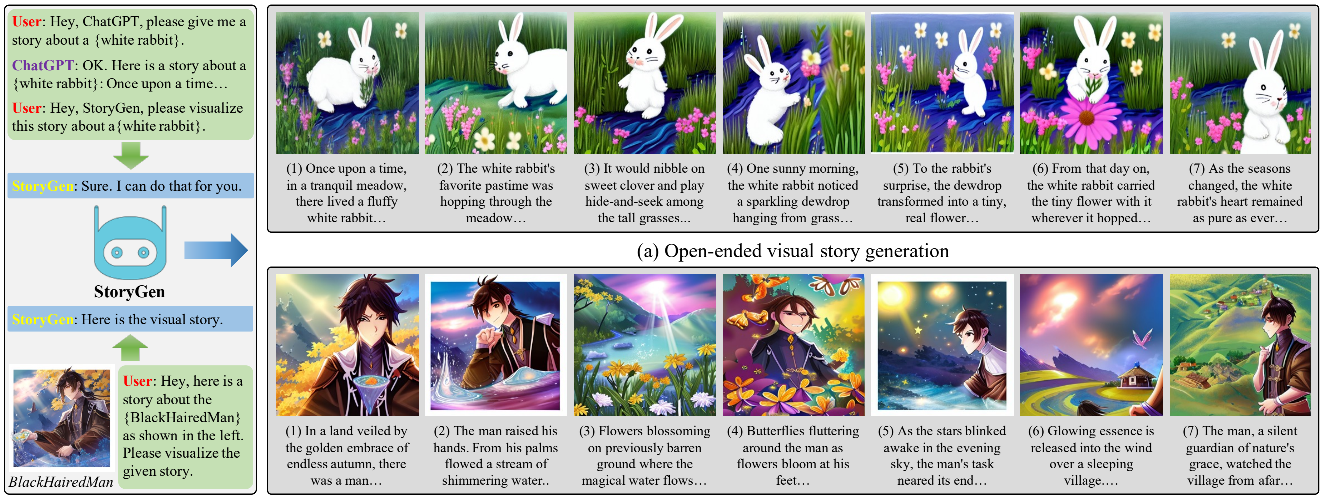
2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-5300-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/CVPR52733.2024.00592

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

Chang Liu[1,3*], Haoning Wu[1*], Yujie Zhong[2], Xiaoyun Zhang[1†], Yanfeng Wang[1,3†], Weidi Xie[1,3]
[1]Coop. Medianet Innovation Center, Shanghai Jiao Tong University, China
[2]Meituan Inc., China [3]Shanghai AI Laboratory, China


User: Hey, ChatGPT, please give me a story about a {white rabbit}.
ChatGPT: OK. Here is a story about a {white rabbit}: Once upon a time…
User: Hey, StoryGen, please visualize this story about a {white rabbit}.
StoryGen: Sure. I can do that for you.
StoryGen
StoryGen: Here is the visual story.
User: Hey, here is a story about the {BlackHairedMan} as shown in the left. Please visualize the given story.
BlackHairedMan
(1) Once upon a time, in a tranquil meadow, there lived a fluffy white rabbit…
(2) The white rabbit's favorite pastime was hopping through the meadow…
(3) It would nibble on sweet clover and play hide-and-seek among the tall grasses...
(4) One sunny morning, the white rabbit noticed a sparkling dewdrop hanging from grass…
(5) To the rabbit's surprise, the dewdrop transformed into a tiny, real flower…
(6) From that day on, the white rabbit carried the tiny flower with it wherever it hopped…
(7) As the seasons changed, the white rabbit's heart remained as pure as ever…
(1) In a land veiled by the golden embrace of endless autumn, there was a man…
(2) The man raised his hands. From his palms flowed a stream of shimmering water..
(3) Flowers blossoming on previously barren ground where the magical water flows…
(4) Butterflies fluttering around the man as flowers bloom at his feet…
(5) As the stars blinked awake in the evening sky, the man's task neared its end…
(6) Glowing essence is released into the wind over a sleeping village.…
(7) The man, a silent guardian of nature's grace, watched the village from afar…

(a) Open-ended visual story generation

(b) Open-ended visual story continuation

Figure 1. **An illustration of open-ended visual storytelling.** In practice, users can feed a unique and engaging story synthesized by a large language model into our proposed **StoryGen** model to generate a sequence of images coherently, denoted as *open-ended visual story generation*. And they can also provide a pre-defined character with its corresponding storyline, to perform *open-ended visual story continuation*. We recommend the reader to zoom in and read the story.

Abstract

Generative models have recently exhibited exceptional capabilities in text-to-image generation, but still struggle to generate image sequences coherently. In this work, we focus on a novel, yet challenging task of generating a coherent image sequence 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 propose a learning-based auto-regressive image generation model, termed as* ***StoryGen****, with a novel vision-language context module, that enables to generate the current frame by conditioning on the corresponding text prompt and preceding image-caption pairs; (ii) to address the data shortage of visual storytelling, we collect paired image-text sequences by sourcing from online videos and open-source E-books, establishing processing pipeline for constructing a large-scale dataset with diverse characters, storylines, and artistic styles, named* ***StorySalon****; (iii) Quantitative experiments and human evaluations have validated the superiority of our StoryGen, where we show it can generalize to unseen characters without any optimization, and generate image sequences with coherent content and consistent character. Code, dataset, and models are available at* `https://haoningwu3639.github.io/StoryGen_Webpage/`.

"Mirror mirror on the wall, who's the fairest of them all?"

—— Grimms' Fairy Tales

1. Introduction

This paper considers an interesting, yet challenging task, namely, *open-ended visual storytelling*. The goal is to train a generative model that effectively captures the relation between visual elements and corresponding text descriptions, to generate a sequence of images that tell a visually coherent story, as shown in Figure 1. The outcome of this task has significant potential for education, as it provides chil-

*: These authors contribute equally to this work.
†: Corresponding author.

2575-7075/24/$31.00 ©2024 IEEE
DOI 10.1109/CVPR52733.2024.00592

dren with an engaging and interactive way to learn complex visual concepts and develop imagination, creativity, emotional intelligence, and language skills, as evidenced by research in psychology [5, 45].

