StoryDALL-E: Adapting Pretrained Text-to-Image Transformers for Story Continuation

Adyasha Maharana, Darryl Hannan, and Mohit Bansal

UNC Chapel Hill, NC 27514, USA
{adyasha,dhannan,mbansal}@cs.unc.edu

Abstract. Recent advances in text-to-image synthesis have led to large pretrained transformers with excellent capabilities to generate visualizations from a given text. However, these models are ill-suited for specialized tasks like story visualization, which requires an agent to produce a sequence of images given a corresponding sequence of captions, forming a narrative. Moreover, we find that the story visualization task fails to accommodate generalization to unseen plots and characters in new narratives. Hence, we first propose the task of story continuation, where the generated visual story is conditioned on a source image, allowing for better generalization to narratives with new characters. Then, we enhance or 'retro-fit' the pretrained text-to-image synthesis models with task-specific modules for (a) sequential image generation and (b) copying relevant elements from an initial frame. We explore full-model finetuning, as well as prompt-based tuning for parameter-efficient adaptation, of the pretrained model. We evaluate our approach StoryDALL-E on two existing datasets, PororoSV and FlintstonesSV, and introduce a new dataset DiDeMoSV collected from a video-captioning dataset. We also develop a model StoryGANc based on Generative Adversarial Networks (GAN) for story continuation, and compare with the StoryDALL-E model to demonstrate the advantages of our approach. We show that our retro-fitting approach outperforms GAN-based models for story continuation. We also demonstrate that the 'retro-fitting' approach facilitates copying of visual elements from the source image and improved continuity in visual frames. Finally, our analysis suggests that pretrained transformers struggle with comprehending narratives containing multiple characters, and translating them into appropriate imagery. Our work encourages future research into story continuation and large-scale models for the task.[1]

1 Introduction

Pretrained text-to-image synthesis models like DALL-E [33] have shown unprecedented ability to convert an input caption into a coherent visualization. Several subsequent approaches have also leveraged powerful multimodal models [4, 32]

[1] Code and data are available at https://github.com/adymaharana/storydalle

for creating artistic renditions of input captions [5], demonstrating their potential for democratizing art. However, these models are designed to process only a single, short caption as input. In contrast, many use cases of text-to-image synthesis require models to process long narratives and metaphorical expressions, condition on existing visuals, and generate more than one image to capture the meaning of the input text. In the past, multiple works have developed specialized Generative Adversarial Networks (GAN) models such as image-to-image translation [15], style transfer [18] etc. For instance, story visualization models [23] convert a sequence of captions into a sequence of images which illustrate the story. However, the recent advent of transformer-based large pretrained models opens up possibilities for leveraging latent knowledge from large-scale pretrained datasets for performing these specialized tasks more effectively. Hence, in this paper, we explore methods to adapt a pretrained text-to-image synthesis model for complex downstream tasks, with a focus on story visualization.

Story visualization is a challenging task that lies at the intersection of image generation and narrative understanding. Given a series of captions, which compose a story, an agent must generate a corresponding sequence of images that depicts the contents of these captions. While prior work in story visualization has discussed potential applications of the task [23, 27, 28, 37], the task itself presents some difficulties when being applied to real world settings. The model is limited to the fixed set of characters, settings, and events on which it is trained and has no way of knowing how to depict a new character that appears in a caption during test time; captions do not contain enough information to fully describe the character's appearance. Therefore, in order to generalize to new story elements, the model must have a mechanism for obtaining additional information about how these elements should be visually represented. First, we make story visualization more conducive to these use cases by presenting a new task called 'story continuation'. In this task, we provide an initial scene that can be leveraged in real world use cases. By including this scene, the model can then copy and adapt elements from it as it generates subsequent images. This has the additional benefit of shifting the focus from text-to-image generation, which is already a task attracting plenty of research, and instead focuses on the narrative structure of a sequence of images, e.g., how an image should change over time to reflect new narrative information in the captions. We introduce a new dataset, DiDeMoSV [11], and also convert two existing visualization datasets PororoSV [23] and FlintstonesSV [8] to the story continuation setting.

Next, in order to adapt a text-to-image synthesis model to this story continuation task, we need to finetune the pretrained model (such as DALL-E [33]) on a sequential text-to-image generation task, with the additional flexibility to copy from a prior input. To do so, we first 'retro-fit' the model with additional layers to copy relevant output from the initial scene. Next, we introduce a self-attention block for generating story embeddings that provide global semantic context of the story during generation of each frame. We name this approach StoryDALL-E and also compare with a GAN-based model StoryGANc for story continuation. We also explore the parameter-efficient framework of prompt-

tuning and introduce a prompt consisting of task-specific embeddings to coax the pretrained model into generating visualizations for the target domain. During training, the pretrained weights are frozen and the new parameters are learned from scratch, which is time as well as memory-efficient.

