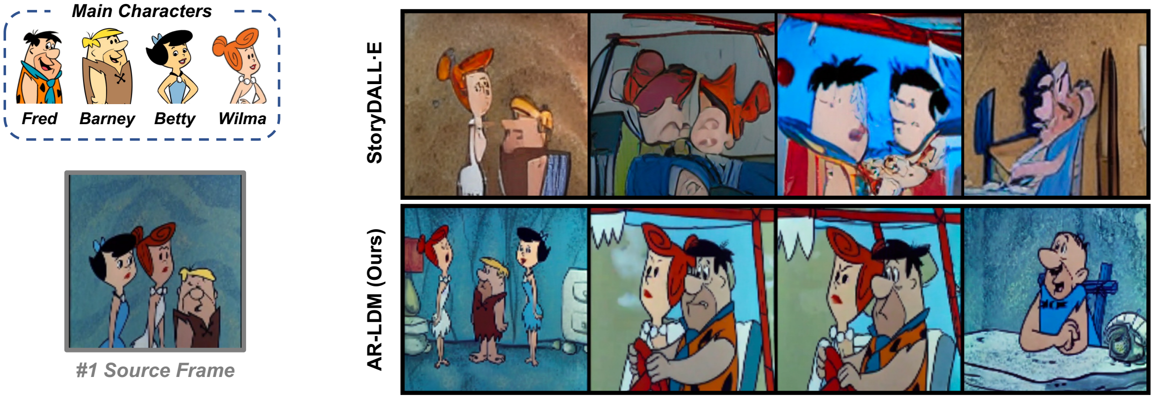
Synthesizing Coherent Story with Auto-Regressive Latent Diffusion Models

Xichen Pan[1]*, Pengda Qin[2], Yuhong Li[2], Hui Xue[2], Wenhu Chen[3]
[1]New York University, [2]Alibaba Group, [3]University of Waterloo



#1. Wilma, Betty and Barney are in the room. They are all three standing while talking to one another.

***#2.** Wilma, Betty, and Barney are standing in the living room. Wilma is talking.*
***#3.** Fred and Wilma are driving in the car.*
***#4.** Fred is driving the car while listening to Wilma who is the passenger. Wilma looks angry while speaking to Fred as she has her arms crossed.*
***#5.** The man in blue with a bow tie is sitting with his hands on a desk in the room. He is talking and then shakes his head while talking.*

Figure 1. Comparison of a visual story example synthesized by AR-LDM (Ours) and StoryDALL·E [20] on FlintstonesSV story continuation dataset. The visual stories are generated with reference to the source frame and captions.

Abstract

Conditioned diffusion models have demonstrated state-of-the-art text-to-image synthesis capacity. Recently, most works focus on synthesizing independent images; While for real-world applications, it is common and necessary to generate a series of coherent images for story-stelling. In this work, we mainly focus on story visualization and continuation tasks and propose AR-LDM, a latent diffusion model auto-regressively conditioned on history captions and generated images. Moreover, AR-LDM can generalize to new characters through adaptation. To our best knowledge, this is the first work successfully leveraging diffusion models for coherent visual story synthesizing. It also extends the text-conditioned method to multimodal conditioning. Quantitative results show that AR-LDM achieves SoTA FID scores on PororoSV, FlintstonesSV, and the adopted challenging dataset VIST containing natural images. Large-scale human evaluations show that AR-LDM has superior performance in terms of quality, relevance, and consistency. Code available at this https URL

1. Introduction

Recently advanced diffusion models [31] such as DALL·E 2 [24], Imagen [29], and Stable Diffusion [26] have shown unprecedented text-to-image synthetic capacities. These models focus on single-image generation, while many real-world use cases like comic drawing require models to generate a series of coherent images according to a long story description. Text-to-image models offer extreme freedom to guide creation through natural language. Simply generating each image according to every single caption will result in poor relevance and consistency. Textual Inversion [4] and DreamBooth [28] have focused on creating a specific unique concept to guide consistent generation results across images. Re-Imagen [2] is able to generate specific entities with reference to retrieved image text pairs in a training-free manner. However, how to generate a series of coherent images illustrating a multi-sentence paragraph is still underexplored.

In this paper, we mainly focus on two tasks: story visualization [16] and story continuation [20]. Story visualization aims at synthesizing a series of images to describe a story containing multiple sentences. Story continuation is a vari-

*Contribution during internship at Alibaba Group.

ant of story visualization with the same goal as story visualization, but additionally based on a source frame (i.e., the first frame). This setting addresses the generalization issue and limited information issue in story visualization, allowing models to generate more meaningful and coherent images. Story visualization and continuation are challenging tasks requiring both vision-language understanding and image generation. Previous works are mainly based on GANs and auto-regressive models, and utilize contextual text encoders to improve consistency. While, as the saying goes, "*A picture is worth a thousand words,*" it is impossible for a single caption to exploit all necessary information for image generation. There are thousands of reasonable illustrations for a given story. For example, for the story shown in Fig. 1, the captions of the third and fourth frames do not describe the detail of the "*car*" or background. The key to generating coherent stories is to preserve as many details across images as possible. The main limitation of existing work is that the generation is guided only by contextual text conditions without leveraging previously generated images.