The recent literature has witnessed tremendous progress in image generation, particularly with the guidance of text as prompt, such as stable diffusion [41], DALL·E [39] and Imagen [14]. However, to generalize the models for open-ended visual storytelling, we are facing three challenges: (i) previous models are designed to only generate images independently, without considering context, for example, preceding frames or overall narrative, resulting in a lack of visual consistency; (ii) most methods generate images by only conditioning on text, which potentially leads to ambiguities or requires unnecessarily long descriptions to maintain character appearances; (iii) existing datasets are limited to a few animations, covering a closed set of vocabulary or characters [25, 31, 36]. Training on such datasets suffers from severe overfitting on seen characters, leading to unsatisfactory generalization capability for open-ended generation.

This paper describes a learning-based model for open-ended visual storytelling, termed as **StoryGen**, that enables to generate unseen characters without any further optimization, while having character consistency. At inference, StoryGen can synthesize frames either by taking text prompts, or along with preceding image-text pairs as conditions, *i.e.*, iteratively creating visual sequences that are aligned with language description, while being consistent with preceding frames in both style and character perspectives. Specifically, to achieve consistency within the generated image sequence, we incorporate a novel **vision-language context module** into the pre-trained stable diffusion model, which provides visual context by conditioning the generation process on extracted diffusion denoising feature of previous frames under the guidance of corresponding captions.

As for training, we construct a dataset called **StorySalon**, that features a rich source of coherent images and stories, primarily comprising children's storybooks collected from videos and E-books. As a result, our dataset includes a diverse vocabulary with different characters, storylines, and artistic styles. The scale and diversity of our collected dataset enable the model for open-vocabulary visual storytelling, *i.e.*, generating 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 them into StoryGen for generation, as shown in Figure 1.

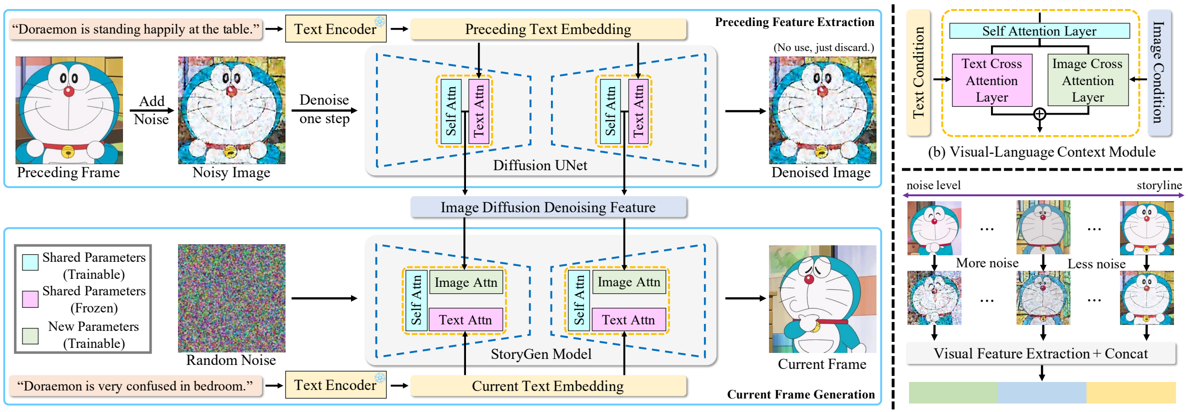
To summarize, we make the following contributions in this paper: (i) we initiate a fun yet challenging task, namely, *open-ended visual storytelling*, that involves generating engaging image sequences aligned to a given storyline; (ii) we propose a learning-based open-ended visual storytelling model, termed as **StoryGen**, which can generalize to unseen characters without any further optimization and generate coherent visual stories, utilizing a novel vision-language context module; (iii) we establish a data processing pipeline and collect a large-scale dataset of storybooks, called **StorySalon**, from online videos and open-source E-books, resulting in a diverse vocabulary with various characters, storylines, and artistic styles; (iv) we conduct quantitative experiments and human evaluations to validate the effectiveness of our proposed modules, demonstrating the superiority of our model, in terms of image quality, consistency, and visual-language alignment of generated contents.

2. Related Works

Text-to-image Generation has been tackled using various generative models, with GAN [8] as the first widely used model. Several GAN-based methods [50, 53, 54] have achieved notable success, and auto-regressive transformers [46], such as DALL·E [39], have also demonstrated the ability to generate high-quality images based on text prompts. Recently, diffusion models, such as Imagen [42] and DALL·E 2 [40], have emerged as a popular approach. Stable Diffusion Models [41] performs diffusion process in latent space, and can generate impressive images after pre-training on a large-scale text-image dataset.

Diffusion Models learn to model a data distribution via iterative denoising and are trained with denoising score matching. Notably, DDPM [13] demonstrates improved performance over other generative models, while DDIM [44] significantly boosts 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 [6, 14, 15, 43], image manipulation [2, 10, 18, 33], grounded generation [26], image restoration [4], and image inpainting [1, 28, 35, 48].

Story Synthesis is first introduced as the task of story visualization by StoryGAN [25], which presents a GAN-based framework and the PororoSV dataset, derived from cartoons. Some works [29, 30] follow the GAN-based framework, whereas others [3, 21] emphasize more on text representation. StoryDALL-E [31] extends story synthesis to story continuation with the initial image given, and exploits a pre-trained DALL·E model [39] to produce coherent images. AR-LDM [36] introduces an auto-regressive latent diffusion model to generate image sequences, but only consistent within a limited character vocabulary. NUWA-XL [52] exploits hierarchical diffusion models to synthesize long videos, but still achieve character consistency by memorizing. TaleCrafter [7] proposes a story visualization system and utilizes LoRA [16] to achieve character consistency. However, large-scale applications will be constrained due to its optimization-based nature. In this paper, we target more ambitious applications, to develop an open-ended visual storytelling model, that can synthesize coherent image



(a) Preceding Feature Extraction Process and Current Frame Generation Process

(b) Visual-Language Context Module

(c) Multi-frame Conditioning

Figure 2. **Architecture Overview.** (a) Our StoryGen model utilizes current text prompt and previous visual-language contexts as conditions to generate an image, iteratively synthesizing a coherent image sequence. Note the parameters of the corresponding attention layers are shared between Diffusion UNet and StoryGen. To avoid potential ambiguity, the parameters are not shared across UNet blocks in a single model. (b) The proposed Visual-Language Context Module can effectively combine the information from current text prompt and contexts from preceding image-caption pairs. (c) We add more noise to reference frames with longer temporal distances to the current frame as positional encoding to distinguish the temporal order. The multiple features can then be directly concatenated to serve as context conditions.

sequences based on storylines of diverse topics.