Results show that our retro-fitting approach in StoryDALL-E is useful for leveraging the latent pretrained knowledge of DALL-E for the story continuation task, and outperforms the GAN-based model on several metrics. Further, we find that the copying mechanism allows for improved generation in low-resource scenarios and of unseen characters during inference. In summary,

- We introduce the task of story continuation, that is more closely aligned with downstream applications for story visualization, and provide the community with a new story continuation dataset.
- We introduce StoryDALL-E, an adaptation of pretrained transformers for story continuation, using retro-fitting. We also develop StoryGANc as a strong GAN baseline for comparison.
- We perform comparative experiments and ablations to show that finetuned StoryDALL-E outperforms StoryGANc on three story continuation datasets along several metrics.
- Our analysis shows that the copying mechanism improves correlation of the generated images with the source image, leading to better continuity in the visual story and generation of low-resource as well as unseen characters.

2 Related Work

Text-to-Image Synthesis. Most work in text-to-image synthesis has focused on the development of increasingly sophisticated generative adversarial networks (GANs) [6]. Recent works have leveraged multi-stage generation [48], attentional generative networks [41], dual learning [31], dynamic memory [24, 49], semantic disentaglement [43], explicit object modelling [12] and contrastive loss [17, 47] to further push performance on this task. DALL-E [33] is a large transformer language model that generates both text tokens and image tokens. VideoGPT [42] adapts the DALL-E architecture for conditional generation of videos from a first frame and trains it from scratch. In contrast, we adapt the pretrained DALL-E by *retro-fitting* the pretrained weights with task-specific modules for conditional generation of a sequence of images from a first frame.

Story Visualization. [23] introduce the CLEVR-SV and PororoSV datasets which are based on the CLEVR [16] visual question answering dataset and Pororo video question answering dataset [19] respectively. [27] adapt the Flintstones text-to-video synthesis dataset [8] into FlintstonesSV. While these datasets have served as challenging benchmarks, they contain recurring characters throughout the dataset. Complex datasets, requiring story visualization models to generalize to a more diverse set of test cases is needed to better guide research in this domain. We introduce the story continuation task and propose a new dataset for the task.

Most story visualization models follow the framework introduced in StoryGAN [23], which comprises a recurrent text encoder, an image generator, and image as well as story discriminators to train the GAN [39]. [46] add textual alignment models and a path-based image discriminator, while [21] add dilated convolution and weighted activation degree to the discriminators. [36] add figure-background segmentation to the model in the form of generators and discriminators. [28] and [27] use dual learning and structured inputs respectively to improve story visualization. We use their models as starting point and add modifications that leverage pretrained transformers for our proposed story continuation task.

Parameter-Efficient Training. Methods like adapter-tuning [10, 13, 26, 38] and prompt-based tuning [20, 22] add a small number of trainable parameters to the frozen weights of a pretrained model, which are then learned for the target task. Sparse updating of parameters [7, 45] and low-rank decomposition matrices [14] also provide parameter-efficient methods for finetuning. [9, 29] combine these approaches for a unified approach to finetuning pretrained models. [1] 'retro-fit' a pre-trained language model with cross-attention layers to retrieve relevant tokens at each timestep of word prediction in natural language generation. We use retro-fitting and prompt-tuning to adapt a pretrained image synthesis model to story continuation.

3 Methods

As discussed in Sec. 1, story visualization has limited applicability in real-world settings because the task formulation does not allow models to generalize to new story elements. Hence, we propose the story continuation task and present our StoryDALL-E and StoryGANc models for the task.

3.1 Story Continuation

Given a sequence of sentences $S = [s_1, s_2, ..., s_T]$ forming a narrative, story visualization is the task of generating a corresponding sequence of images $\hat{X} = [\hat{x}_1, \hat{x}_2, ..., \hat{x}_T]$, following [23]. S contains a story, where the captions are temporally ordered and describe the same narrative. This task has many different potential applications such as facilitating the creation of comics or creating visualizations in an educational setting. However, due to the way that the story visualization task is formulated, current models are far from being applied to these settings. The models rely on the images seen in the training data, to generate new visualizations for input stories during the inference phase. Thus, they can only recreate the characters as already found in the training set. Additionally, the captions in story visualization datasets are focused on the narrative, which limits the amount of information that is provided to the model, including descriptions of characters or settings, background etc. Much of this is inferred by the model, leading to generations that might be drastically different than expected, and it is unrealistic to expect the models to generate completely new

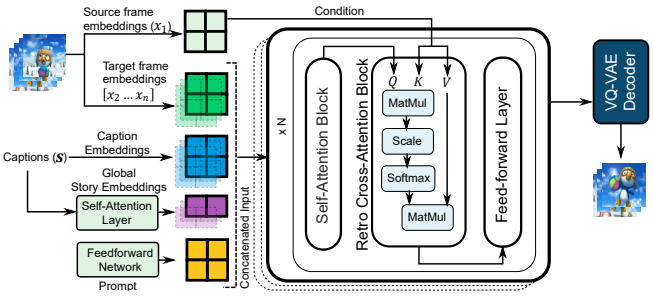

Fig. 1. Illustration of our StoryDALL-E architecture for the prompt-tuning setting. The frames are encoded using pretrained VQVAE and sent as inputs to the pretrained DALL-E. The inputs are prepended with input-agnostic prompt (in prompt-tuning setting only) and global story embeddings corresponding to each sample in the story continuation dataset. The output of StoryDALL-E is decoded using VQ-VAE to generate the predicted image.

visual attributes without sufficient instructions in the caption. Story continuation addresses these issues by providing initial information about the story setting and characters.