In this work, we propose Auto-Regressive Latent Diffusion Model (AR-LDM) to leverage diffusion models to synthesize coherent stories. Specifically, we employ a history-aware encoding module containing a CLIP text encoder [22], and a BLIP multimodal encoder [15]. For each frame, AR-LDM is guided by a multimodal condition containing both the current caption embedding and previously generated image-caption history embedding. This allows AR-LDM to generate relevant and coherent images. As shown in Fig. 1, AR-LDM shows strong multimodal understanding and image generation ability. It is able to precisely generate the scene as captions described in high quality, as well as keeping a strong consistency across frames. Additionally, we also explore adapting AR-LDM to preserve consistency for unseen characters (i.e., characters referred by a pronoun, like the man in the last frame of Fig. 1) within the stories. This adaptation can largely alleviate the inconsistent generation results caused by uncertain descriptions of unseen characters.

To evaluate our method, we utilize two widely accepted datasets, FlintstonesSV and PororoSV, as our test bed. While all existing story visualization and continuation datasets are cartoon images[1], we adopt the VIST [10] dataset for this task to better evaluate real-world story synthesis capacity. VIST contains story-in-sequence (SIS) captions that better match real-world use cases, and also provides description-in-isolation (DII) style captions. Quantitative evaluation results show our method achieves SoTA performance in both story visualization and continuation tasks. In particular, AR-LDM achieves an FID score of 16.59 on PororoSV, with a relative improvement of 55% over previous story visualization methods. AR-LDM also boosts story continuation performance with a relative improvement of approximately 20% on all evaluation datasets. We also conduct large-scale human evaluations to test our method's visual quality, relevance, and consistency, which shows that humans mostly prefer our synthesized stories over previous methods.

Our contribution can be summarized as follows:

1. We propose a history-aware auto-regressive conditioned latent diffusion model AR-LDM, which first leverages diffusion models for story synthesis.
2. We propose a multimodal conditioning module, extend the use of text conditioned diffusion model to a broader field as a general-purpose image decoder.
3. We go beyond cartoon stories and adopt the VIST dataset for real-world story synthesis.
4. For more practical application, we additionally propose a simple but efficient adaptation method to allow AR-LDM generalizing to unseen characters.