3. Method

In this section, we start by formulating the problem of open-ended visual storytelling in Section 3.1; then we elaborate on the proposed StoryGen architecture in Section 3.2; lastly, we present details for model training in Section 3.3.

3.1. Problem Formulation

In this paper, we focus on a challenging task, termed as *open-ended visual storytelling*, the goal is to generate continuous image sequence from a given story in the form of natural language. Specifically, we propose a learning-based auto-regressive image generation model, called **StoryGen**, that generates the current frame $\hat{\mathcal{I}}_k$ by conditioning on the current text prompt $\mathcal{T}_k$, and image-text pairs $(\hat{\mathcal{I}}_{<k}, \mathcal{T}_{<k})$ of previous frames, as illustrated in Figure 2 (a). The model is formulated as follows:

$$\begin{aligned}\{\hat{\mathcal{I}}_1, \hat{\mathcal{I}}_2, \dots, \hat{\mathcal{I}}_L\} &= \Phi_{\text{StoryGen}}(\{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_L\}; \Theta) \\ \hat{\mathcal{I}}_k &:= \Phi_{\text{StoryGen}}(\hat{\mathcal{I}}_k | \mathcal{T}_k, (\hat{\mathcal{I}}_{<k}, \mathcal{T}_{<k}))\end{aligned}$$

Here, $\{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_L\}$ refer to the given storylines, and $\{\hat{\mathcal{I}}_1, \hat{\mathcal{I}}_2, \dots, \hat{\mathcal{I}}_L\}$ denote the generated image sequence. $\Phi_{\text{StoryGen}}(\cdot)$ represents our proposed StoryGen model. In one-step generation, StoryGen takes the current text prompt, and preceding image-caption pairs as conditions, and generates an image consistent with both the story's narrative and previous frames. The whole image sequence can then be synthesized with step-by-step inference.

Relation to Existing Tasks. In contrast to existing story visualization works, this paper makes improvements from two aspects: (i) conventional generation/continuation tasks are limited to training on specific characters/stories, for example, [25, 31, 36] only exploits datasets from animation *The Flintstones* and *Pororo*, while our model enables to generate visual stories based on any given storyline, such as a brand-new one generated by ChatGPT; and any pre-defined character, for example, *'Doraemon'* from the Internet; (ii) unlike existing work that requires costly character-specific optimization, for example, [7, 36] rely on LoRA-based [16] optimization to adapt to new characters, our model is learning-based and expected to generalize to any unseen character without any further optimization.

3.2. Architecture

To tackle the problem of open-ended visual storytelling, we expect the model to not only condition on the current text prompt, but also preceding image-text pairs. In this section, we describe the procedure for one-step generation, *i.e.*, generating the k-th frame ($k > 1$) by conditioning on $\{(\hat{\mathcal{I}}_1, \mathcal{T}_1), \dots, (\hat{\mathcal{I}}_{k-1}, \mathcal{T}_{k-1}), \mathcal{T}_k\}$. Generally speaking, our proposed **StoryGen** model comprises four components: (i) Input Initialization, (ii) Context Encoding, (iii) Visual-Language Contextual Fusion, (iv) Conditional Generation.

Input Initialization. Our model is built upon the foundation of a pre-trained stable diffusion model (SDM), which randomly samples a noisy latent $\mathbf{x}$ from the latent space of the VAE [19] encoder. Moreover, for a given text prompt $\mathcal{T}_k$, the text condition will be extracted by a pre-trained CLIP [37] text encoder ϕ_{CLIP} via $\mathcal{C}^{\text{T}} = \phi_{\text{CLIP}}(\mathcal{T}_k)$.

Context Encoding. In standard SDM, the noisy latent is recursively denoised with a UNet, conditioning on the text prompt. However, in our case, it is crucial for the generation procedure to also condition on context features of preceding frames, to maintain consistency in characters and storyline.

In practice, to extract the contextual features, we add noise to the preceding frames and exploit the pre-trained SDM to denoise for one diffusion step under the guidance of their corresponding captions. The diffusion features after every self-attention layer in the UNet blocks can be directly selected to serve as the conditioning visual context features, thus constituting a pyramid of visual context features. The visual condition features for $\hat{\mathcal{I}}_k$ can be expressed as:

$$\mathcal{C}^{\text{V}} = [\phi_{\text{SDM}}(\hat{\mathcal{I}}_1, \phi_{\text{CLIP}}(\mathcal{T}_1)), \ldots, \phi_{\text{SDM}}(\hat{\mathcal{I}}_{k-1}, \phi_{\text{CLIP}}(\mathcal{T}_{k-1}))]$$

Experimentally, we notice that, the magnitude of noise added to the preceding frames can greatly affect the conditional generation quality, *i.e.*, large-scale noise on preceding frames incurs severe information loss. Thus, we propose to use a much smaller diffusion timestep t' for preceding frames compared with the diffusion timestep t of the current image $\hat{\mathcal{I}}_k$, and follow a $t' = t/10$ rule. As depicted in Figure 2 (c), in case of multiple preceding image-caption pairs, we use larger t' for frames with longer temporal distances to $\hat{\mathcal{I}}_k$. Therefore, the extracted multi-frame visual context features can be directly concatenated, and their different noise level will serve as temporal positional embedding. Such design reflects the intuition that frames with longer distances will incur less effect on generating the current frame.

Vision-Language Contextual Fusion. Here, our vision-language context module is designed to fuse information from current text prompt and contextual information from preceding image-caption pairs. This is achieved by augmenting the transformer decoder in SDM with an additional image cross-attention layer. Note that, the math expression in this section is not strict, we omit the footnote of diffusion timestep t and UNet block level l for simplicity.