In the story continuation task, the first image of the sequence x_1 is provided as additional input to the model. By including an initial ground truth scene as input, the model has access to the appearances of characters, the setting in which the story takes place, and more. When making subsequent scenes, the model then no longer needs to create all the visual features from scratch, but can instead copy from the initial frame. This first image addresses both the generalization issue and the limited information issue in current story visualization models. We refer to this first frame as *source frame* and the remaining frames in the sequence $[x_2, ..., x_t]$ as *target frames*.

3.2 StoryDALL-E

The DALL-E generative network is trained using a simple language-modelling objective on the sequence of discrete image tokens for the task of text-to-image synthesis. With massive amounts of data, such models learn the implicit alignment between text tokens and image tokens, which can be leveraged for downstream tasks like story continuation. The two main aspects that differentiate the story continuation task from text-to-image synthesis are: (1) sequence of captions vs. single caption, and (2) source frame vs. no source frame. Hence, in order to convert the text-to-image synthesis model into a story continuation model, we add three task-specific modules to the native DALL-E architecture. First, we use a global story encoder to pool information from all captions and produce a story embedding, which provides global context of the story at each timestep. Next, we 'retro-fit' the model with cross-attention layers in order to accept the source frame as additional input. Finally, we learn a sequence of embeddings for the story continuation task and provide it as prompt to the model

for task-specific instructions. During finetuning, the pretrained model weights are frozen and these task-specific modules are trained from scratch, leading to a parameter-efficient adaptation of DALL-E for story continuation. We refer to our proposed model as StoryDALL-E (see Figure 1).

Global Story Encoder. Most previous works in story visualization utilize recurrent encoders in the form of LSTM networks [23] or memory-augmented encoders [28], [27], to accept a sequence of captions as input. However, recurrent architectures are memory as well as time-intensive because of sequential processing. Hence, we propose to use a self-attention (f_{self}) based global story encoder, which takes the sentence embeddings for all captions as input and generates contextualized story embeddings for each time-step using parallel processing (see Figure 1). Additionally, we initialize sinusoid positional embeddings (S_{pos}) to provide information about the position of the target frame within the story, and add those to the story embeddings: $S_{global} = f_{self}(S + S_{pos})$. These embeddings are prepended to the word embeddings for the caption at that timestep and sent as input to the generative model.

Retro-fitted Cross-Attention Blocks. Next, we want to 'retro-fit' the DALL-E model with the ability to copy relevant elements from the source image, in order to promote generalizability to unseen visual attributes. This will allow the model to generate visual stories with completely new characters, as long as they are present in the source frame. Hence, we adapt the model to 'condition' the generation of target frame on the source frame by adding a cross-attention block to each self-attention block of the native DALL-E architecture. The image embeddings of the source frame are used in the cross-attention layer as *key* (K) and *value* (V), while the output from the preceding self-attention layer is used as *query* (Q). As shown in Figure 1, the DALL-E self-attention block consists of the self-attention (f^i_{self}), feed-forward (f^i_{dense}) and normalization (f_{norm}) layers. Given an input z_i to the ith self-attention block, the output z^{i+1} is: $z^{i+1} = f_{norm}(f^i_{dense}(f^i_{self}(z_i)))$. In StoryDALL-E, we insert a cross-attention layer such that the output z^{i+1} is:

$$z^{i+1} = f_{norm}(f^i_{dense}(f^i_{cross}(f^i_{self}(z^i), c_{img}))) \tag{1}$$

where f^i_{cross} is the cross-attention layer in the ith Transformer block and c_{image} is sequence of embedding representations for the conditioning image. The self-attention layers are constrained to perform causal masking for computing attention weights due to the nature of the image synthesis task. However, within the cross-attention layer, the input is free to attend over the entire source frame which eases the next token prediction task by augmenting the model with relevant information. The cross-attention layers are trained from scratch.

The StoryDALL-E architecture can be fully fine-tuned to learn the weights of the above-mentioned task-specific modules, while updating the weights of the pretrained model as necessary, on the target task as well as dataset. However, [1]

show that freezing of pretrained weights during training of retro-fitted models can also lead to similar performance as models trained from scratch, with lesser training data. Further, it provides a parameter-efficient approach that can be trained/deployed with a smaller amount of computational resources. Hence, we additionally explore prompt-tuning [22] of the StoryDALL-E model.

Prompt. Prompt-tuning is an alternative [22] to full model fine-tuning where the pretrained model weights are frozen and instead, a small sequence of task-specific vectors is optimized for the downstream task. We initialize a parameterization network $MLP(.)$, which takes a matrix of trainable parameters $P_\theta^{'}$ of dimensions P_{idx} and $dim(h^i)$ as input and generates the prompt P_θ. These trainable matrices are randomly initialized and trained from scratch on the downstream task and dataset. P_θ is appended to the word embeddings of input caption, along with the global story embeddings. Together, these additional embedding vectors act as 'virtual tokens' of a task-specific prompt, and are attended to by each of the caption as image tokens. Formally, the input h^i to the ith self-attention layer in the auto-regressive transformer is organized as follows:

$$h^i = \begin{cases} P_\theta[j,:] & \text{if } j \in [0, P_{idx}) \\ S_{global} & \text{if } j == P_{idx} \\ f^i(z_j, h_{<j}) & \text{otherwise} \end{cases} \tag{2}$$

where $f^i(.)$ is the ith transformer block in StoryDALL-E.

