arXiv preprint arXiv:2302.03011*, 2023. 3
- [6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 2020. 3
- [7] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3
- [8] 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. 8
- [9] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1, 3
- [10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020. 2
- [11] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5
- [12] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 3
- [13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 5
- [14] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022. 3
- [15] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. 6
- [16] Bowen Li. Word-level fine-grained story visualization. In *Proceedings of the European Conference on Computer Vision*, 2022. 3
- [17] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2, 3
- [18] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Guiding text-to-image diffusion model towards grounded generation. *arXiv preprint arXiv:2301.05221*, 2023. 3
- [19] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 3
- [20] Adyasha Maharana and Mohit Bansal. Integrating visuospatial, linguistic, and commonsense structure into story visualization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021. 3

- [21] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. Improving generation and evaluation of visual stories via semantic consistency. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021. 3
- [22] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. In *Proceedings of the European Conference on Computer Vision*, 2022. 2, 3
- [23] Caron Mathilde, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision*, 2021. 6
- [24] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021. 3
- [25] Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, 2007. 6
- [26] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3
- [27] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 2021. 2
- [28] Xichen Pan, Pengda Qin, Yuhong Li, Hui Xue, and Wenhui Chen. Synthesizing coherent story with auto-regressive latent diffusion models. *arXiv preprint arXiv:2211.10950*, 2022. 2, 3
- [29] 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. 3
- [30] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022. 6
- [31] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 2020. 3
- [32] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3
- [33] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 2021. 1, 3
- [34] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. *arXiv preprint arXiv:2302.01721*, 2023. 3
- [35] 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. 1, 3
- [36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III* 18, 2015. 3
- [37] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 2022. 3
- [38] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 3
- [39] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2015. 2

- 416 [40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint*
417 *arXiv:2010.02502*, 2020. 2
- 418 [41] Gabrielle A. Strouse, Angela Nyhout, and Patricia A. Ganea. The role of book features in young children’s
419 transfer of information from picture books to real-world contexts. *Frontiers in Psychology*, 2018. 1
- 420 [42] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural*
421 *information processing systems*, 2017. 3
- 422 [43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz
423 Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*,
424 2017. 3
- 425 [44] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies
426 sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022. 6
- 427 [45] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings*
428 *of the 28th international conference on machine learning (ICML-11)*, 2011. 2
- 429 [46] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, Wynne Hsu, Ying Shan, Xiaohu
430 Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video
431 generation. *arXiv preprint arXiv:2212.11565*, 2022. 3
- 432 [47] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided
433 object inpainting with diffusion model. *arXiv preprint arXiv:2212.05034*, 2022. 3
- 434 [48] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He.
435 Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In
436 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3
- 437 [49] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang,
438 Linjie Li, Shuguang Liu, Fan Yang, et al. Nuwa-xl: Diffusion over diffusion for extremely long video
439 generation. *arXiv preprint arXiv:2303.12346*, 2023. 3
- 440 [50] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N
441 Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks.
442 In *Proceedings of the International Conference on Computer Vision*, 2017. 3
- 443 [51] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N
444 Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE*
445 *transactions on pattern analysis and machine intelligence*, 2018. 3
- 446 [52] Deyao Zhu, Jun Chen, Kilichbek Haydarov, Xiaoqian Shen, Wenxuan Zhang, and Mohamed Elhoseiny.
447 Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. *arXiv preprint*
448 *arXiv:2303.06594*, 2023. 6