Specifically, on visual context conditioning, the noisy latent $\mathbf{x}$ is projected into query, and cross-attends to the visual context features from the corresponding-level UNet block that act as key and value, denoted as:

$$\mathbf{Q}_I = \mathbf{x}\mathbf{W}_I^Q, \quad \mathbf{K}_I = \mathcal{C}^{\text{V}}\mathbf{W}_I^K, \quad \mathbf{V}_I = \mathcal{C}^{\text{V}}\mathbf{W}_I^V$$

where $\mathbf{W}_I^Q$, $\mathbf{W}_I^K$, and $\mathbf{W}_I^V$ represent different projection matrices, respectively.

On text conditioning, the noisy latent $\mathbf{x}$ is again projected to query, and cross-attends to the text features of the current prompt encoded by CLIP text encoder, *i.e.*,

$$\mathbf{Q}_T = \mathbf{x}\mathbf{W}_T^Q, \quad \mathbf{K}_T = \mathcal{C}^{\text{T}}\mathbf{W}_T^K, \quad \mathbf{V}_T = \mathcal{C}^{\text{T}}\mathbf{W}_T^V$$

where $\mathbf{W}_T^Q$, $\mathbf{W}_T^K$, and $\mathbf{W}_T^V$ also represent corresponding projection matrices.

As depicted in Figure 2 (b), the image cross-attention layer is inserted in parallel to the text cross-attention layer in the transformer decoder of UNet blocks. Drawing inspiration from ControlNet [55], the results from these two cross-attention layers are simply summed up as the final output $\mathbf{O}$. The final output can thus be expressed as:

$$\mathbf{O} = \text{Softmax}(\frac{\mathbf{Q}_I(\mathbf{K}_I)^\top}{\sqrt{d}})\mathbf{V}_I + \text{Softmax}(\frac{\mathbf{Q}_T(\mathbf{K}_T)^\top}{\sqrt{d}})\mathbf{V}_T$$

Conditional Generation. With the fused vision-language condition features from above, our StoryGen can now generate visual stories that achieve both content coherence and character consistency. Here, our conditional generation procedure can be represented as:

$$\hat{\mathcal{I}}_k = \Phi_{\text{StoryGen}}(\hat{\mathcal{I}}_k | \mathcal{T}_k, (\hat{\mathcal{I}}_{<k}, \mathcal{T}_{<k})) = \Phi_{\text{StoryGen}}(\mathbf{x}, \mathcal{C}^{\text{T}}, \mathcal{C}^{\text{V}})$$

With the new conditioning modality introduced, we also adopt another classifier-free guidance term [12], as has been done in [2]. Concretely, we exploit two different guidance scales, w_v and w_t for the visual condition and the text condition. The relation between the final noise for inference $\bar{\boldsymbol{\epsilon}}_\theta$ and UNet-predicted noise $\boldsymbol{\epsilon}_\theta$ is now expressed as:

$$\begin{aligned}\bar{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t, \mathcal{C}^{\text{V}}, \mathcal{C}^{\text{T}}) = {} & \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \varnothing, \varnothing) \\ & + w_v(\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathcal{C}^{\text{V}}, \varnothing) - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \varnothing, \varnothing)) \\ & + w_t(\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathcal{C}^{\text{V}}, \mathcal{C}^{\text{T}}) - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathcal{C}^{\text{V}}, \varnothing))\end{aligned}$$

Discussion. Our work differs from previous ones from two aspects. First, our StoryGen is a learning-based method, which can directly generalize to unseen characters by attending to reference images. Second, we propose to condition the generation process on diffusion features of preceding image-text pairs from the same SDM, which preserves more visual details, greatly differing from existing works [22, 49, 51] using CLIP, BLIP [24], or VAE features.

3.3. Model Training

Training Objective. At training stage, we randomly sample a triplet each time, *i.e.*, $\{\mathcal{I}_k, \mathcal{T}_k, (\mathcal{I}_{<k}, \mathcal{T}_{<k})\}$. The objective function can be expressed as:

$$\mathcal{L}_t = \mathbb{E}_{t \sim [1,T], \mathbf{x}_0, \boldsymbol{\epsilon}_t, \mathcal{C}^{\text{V}}, \mathcal{C}^{\text{T}}}\left[\|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathcal{C}^{\text{V}}, \mathcal{C}^{\text{T}})\|^2\right]$$

Two-stage Training Strategy. Our two-stage training strategy includes single-frame pre-training and multiple-frame fine-tuning. To be specific, at the first stage, we do not introduce additional image cross-attention layers, and only train self-attention layers in standard SDM to ensure the single-frame generation ability. In multiple-frame fine-tuning, we train additional image cross-attention layers in vision-language context module on our dataset, with all other parameters frozen. This enables the generation procedure to utilize information from not only current prompt, but also preceding image-caption pairs.

Inference. As shown in Figure 1, at inference time, we can prompt ChatGPT to generate novel storylines, and synthesize the first image directly or attending to a pre-defined

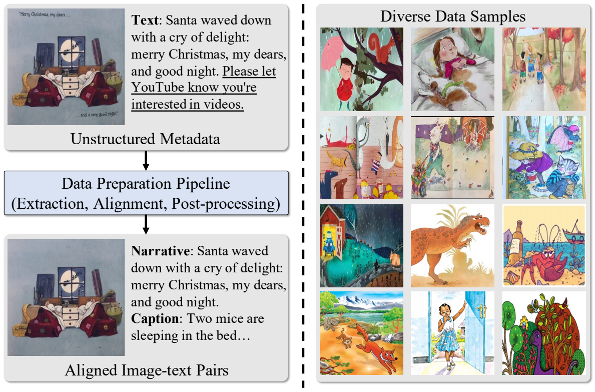

Text: Santa waved down with a cry of delight: merry Christmas, my dears, and good night. Please let YouTube know you're interested in videos.
Unstructured Metadata
Data Preparation Pipeline
(Extraction, Alignment, Post-processing)
Narrative: Santa waved down with a cry of delight: merry Christmas, my dears, and good night.
Caption: Two mice are sleeping in the bed…
Aligned Image-text Pairs
Diverse Data Samples

Figure 3. **Dataset Pipeline and Visualization**. **Left**: Metadata sourced from the Internet undergoes a three-step pipeline including frame extraction, visual-language alignment and post-processing, resulting in properly aligned image-text pairs. **Right**: Our StorySalon dataset contains diverse styles and characters.

character. Then the previously synthesized frames, along with the story descriptions, are treated as conditions to synthesize the image sequence in an auto-regressive manner. Experimentally, our proposed StoryGen is shown to generate images that align with the storyline, as well as maintain consistency with previously generated frames.