With the aforementioned additions, we convert the pretrained DALL-E into StoryDALL-E model for the story continuation task. A pretrained VQVAE encoder [30] is used to transform RGB images into small 2D grids of image tokens, which are flattened and concatenated with the modified inputs in StoryDALL-E (see supplemen. for details). Finally, StoryDALL-E is trained to model the joint distribution over the tokens of text s and image x: $p(x) = \prod_{j=1}^{d} p(x_j|x_{<i}; s)$. New parameters as well as pretrained weights are optimized in full-model finetuning whereas only the parameters of the prompt, story encoder and cross-attention layers are optimized during prompt-tuning.

3.3 StoryGANc

Generative Adversarial Networks (GANs) have enjoyed steady progress at many image generation tasks such as style transfer [18], conditional image generation [41], image-to-image translation [15] over the last decade. Unlike transformers, they do not need to be pretrained on massive datasets, and can be trained for narrow domains with smaller datasets, which makes it an appealing method. Several recent works in story visualization have demonstrated the effectiveness of GANs for this task [23,28,37]. Hence, we also develop a GAN-based model, StoryGANc, for the story continuation task and compare its performance to that of StoryDALL-E on the proposed datasets (see supplemen. for figure and details). StoryGANc follows the general framework of the StoryGAN model [23]

i.e., it is composed of a recurrent text encoder, an image generation module, and two discriminators - image and story discriminator. We modify this framework to accept the source frame as input for the story continuation task, and use it for improving the generation of target frames. Our StoryGANc model is implemented as follows:

Pre-trained Language Model Encoder. We use a pretrained language model (such as RoBERTa [25] or CLIP text encoder [32]) as the caption encoder. These models are pretrained on large unimodal or multimodal datasets of language, which is of great utility for understanding the semantic concepts present in input captions. To ensure that the model has access to all captions, we append the captions together and use a special token to denote which caption is currently being generated.

Contextual Attention. The story representation from the encoder is combined with the image embeddings of the first frame of the image sequence using contextual attention [44] between the two inputs. The resulting representation is fed through a generator module which recurrently processes each caption, and produces a corresponding image.

Discriminators. The story discriminator takes all of the generated images and uses 3D convolution to create a single representation and then makes a prediction as to whether the generated story is real or fake. The image discriminator performs the same function but only focuses on individual images. The KL-Divergence loss enforces gaussian distribution on the latent representations learnt by GAN. Finally, the model is trained end-to-end using the objective function: $\min_{\theta_G} \max_{\theta_I, \theta_S} \mathcal{L}_{KL} + \mathcal{L}_{img} + \mathcal{L}_{story}$, where θ_G, θ_I and θ_S denote the parameters of the text encoder + image generator, and image and story discriminators respectively. During inference, the trained weights θ_G are used to generate a visual story for a given input of captions.

4 Datasets

Since story continuation is a reframing of the story visualization tasks, existing story visualization datasets can be adapted for story continuation by assigning the first frame in the sequence as source frame and the rest as target frames. However, such existing story visualization datasets like PororoSV [23] and FlintstonesSV [8] are also homogeneous datasets with recurring characters i.e., the characters used during evaluation already appear in the training set. It is not possible to evaluate the generalization capacity of story continuation models using these datasets. Hence, we propose a new dataset in this paper.

DiDeMoSV. DiDeMo [11] is a video captioning dataset containing 10,000 short clips with more than 40,000 text descriptions temporally localized with the videos. Each of the clips were randomly sampled from the YFCC100M [40]

Fig. 2. Examples from the PororoSV (top), FlintstonesSV (middle) and DiDeMoSV (bottom) datasets. In the story continuation setting, the first frame is used as input to the generative model.

dataset which is based upon Flickr. This results in videos that cover a large breadth of real-world scenarios, containing many different settings, actions, entities, and more. The dataset contains 11550/2707/3378 samples in training, validation and test respectively, with each sample containing three consecutive frames. This dataset challenges story continuation models to generate diverse inputs, covering many more story elements, in contrast to existing story visualization datasets. In order to do this, models must maximize their usage of the initial scene input and need to incorporate additional general visual knowledge, whether this is done through transfer learning or additional data.

We also use the existing PororoSV [23] and FlintstonesSV datasets [8], containing 10191/2334/2208 and 20132/2071/2309 samples respectively, to evaluate our story continuation models. Each sample contains 5 consecutive frames. There are 9 and 7 main characters in PororoSV and FlintstonesSV respectively, that appear throughout the dataset. For story continuation, we use the first frame as source frame and the rest of the four frames in the sequence as target frames. Evaluation is only performed on the generation of target frames. See Figure 2 for examples from the three story continuation datasets.

5 Experiments

We use the pretrained weights from popular open-source minDALL-E (1.3B parameters) which is trained on 14 million text-image pairs from the CC3M [35] and CC12M [2] datasets, to initialize our models.[2] minDALL-E uses the

[2] https://github.com/kakaobrain/minDALL-E

Table 1. Results on the test sets of PororoSV, FlintstonesSV and DiDeMoSV (DSV) datasets from various models. Scores are based on FID (lower is better), character classification F1 and frame accuracy (F-Acc.; higher is better) evaluations.

Model	PororoSV			FlintstonesSV			DSV
	FID ↓	Char-F1↑	F-Acc↑	FID ↓	Char-F1↑	F-Acc↑	FID↓
StoryGANc (BERT)	72.98	**43.22**	17.09	91.37	70.45	55.78	91.43
StoryGANc (CLIP)	74.63	39.68	16.57	90.29	72.80	**58.39**	92.64
StoryDALL-E (prompt)	61.23	29.68	11.65	53.71	42.48	32.54	64.58
StoryDALL-E (finetuning)	**25.90**	36.97	**17.26**	**26.49**	**73.43**	55.19	**32.92**

pretrained VQGAN-VAE [4] for discretizing image inputs. We experiment with pretrained CLIP [32] (38M parameters) and distilBERT [34] (110M parameters) text encoders for the StoryGANc models. The StoryDALL-E models are trained for 5 epochs with learning rates of 1e-04 (AdamW, Cosine Scheduler) and 5e-04 (AdamW, Linear Decay Scheduler) for full-model fine-tuning and prompt-tuning setups respectively. Checkpoints are saved at the end of every epoch. The StoryGANc models are trained for 120 epochs with learning rates 1e-04 and 1e-05 for the generator and discriminators respectively. Checkpoints are saved every 10 epochs. These models are trained on single A6000 GPUs.

We use the FID score for saving the best checkpoints in our experiments. The FID score calculates the difference between the ground truth and generated images by computing the distance between two feature vectors. Following [23] and [28], we also compute the character classification scores (F1 Score and Frame Acc.) for the PororoSV and FlintstonesSV datasets. See supplemen. for details.

6 Results

Main Quantitative Results. Table 1 contains the FID, character classification F1 score and frame accuracy results on the test sets of PororoSV and FlintstonesSV datasets using various models in our experiments. We train two variations of the StoryDALL-E model with the distilBERT and CLIP text encoders. Our model StoryDALL-E is trained under two settings, one where the pretrained weights are frozen during training and the other where the pretrained weights are also finetuned on the target dataset. In practice, we find it necessary to finetune the pretrained text and image embeddings within the Transformers, which are pretrained on real-world images, in order to adapt them to different domains such as cartoons. This results in nearly 30% trainable parameters during prompt-tuning, as compared to full-model finetuning. With StoryDALL-E, we see drastic improvements in FID score for the PororoSV and FlinstonesSV datasets, over the StoryGANc model, demostrating the superior visual quality of the generated visual stories. The character classification scores remain the same for FlintstonesSV and drop by 6% and 14% for PororoSV with use of finetuned and prompt-tuned StoryDALL-E respectively. GAN-based models like StoryGANc are able to recreate distinct and finer details of a character which leads to higher accuracy scores using a classification model, such as

Table 2. Ablation results of StoryDALL-E on validation sets of PororoSV, FlintstonesSV and DiDeMoSV (DSV) datasets. Scores are based on FID (lower is better), character classification F1 and frame accuracy (F-Acc.; higher is better) evaluations.

Model	PororoSV			FlintstonesSV			DSV
	FID ↓	Char-F1↑	F-Acc↑	FID ↓	Char-F1↑	F-Acc↑	FID↓
StoryDALL-E	21.64	40.28	20.94	28.37	74.28	52.35	41.58
- Cross-Attention	30.45	39.32	34.65	35.04	73.94	53.28	55.89
- Story Embeddings	23.27	40.25	18.16	29.21	72.18	52.72	42.34
- Story Embeddings & Cross-Attention	31.68	35.29	16.73	36.28	72.44	51.32	58.14

the Inception-v3 used in our experiments [28]. With prompt-tuning, we observe that StoryDALL-E models manage to capture the background elements of the scene but fail to properly recreate the characters in the frame. The frame accuracy score, which is based on exact match overlap of multiple characters in the predicted scene with those in ground truth, remains low for all models, suggesting that both methods struggle to compose multiple roles in a single image [3].

For the more challenging DiDeMoSV dataset, the fully finetuned StoryDALL-E model outperforms the GAN models by a wide margin in terms of FID score. It should be noted here that PororoSV and FlintstonesSV have a finite set of recurring animated characters throughout the dataset, whereas DiDeMoSV is derived from a multitude of real-world scenarios with no overlap in characters between training and evaluation sets. While the addition of a source frame makes it easier for the model to replicate it in the target frames, the generation is significantly more difficult due to the diversity in evaluation samples. However, since the DiDeMoSV dataset contains images from the real-world domain, the pretrained knowledge of StoryDALL-E derived from Conceptual Captions is useful for generating relevant and coherent images for the dataset, while StoryGANc largely fails to do so.