4. StorySalon Dataset

In order to train our proposed *open-ended visual storytelling* model, we construct a large-scale dataset, termed as **StorySalon**. The dataset contains videos and E-books with diverse characters, storylines, and artistic styles. Specifically, we download a large number of videos and subtitles from YouTube, by querying keywords related to story-telling for children, for instance, *storytime*. Additionally, we collect E-books (partially with corresponding audios available) from six open-source libraries which are all registered under the Creative Commons 4.0 International Attribution (CC BY 4.0) license. In the following, we elaborate on the data processing pipeline and statistics of our collected dataset.

Visual Frame Extraction. We extract keyframes from the videos, along with the corresponding subtitles and their timestamps. To remove duplicate frames, we extract ViT features for each frame using pre-trained DINO [32]. For the image groups with high similarity scores, we only keep one of each. Then, we use YOLOv7 [47] to segment and remove real-person frames and headshots, as they often correspond to the story-teller and are unrelated to the content of the storybook. Similarly, we extract images from the downloaded E-books, except for those with extraneous information, for example, the authorship page. We acquire the corresponding text description with Whisper [38] from the audio file, and for E-books that do not have corresponding audio files, but with available storyline text, we use OCR algorithms, to directly recognize the text on each page.

Dataset	Style	#Frames	Avg.Length	#Categories
PororoSV [25]	Animation	73,665	5	9
FlintstonesSV [9]	Animation	122,560	5	7
DiDeMoSV [31]	Real	52,905	3	-
VIST [17]	Real	145,950	5	-
StorySalon	Animation	**159,778**	**14**	**446**

Table 1. **Dataset Statistics.** Our StorySalon dataset far exceeds previous story generation datasets in terms of the total number of images, average length, and categories of characters included.

Visual-Language Alignment. As shown in Figure 3, for each of the image, we can collect two types of text descriptions, *e.g.*, story-level narration, and descriptive captions. This is based on our observation that there actually exists a semantic gap between narrative storyline and descriptive text, for example, the same image can be well described as *"The cat is isolated by others, sitting alone in front of a village."* in the story, or *"A black cat sits in front of a number of houses."* as descriptive caption, therefore, directly fine-tuning stable diffusion models with story narration may be detrimental to its pre-trained text-image alignment. In practice, to get story-level paired image-text samples, we align the subtitles with visual frames by using Dynamic Time Warping (DTW) algorithm [34]. To get visual descriptions, we use TextBind [23] to generate captions for each image, with both the image and the corresponding narrative text as inputs. At training time, this allows us to substitute the original story with more accurate and descriptive captions.

Visual Frame Post-processing. In practice, we observe that book pages and borders in images can potentially interfere with our generative model by having story texts printed on them. To tackle this, we use an OCR detector to identify text regions in images and an image inpainting model [41] to fill in the text and headshot regions, resulting in more precise image-text pairs that are suitable for model training.

Discussion. After the three-step pipeline above, we obtain our StorySalon dataset. As shown in Table 1, our dataset has nearly 160K animation-style images in total with an average length of 14 frames per story, which is conducive to building long-range semantic correspondence. Finally, we query MiniGPT-4 [56] about the main character category of each image in our dataset, like *Dog* and *Cat*, then count the categories and filter out those appear less than 3 times. Compared with previous datasets with less than 10 characters, our dataset comprises hundreds of character categories, and even more character instances, which provides a data basis for training open-ended visual storytelling models, showing a significantly broader range of visual styles and character appearances over existing datasets.

5. Experiments

In this section, we start by describing our experimental settings, then compare with other models from three different

Model	FID ↓	CLIP-I ↑	CLIP-T ↑
GT	-	1.0	0.2668
SDM	73.50	0.6155	0.3218
Prompt-SDM	67.35	0.6272	**0.3225**
Finetuned-SDM	42.01	0.6970	0.3005
StoryDALL·E	38.34	0.6823	0.2366
AR-LDM	39.55	0.6864	0.2614
StoryGen	**33.90**	**0.7467**	0.2875

Table 2. **Comparison of automatic metrics** on StorySalon test set. Prompt-SDM denotes Stable Diffusion model with cartoon-style-directed prompts and Finetuned-SDM represents a Stable Diffusion model with all parameters fine-tuned on our StorySalon dataset.

Story Generation						
Model	Align. ↑	Style ↑	Cont. ↑	Char. ↑	Qual. ↑	Pref. ↑
GT	4.04	4.66	4.41	4.54	4.29	–
SDM	3.61	2.88	2.90	2.51	3.74	14.05%
Prompt-SDM	3.39	2.56	2.68	2.10	3.44	8.57%
StoryGen-S	3.50	2.73	2.81	2.21	3.19	10.24%
StoryGen	**3.78**	**4.79**	**4.26**	**4.64**	**3.76**	**67.14%**
Story Continuation						
StoryDALL·E	1.18	1.55	1.20	1.14	1.19	0.63%
AR-LDM	2.47	2.82	2.40	1.87	2.54	2.50%
StoryGen	**4.23**	**4.70**	**4.35**	**4.38**	**4.18**	**96.87%**

Table 3. **Comparison results of human evaluation.** GT stands for ground truth from the test set. StoryGen-S represents StoryGen without context conditions. The abbreviated metrics are Text-image alignment, Style consistency, Content consistency, Character consistency, image quality, and Preference, respectively.

perspectives: image-text alignment, consistency and image quality with subjective human evaluation and quantitative metrics. Additionally, we present results for ablation experiments to prove the effectiveness of our proposed modules. Please refer to ArXiv version for more experimental details.

5.1. Experimental Settings

Training Details. Our model is built on the stable diffusion v1.5 model, and trained with a learning rate of 1×10^{-5} and a batch size of 256. We begin with a single-frame self-attention pre-training stage, which involves 3,000 iterations on 8 NVIDIA RTX3090. Next, we incorporate our proposed vision-language context module, and train it for 5,000 iterations using a single preceding image-caption pair as context condition, then continue to train it for another 5,000 iterations with multiple image-caption pairs for multi-frame conditioning. To maintain our model's unconditional denoising ability for classifier-free guidance, we randomly drop the current text and the context image-caption pairs with a probability of 5% and 15%, respectively. During inference, we utilize DDIM [44] with 40 steps of sampling and select the guidance weight $w_v = 7.0$ and $w_t = 3.5$.

Baselines. We consider two scenarios of our proposed open-ended storytelling task, namely, story generation and story continuation. For **story generation**, we need the model to be able to generate a complete visual story only based on a given storyline. So we present a comparison with Stable Diffusion Model (**SDM**) and **Prompt-SDM**, which conditions on an additional cartoon-style-directed prompt *"A cartoon style image"*. For **story continuation**, the first frame or the main character is given, and the model is expected to generate coherent images based on the storyline. In this scenario, we compare our model with two closed-set story continuation models: namely, **StoryDALL·E** [31] and **AR-LDM** [36] re-trained on our StorySalon dataset.

Automatic Metrics. To evaluate the quality of generated image sequences, we adopt three widely-used metrics, including Fréchet Inception Distance score (FID) [11], CLIP image-image similarity (CLIP-I), and CLIP text-image similarity (CLIP-T). Notably, in order to avoid the impact of randomness in synthesis quality, we utilize a CLIP-based scoring function trained exclusively on text-to-image generated images, namely, PickScore [20], to automatically select the generated images with better quality. Each chosen image is selected from a pool of 10 candidates.

5.2. Quantitative Evaluation Results

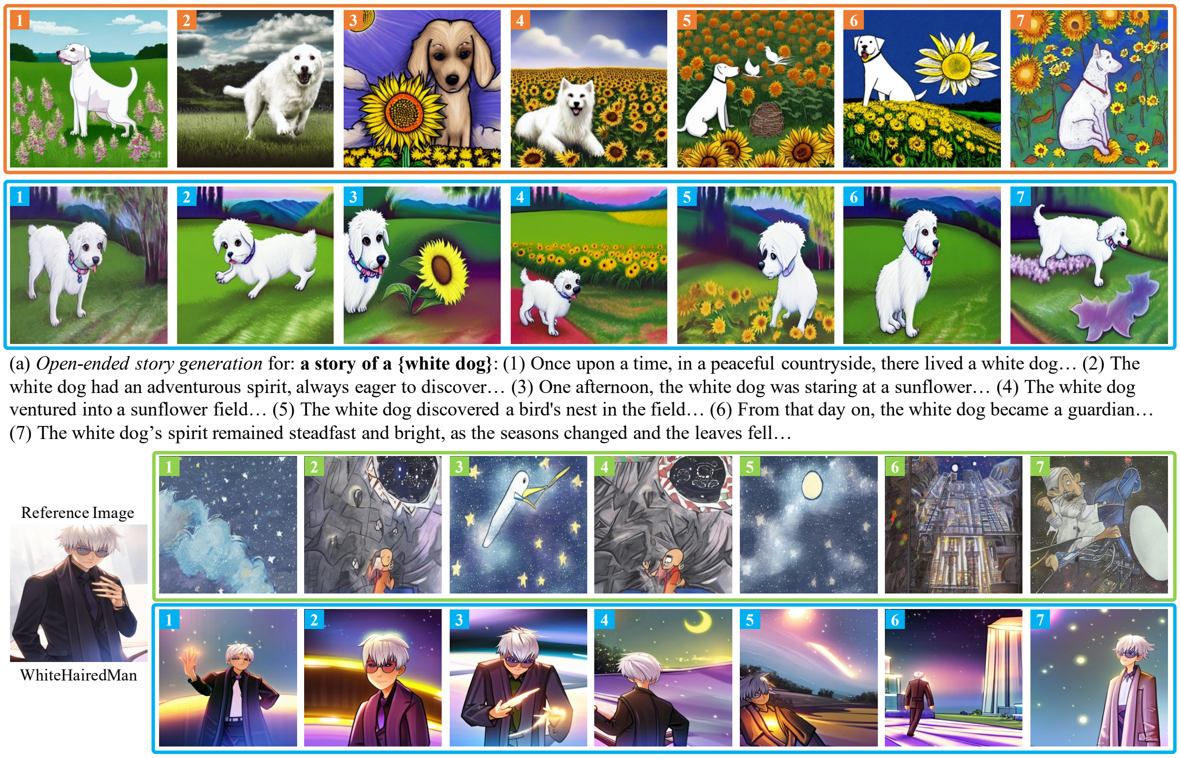
We compare our StoryGen model with other baselines on StorySalon test set, which contains 5% of total data (nearly 7K pairs). Each contains a current prompt and the image-text context of the previous frame. The models are expected to generate the current frame based on given conditions.

The quantitative results in Table 2 demonstrate that our StoryGen model exhibits significant performance improvement in terms of FID score and CLIP-I similarity compared to existing models, while maintaining comparable CLIP-T similarity. This confirms that our model can effectively exploit contextual information, thus generating animation-style visual stories based on the given storyline. Notably, CLIP trained on natural images tends to have an understanding bias towards animation-style images, and the slight decline in CLIP-T is an inevitable result of the conflict between text condition and newly introduced image condition.

5.3. Human Evaluation Results

Considering that the above metrics may not reflect the quality of the generated stories accurately, and there is no standardized metric for evaluating the consistency within the visual story, we further include human evaluation for comparison of image-text alignment, image style, story consistency, character consistency and synthesis quality.