Ablations. Table 2 contains results from ablation experiments on finetuned StoryDALL-E on the validation sets of the three story continuation datasets. The primary modifications we make to DALL-E in order to adapt it into StoryDALL-E, are the cross-attention layers and global story embeddings. We perform minus-one experiments on StoryDALL-E by removing each of these components and observing the effect on FID results on validation sets. First, we remove the cross-attention layers from StoryDALL-E, which reverts the model to the story visualization setting where the model no longer receives the first image as input, and is evaluated on generation of rest of the frames in the visual story. With this ablation, we see large increase in FID scores across all datasets. Without a source image to guide the generated output, the quality of illustration drops rapidly, especially for the new DiDeMo dataset. The removal of global story embeddings results in a text-to-image synthesis setting with the first frame as additional input. In this scenario, we see smaller drops in FID, indicating that the global context is not as important as the ability to copy from an initial image. In the third row, we remove both, cross-attention layers and story embeddings, which

Table 3. Results from human evaluation (Win% / Lose% / Tie%). Win% = % times stories from StoryDALL-E was preferred over StoryGANc, Lose% for vice-versa. Tie% represents remaining samples.

Dataset	Visual Quality	Relevance	Consistency
PororoSV	94/0/6	44/28/28	56/26/18
FlintstonesSV	90/2/8	32/38/30	42/32/26
DiDeMoSV	64/0/36	38/0/62	32/48/20

relegates the setting to a text-to-image synthesis task, and observe large increase in FID scores across all datasets.

6.1 Human Evaluation

We additionally conduct human evaluation on our model's outputs hoping to better capture the overall quality of the generated stories. We have a human annotator compare generated visual stories from our StoryDALL-E (finetuning) and StoryGANc (BERT) models. They are provided with predictions from each dataset and the corresponding ground truth captions, and asked to pick the better prediction (or tie) in terms of visual quality, consistency, and relevance [23]. Results are presented in Table 3. The StoryDALL-E model outperforms StoryGANc model in terms of visual quality and relevance, achieving higher % of wins in each of the three datasets (except relevance in FlintstonesSV). These results follow from the fact that StoryDALL-E uses the VQGAN-VAE [4] which is designed for reconstructing higher resolution images. Moreover, it has access to large pretraining data, which improves alignment between semantic concepts in captions and regions in images. We see wins in terms of consistency for PororoSV and DiDeMoSV predictions from StoryDALL-E models. But, the absolute numbers for consistency and relevance show that there is still room for improvement.

7 Analysis

In this section, we perform experiments to analyze aspects of the StoryDALL-E model and the story continuation task. First, we perform qualitative analyses of the predictions from StoryDALL-E. Next, we quantify the effect of the retrofitted cross-attention layers and visualize the attention heads. See supplemen. for an analysis of the diverse semantic content in the DiDeMoSV dataset.

7.1 Qualitative Analysis

Figure 3 contains sampled outputs from both of our models for the three story continuation datasets. In each of these examples, StoryDALL-E generates higher quality images than StoryGANc. The difference is especially stark for PororoSV and FlintstonesSV datasets since StoryDALL-E is exposed to the

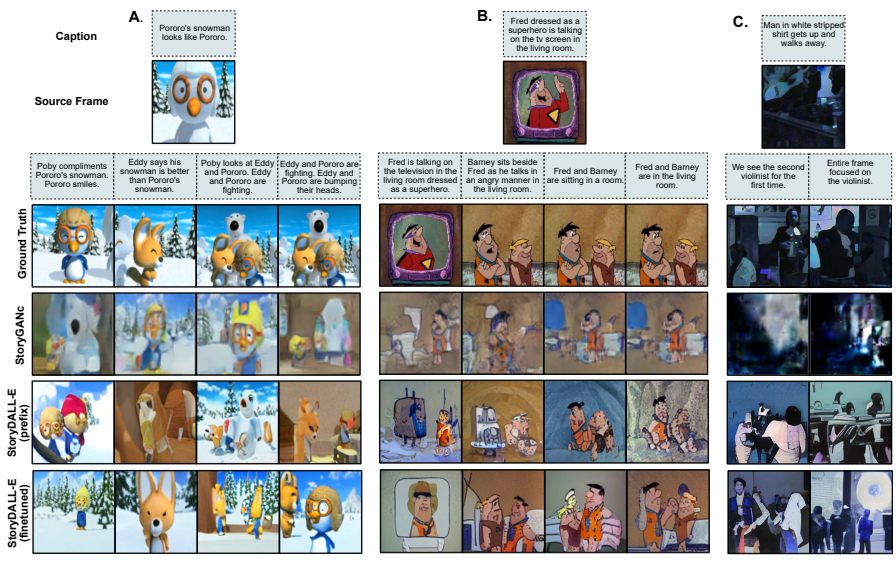

Fig. 3. Examples of predictions for (A) PororoSV (B) FlintstonesSV and (C) DiDeMoSV story continuation datasets from StoryDALL-E and StoryGANc models. Source frame refers to the initial frame provided as additional input to the model.

characters during training and has additional guidance from source frame during inference. In the case of DiDeMoSV, the generations from StoryGANc are largely incomprehensible, which could be attributed to the unseen semantic concepts such as 'violinist' which did not appear in the training set. In contrast, StoryDALL-E is exposed to various real-world concepts during pretraining, which can be leveraged during generation. For instance, the pretrained knowledge as well as the copying mechanism help the StoryDALL-E model comprehend 'television' and generate an image for 'Fred is talking in the television' (see Fig.3(b)). However, the overall quality of the images from StoryDALL-E also do not approach human produced images. As discussed in Sec. 6, it is especially true for frames containing multiple characters. This suggests that while current models are able to attempt the task, there is still much work to be done before consistent and coherent images are commonly produced by the models.