For the two scenarios mentioned above, we respectively conduct two types of human evaluation to assess the quality of generated visual stories. To mitigate bias, participants are



(a) *Open-ended story generation* for: **a story of a {white dog}**: (1) Once upon a time, in a peaceful countryside, there lived a white dog… (2) The white dog had an adventurous spirit, always eager to discover… (3) One afternoon, the white dog was staring at a sunflower… (4) The white dog ventured into a sunflower field… (5) The white dog discovered a bird's nest in the field… (6) From that day on, the white dog became a guardian… (7) The white dog's spirit remained steadfast and bright, as the seasons changed and the leaves fell…

(b) *Open-ended story continuation* for: **a story of a {a white-haired man}**: (1) In a perpetual twilight, the white-haired man reached towards the twilight sky, stars appearing at his touch… (2) One twilight, the white-haired man looked concerned at a dark void in the sky... (3) The white-haired man drew stars in the sky with a silver quill… (4) The white-haired man was observing new constellations shining where the void once was… (5) The white-haired man with a serene expression was watching the peaceful starry sky... (6) The white-haired man walked towards a tower observatory under the starry sky… (7) Alone but content, the white-haired man's gaze traversed the depths of space…

Figure 4. **Qualitative Comparison with other methods**. The image sequences in orange, green, and blue boxes are generated by Prompt-SDM, AR-LDM and StoryGen respectively. Our synthesis results exhibit impressive performance superiority in terms of style, content and character consistency, text-image alignment, and image quality. Please refer to the Appendix for more qualitative results.

unaware of the type of storybooks they are evaluating. Concretely, we prompt GPT-4 to produce multiple storylines for both test modes, and for story continuation, we search the Internet for multiple characters that have never appeared in our dataset. Then we utilize our StoryGen along with other baselines to generate corresponding sequences of images.

Protocol-I. We randomly select an equal number of samples from the generated results of our StoryGen and other baselines. Each time we randomly sample a visual story from these sources, and participants are then invited to rate the sample with a score ranging from 1 to 5, taking into account text-image alignment, style consistency, content consistency, character consistency and image quality. Higher scores indicate better samples. We also evaluate the same number of samples from StorySalon test set as a reference.

Protocol-II. Each time we randomly sample a storyline and its corresponding visual storybooks generated by StoryGen and other methods. Participants are invited to select their preferred generated result among these different image sequences of the same storyline.

Results. The results of human evaluation presented in Table 3 illustrate that our StoryGen model demonstrates excellent performance in overall score, especially in terms of consistency and quality. This indicates that our model can generate coherent image sequences that are highly consistent with given text prompts and visual-language contexts.

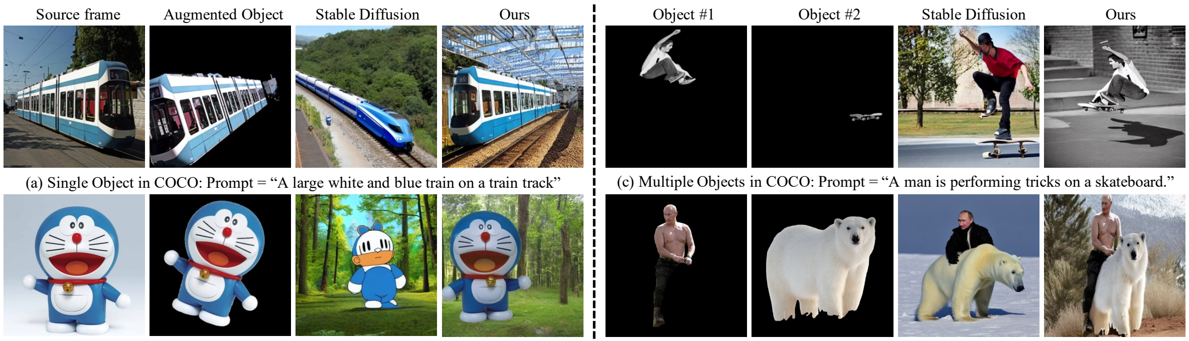
5.4. Qualitative Results

In Figure 4, we present visualization results of both open-ended visual story generation and visual story continuation, showing that our StoryGen can generate visual stories with a broad vocabulary, while maintaining content coherence and character consistency throughout the narrative, whereas other methods fail to do so. Moreover, our model can stably maintain the animation style of generated images, which satisfies the requirements of visual storytelling for children. More results can be found in our ArXiv version.

5.5. Ablation Studies

In order to demonstrate the effectiveness of our proposed modules, we conduct ablation studies from both quantitative metrics and qualitative visualization.

On Variants of StoryGen. We evaluate the performance of



(a) Single Object in COCO: Prompt = "A large white and blue train on a train track"

(c) Multiple Objects in COCO: Prompt = "A man is performing tricks on a skateboard."

(b) Single Object in Real-Data: Prompt = "Doraemon is standing in a forest."

(d) Multiple Objects in Real-Data: Prompt = "Putin is riding a polar bear in the wild."

Figure 5. **Ablation studies on consistency**. We incorporate our proposed Visual-Language Context Module into a pre-trained SDM, and train it on MS-COCO [27] with other parameters frozen. The content consistency of single-object and multi-object generation on COCO and real data has demonstrated the effectiveness of our module. Please refer to the Appendix for experiment details and quantitative results.

multiple model variants on the StorySalon test set, including (i) our model without the context module, marked as **StoryGen-Single**, which solely fine-tunes the self-attention layers on our dataset. (ii) our model with context features encoded by the VAE of SDM as context condition, without text-guided diffusion process, denoted as **StoryGen-VAE**; (iii) our model with CLIP image embedding as context condition (**StoryGen-CLIP**); (iv) our model with context features extracted by BLIP image encoder (**StoryGen-BLIP**); (v) our model with naive denoising features at **L**arge-scale diffusion **T**imestep, satisfying $t' = t$, as condition (**StoryGen-LT**); and (vi) our full model (**StoryGen**). We also employ PickScore to filter generation results of all these models. The findings presented in Table 4 illustrate the inclusion of our context module can significantly improve the model performance, in terms of CLIP-I and FID. As for the slight inferiority in CLIP-T, we have claimed above that this is due to the understanding bias towards animation-style images for CLIP trained on natural images.

Model	FID ↓	CLIP-I ↑	CLIP-T ↑
StoryGen-Single	38.81	0.6869	**0.3140**
StoryGen-VAE	36.98	0.6846	0.3061
StoryGen-CLIP	36.66	0.6934	**0.3140**
StoryGen-BLIP	34.78	0.7026	0.2838
StoryGen-LT	36.41	0.7141	0.3025
StoryGen	**33.90**	**0.7467**	0.2875

Table 4. **Ablation studies** on Visual-Language Context Module.