We also examine the ability of StoryDALL-E to recreate scarce characters from the training set (see Fig. 4(a)) and generate unseen characters (see Fig. 4(b)), when guided by the copying mechanism via cross-attention layers. We find that the copying mechanism allows for better generation of shape and form for less-frequent characters in PororoSV. Similarly, we identified non-recurring characters in the FlintstonesSV dataset and observed the corresponding generated images, when StoryDALL-E has access to a previous frame where they appear. StoryDALL-E succeeds at partially copying visual aspects of the characters, such as the purple skirt (top) and blue uniform (bottom).

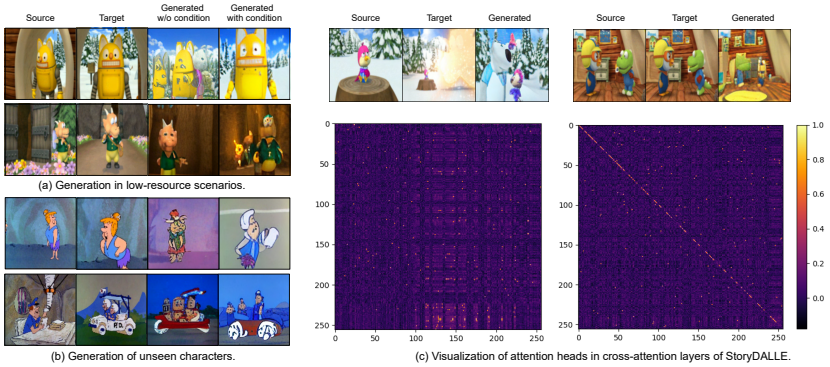

Fig. 4. Examples of generation from StoryDALL-E in (a) low-resource scenarios and (b) of unseen characters. (c) Plots of attention scores computed in retro cross-attention layers for examples of source frames (x-axis) and target frames (y-axis).

7.2 Retro-fitted Cross-Attention

We examine the attention scores computed in the retro cross-attention layer and present examples in Fig. 4(c). The cross-attention layer in StoryDALL-E receives vector representations for the source image and computes the cross-attention output using source frame as key/value and target frame as query. In the first example (left), the target frame is copying visual attributes of the pink bird with the most emphasis, as be seen from the higher attention scores for the image tokens roughly in the center of the source frame. For the second example (right), the source frame and target frames are nearly similar; the attention scores are highest in the diagonal of the plot. The resulting images in both samples contain many visual attributes already found in the source image, demonstrating that the cross-attention layer is effective at enabling conditional image generation. See supplementary for correlation scores between source image and frames generated with and without condition using StoryDALL-E.

8 Conclusion

We introduce a new task called story continuation in order to make the story visualization task more conducive for real-world use cases. We present a new dataset DiDeMoSV, in addition to reformatting two existing story visualization datasets for story continuation. Our model StoryDALL-E, based on a retro-fitting approach for adapting pretrained transformers, out-performs GAN-based models on the story continuation datasets. We hope that the dataset and models motivate future work in this area.

Acknowledgement. We thank the reviewers for their useful feedback. This work was supported by ARO Award W911NF2110220, DARPA KAIROS Grant FA8750-19-2-1004, NSF-AI Engage Institute DRL-211263. The views, opinions, and/or findings contained in this article are those of the authors, not the funding agency.

References

1. Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Driessche, G.v.d., Lespiau, J.B., Damoc, B., Clark, A., et al.: Improving language models by retrieving from trillions of tokens. arXiv preprint arXiv:2112.04426 (2021) 4, 6
2. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3558–3568 (2021) 9
3. Cho, J., Zala, A., Bansal, M.: Dall-eval: Probing the reasoning skills and social biases of text-to-image generative transformers. arXiv preprint arXiv:2202.04053 (2022) 11
4. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12873–12883 (2021) 1, 10, 12
5. Frans, K., Soros, L., Witkowski, O.: Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. arXiv preprint arXiv:2106.14843 (2021) 2
6. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: NeurIPS (2014) 3
7. Guo, D., Rush, A.M., Kim, Y.: Parameter-efficient transfer learning with diff pruning. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 4884–4896 (2021) 4
8. Gupta, T., Schwenk, D., Farhadi, A., Hoiem, D., Kembhavi, A.: Imagine this! scripts to compositions to videos. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 598–613 (2018) 2, 3, 8, 9
9. He, J., Zhou, C., Ma, X., Berg-Kirkpatrick, T., Neubig, G.: Towards a unified view of parameter-efficient transfer learning. arXiv preprint arXiv:2110.04366 (2021) 4
10. Henderson, J., Ruder, S., et al.: Compacter: Efficient low-rank hypercomplex adapter layers. In: Advances in Neural Information Processing Systems (2021) 4
11. Hendricks, L.A., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with natural language. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017) 2, 8
12. Hinz, T., Heinrich, S., Wermter, S.: Semantic object accuracy for generative text-to-image synthesis. IEEE transactions on pattern analysis and machine intelligence (2020) 3
13. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp. In: International Conference on Machine Learning. pp. 2790–2799. PMLR (2019) 4
14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021) 4
15. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017) 2, 7
16. Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2901–2910 (2017) 3