Qualitative Visualization. As mentioned above, consistency is a crucial factor in visual story generation. We hope to more intuitively demonstrate that our proposed context module can accurately capture the image content of the previous frame. To this end, we incorporate our context module into SDM and train it from scratch on the MS-COCO [27] with other parameters frozen. Specifically, we crop the object and perform data augmentations such as translation and rotation to use it as image condition. The category of the cropped object is used as its corresponding text, and the caption of the original image serves as the text prompt. We expect the model to reconstruct the original image relying on the conditions above, which enables the context module to learn how to leverage the previous image. As shown in Figure 5, our model can make full use of the objects in the reference frame and generate new images that are consistent with them, while SDM fails to do so. In addition, this can also be transferred to any real-world reference image, which strongly illustrates the robustness and capability of our context module to assist diffusion models in generating images based on any given object.

6. Conclusion

In this paper, we consider an interesting, yet challenging task, termed as *open-ended visual storytelling*, which involves generating a sequence of images that tell a coherent visual story based on the given storyline. Our proposed learning-based **StoryGen** model can take input from the preceding image-caption context along with the text prompt to generate coherent image sequences in an auto-regressive manner, *i.e.*, without test-time optimization. On the data side, we establish a data processing pipeline to collect a large-scale dataset named **StorySalon** that comprises storybooks with diverse characters, storylines, and artistic styles sourced from videos and E-books. Extensive human evaluation and quantitative comparison have illustrated that our proposed model substantially outperforms existing models, from the perspective of image quality, content coherence, character consistency, and visual-language alignment.

Acknowledgments

This work is supported by National Natural Science Foundation of China (62271308), STCSM (22511105700, 22DZ2229005), 111 plan (BP0719010), State Key Laboratory of UHD Video and Audio Production and Presentation, and National Key R&D Program of China (No. 2022ZD0161400).

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2

[2] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2, 4

[3] Hong Chen, Rujun Han, Te-Lin Wu, Hideki Nakayama, and Nanyun Peng. Character-centric story visualization via visual planning and token alignment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processinng*, 2022. 2

[4] Zheng Chen, Yulun Zhang*, Ding Liu, Bin Xia, Jinjin Gu, Linghe Kong*, and Xin Yuan. Hierarchical integration diffusion model for realistic image deblurring. In *Advances in Neural Information Processing Systems*, 2023. 2

[5] 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. 2

[6] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the International Conference on Computer Vision*, 2023. 2

[7] Yuan Gong, Youxin Pang, Xiaodong Cun, Menghan Xia, Yingqing He, Haoxin Chen, Longyue Wang, Yong Zhang, Xintao Wang, Ying Shan, and Yujiu Yang. Talecrafter: Interactive story visualization with multiple characters. *SIGGRAPH Asia*, 2023. 2, 3

[8] 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. 2

[9] Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. Imagine this! scripts to compositions to videos. In *Proceedings of the European Conference on Computer Vision*, 2018. 5

[10] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *Proceedings of the International Conference on Learning Representations*, 2023. 2

[11] 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. In *Advances in Neural Information Processing Systems*, 2017. 6

[12] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 4

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020. 2

[14] 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. 2

[15] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, 2022. 2

[16] 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. In *Proceedings of the International Conference on Learning Representations*, 2022. 2, 3

[17] Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016. 5

[18] 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. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2

[19] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations*, 2014. 3

[20] 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. In *Advances in Neural Information Processing Systems*, 2023. 6

[21] Bowen Li. Word-level fine-grained story visualization. In *Proceedings of the European Conference on Computer Vision*, 2022. 2

[22] Dongxu Li, Junnan Li, and Steven CH Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In *Advances in Neural Information Processing Systems*, 2023. 4

[23] Huayang Li, Siheng Li, Deng Cai, Longyue Wang, Lemao Liu, Taro Watanabe, Yujiu Yang, and Shuming Shi. Textbind: Multi-turn interleaved multimodal instruction-following. *arXiv preprint arXiv:2309.08637*, 2023. 5

[24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning*, 2022. 4

[25] 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, 5

[26] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Guiding text-to-image diffusion

model towards grounded generation. In *Proceedings of the International Conference on Computer Vision*, 2023. 2

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, 2014. 8

[28] 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 Conference on Computer Vision and Pattern Recognition*, 2022. 2

[29] Adyasha Maharana and Mohit Bansal. Integrating visuospatial, linguistic, and commonsense structure into story visualization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processinng*, 2021. 2

[30] 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. 2

[31] 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, 5, 6

[32] 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. 5

[33] 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 *Proceedings of the International Conference on Learning Representations*, 2021. 2

[34] Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, 2007. 5

[35] 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. In *Proceedings of the International Conference on Machine Learning*, 2022. 2

[36] Xichen Pan, Pengda Qin, Yuhong Li, Hui Xue, and Wenhu Chen. Synthesizing coherent story with auto-regressive latent diffusion models. In *Winter Conference on Applications of Computer Vision*, 2024. 2, 3, 6

[37] 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

[38] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the International Conference on Machine Learning*, 2023. 5

[39] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the International Conference on Machine Learning*, 2021. 2

[40] 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. 2

[41] 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, 5

[42] 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. In *Advances in Neural Information Processing Systems*, 2022. 2

[43] 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. In *Proceedings of the International Conference on Learning Representations*, 2023. 2

[44] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proceedings of the International Conference on Learning Representations*, 2020. 2, 6

[45] Gabrielle A. Strouse, Angela Nyhout, and Patricia A. Ganea. The role of book features in young children's transfer of information from picture books to real-world contexts. *Frontiers in Psychology*, 2018. 2

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 2

[47] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 5

[48] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2

[49] Li Xin, Chu Wenqing, Wu Ye, Yuan Weihang, Liu Fanglong, Zhang Qi, Li Fu, Feng Haocheng, Ding Errui, and Wang Jingdong. Videogen: A reference-guided latent diffusion approach for high definition text-to-video generation. *arXiv preprint arXiv:2309.00398*, 2023. 4

[50] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[51] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arxiv:2308.06721*, 2023. 4

[52] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, et al. Nuwa-xl: Diffusion over diffusion for extremely long video generation. In *Association for Computational Linguistics*, 2023. 2

[53] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the International Conference on Computer Vision*, 2017. 2

[54] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 2

[55] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the International Conference on Computer Vision*, 2023. 4

[56] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 5