17. Kang, M., Park, J.: Contragan: Contrastive learning for conditional image generation. In: NeurIPS (2020) 3
18. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019) 2, 7
19. Kim, K.M., Heo, M.O., Choi, S.H., Zhang, B.T.: Deepstory: video story qa by deep embedded memory networks. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. pp. 2016–2022 (2017) 3
20. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 3045–3059 (2021) 4
21. Li, C., Kong, L., Zhou, Z.: Improved-storygan for sequential images visualization. Journal of Visual Communication and Image Representation **73**, 102956 (2020). https://doi.org/https://doi.org/10.1016/j.jvcir.2020.102956, http://www.sciencedirect.com/science/article/pii/S1047320320301826 4
22. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 4582–4597 (2021) 4, 7
23. Li, Y., Gan, Z., Shen, Y., Liu, J., Cheng, Y., Wu, Y., Carin, L., Carlson, D., Gao, J.: Storygan: A sequential conditional gan for story visualization. In: Proceedings of the IEEE Conference on CVPR. pp. 6329–6338 (2019) 2, 3, 4, 6, 7, 8, 9, 10, 12
24. Liang, J., Pei, W., Lu, F.: Cpgan: full-spectrum content-parsing generative adversarial networks for text-to-image synthesis. arXiv preprint arXiv:1912.08562 (2019) 3
25. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019) 8
26. Mahabadi, R.K., Ruder, S., Dehghani, M., Henderson, J.: Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 565–576 (2021) 4
27. Maharana, A., Bansal, M.: Integrating visuospatial, linguistic, and commonsense structure into story visualization. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 6772–6786 (2021) 2, 3, 4, 6
28. Maharana, A., Hannan, D., Bansal, M.: 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. pp. 2427–2442 (2021) 2, 4, 6, 7, 10, 11
29. Mao, Y., Mathias, L., Hou, R., Almahairi, A., Ma, H., Han, J., Yih, W.t., Khabsa, M.: Unipelt: A unified framework for parameter-efficient language model tuning. arXiv preprint arXiv:2110.07577 (2021) 4
30. Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al.: Conditional image generation with pixelcnn decoders. Advances in neural information processing systems **29** (2016) 7
31. Qiao, T., Zhang, J., Xu, D., Tao, D.: Mirrorgan: Learning text-to-image generation by redescription. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1505–1514 (2019) 3

32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021) 1, 8, 10
33. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning. pp. 8821–8831. PMLR (2021) 1, 2, 3
34. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In: 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing (NeurIPS) (2019) 10
35. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2556–2565 (2018) 9
36. Song, Y.Z., Rui Tam, Z., Chen, H.J., Lu, H.H., Shuai, H.H.: Character-preserving coherent story visualization. In: European Conference on Computer Vision. pp. 18–33. Springer (2020) 4
37. Song, Y.Z., Tam, Z.R., Chen, H.J., Lu, H.H., Shuai, H.H.: Character-preserving coherent story visualization. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020) 2, 7
38. Sung, Y.L., Cho, J., Bansal, M.: Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. arXiv preprint arXiv:2112.06825 (2021) 4
39. Szűcs, G., Al-Shouha, M.: Modular storygan with background and theme awareness for story visualization. In: International Conference on Pattern Recognition and Artificial Intelligence. pp. 275–286. Springer (2022) 4
40. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Communications of the ACM **59**(2), 64–73 (2016) 8
41. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1316–1324 (2018) 3, 7
42. Yan, W., Zhang, Y., Abbeel, P., Srinivas, A.: Videogpt: Video generation using vq-vae and transformers. arXiv preprint arXiv:2104.10157 (2021) 3
43. Yin, G., Liu, B., Sheng, L., Yu, N., Wang, X., Shao, J.: Semantics disentangling for text-to-image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2327–2336 (2019) 3
44. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5505–5514 (2018) 8
45. Zaken, E.B., Ravfogel, S., Goldberg, Y.: Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. arXiv preprint arXiv:2106.10199 (2021) 4
46. Zeng, G., Li, Z., Zhang, Y.: Pororogan: An improved story visualization model on pororo-sv dataset. In: Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence. pp. 155–159 (2019) 4
47. Zhang, H., Koh, J.Y., Baldridge, J., Lee, H., Yang, Y.: Cross-modal contrastive learning for text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 833–842 (2021) 3

48. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: ICCV (2017) 3
49. Zhu, M., Pan, P., Chen, W., Yang, Y.: Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5802–5810 (2019) 3