2. Related work

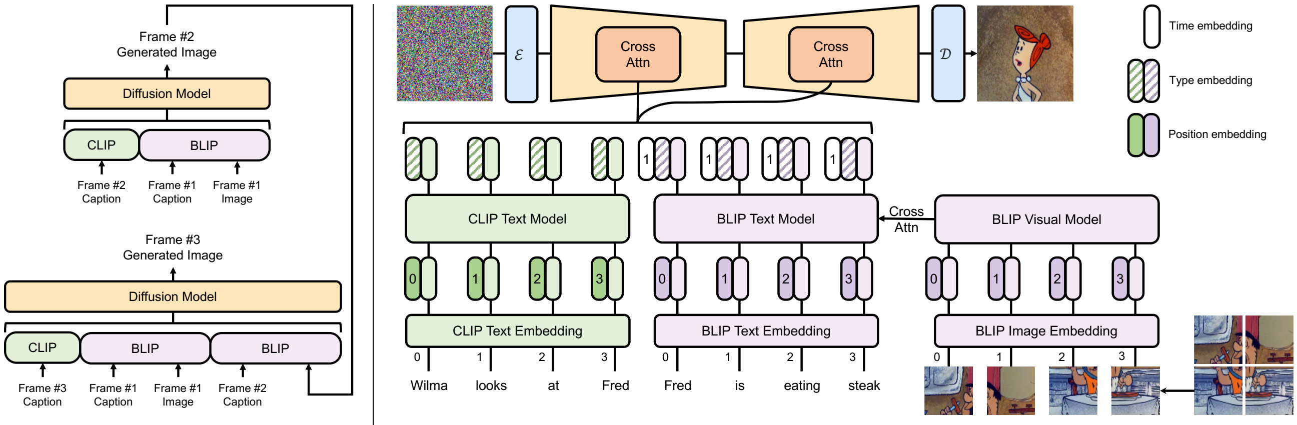
2.1. Text-to-Image Synthesis

Recent advances in text-to-image synthesis mainly focus on generative adversarial networks (GANs) [5], auto-regressive models, and diffusion models. GANs like Stackgan [41], Attngan [38], Mirrorgan [21], and MXC-GAN [40] perform adversarial training between generators and discriminators to learn to generate high-quality images. Large auto-regressive models like DALL·E [25], Make-A-Scene [3], and Parti [39] can be easily scaled up and have also shown their excellent image synthetic capacity. Recently, success in diffusion models has attracted many researchers' attention. As likelihood-based models, diffusion models do not suffer from mode-collapse and potentially unstable training as GANs, and can generate more diversified images. Additionally, diffusion models are more parameter-effective than auto-regressive models. Unlike prior diffusion models mostly relied on text conditioning, AR-LDM extends the text conditional generation method to multimodal conditioning. Concurrent work M-VADER [37] shares a similar idea with us, but they focus on single image synthesis and do not utilize the auto-regressive method.

2.2. Story Synthesis

StoryGAN [16] firstly proposes the story visualization task. Most story visualization models are based on GANs and comprise context text encoder, image generator, separate image, and story discriminators. The context text encoder and story discriminators mainly aim to preserve image consistency. DUCO-StoryGAN [19] uses dual learning and copy-transform to improve story visualization. The copy-transform mechanism first incorporates features from

[1] the DiDeMoSV [20] dataset is a cartoon-style real-world dataset



(a) Auto-regression process.

(b) Model Architecture.

Figure 2. Overview of proposed AR-LDM. The blue blocks represent the perceptual compression models; the orange blocks denote the generative network. The green blocks and the purple blocks are the history-aware conditioning network. Illustration inspired by [12].

previously generated images through the attention mechanism to improve consistency. VLC-StoryGAN [18] and Word-Level SV [14] focus on text inputs, and propose to use structured input and sentence representation to better guide visual story generation. VP-CSV [1] leverages transformer [35], and VQ-VAE [34] to preserve characters across generated images. StoryDALL·E [20] turns to story continuation, a variant of story visualization based on a given source image. They retrofit the pre-trained transformers DALL·E [25] and achieve a drastic improvement over GAN-based models. Concurrent work Make-A-Story [23] also leverages diffusion model for story visualization. They maintain consistency through a visual auto-regressive plugin module. While AR-LDM utilizes a multimodal auto-regressive conditioning module which does not require any modifications to the U-Net architecture. This allows vision language cross attention and implicit grounding, thereby leading to superior performance.

3. Method

3.1. Preliminaries

Diffusion models [31] define a Markov chain of forward diffusion process q to gradually add Gaussian noise sampled to real data $\mathbf{z}_0 \sim q(\mathbf{z})$ in T steps. In particular, $\mathbf{z}$ in this paper denotes latent representations instead of pixels. The forward process $q(\mathbf{z}_t|\mathbf{z}_{t-1})$ at each time step t is:

$$
\begin{aligned}
q(\mathbf{z}_t|\mathbf{z}_{t-1}) &= \mathcal{N}(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t\mathbf{I}) \\
q(\mathbf{z}_{1:T}|\mathbf{z}_0) &= \prod_{t=1}^{T} q(\mathbf{z}_t|\mathbf{z}_{t-1})
\end{aligned}
\tag{1}
$$

in which $\beta_t \in (0,1)$ denotes the step size. Note $\beta_{t-1} < \beta_t$.

Diffusion models learn a UNet [27] denoted as $\boldsymbol{\epsilon}_\theta$ to reverse the forward diffusion process, constructing desired data samples from the noise. Let $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$. We can reparameterize the denoising process $p(\mathbf{z}_{t-1}|\mathbf{z}_t)$ also as a Gaussian distribution. It can be estimated by $\boldsymbol{\epsilon}_\theta$ and has a form of the following:

$$
\begin{aligned}
& p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t) = \mathcal{N}(\mathbf{z}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{z}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{z}_t, t)) \\
\text{with} \quad & \boldsymbol{\mu}_\theta(\mathbf{z}_t, t) = \frac{1}{\sqrt{\alpha_t}}(\mathbf{z}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t))
\end{aligned}
\tag{2}
$$

The learning objective of diffusion models is to approximate the mean $\boldsymbol{\mu}_\theta(\mathbf{z}_t, t)$ in the reverse diffusion process. We can use variational lower bound (ELBO) [13] to minimize the negative log-likelihood of $p_\theta(\mathbf{z}_0)$ [8], the simplified objective can be written as a denoising objective:

$$
\mathcal{L} = \mathbb{E}_{\mathbf{z}_0, \boldsymbol{\epsilon} \sim \mathcal{N}(0,1), t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t)\|^2\right] \tag{3}
$$

During inference, [9] proposes to use classifier-free guidance to obtain more relevant generation results.

$$
\hat{\boldsymbol{\epsilon}} = w \cdot \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \varphi, t) - (w-1) \cdot \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t) \tag{4}
$$

where w is guidance scale, φ denotes the condition.

3.2. Auto-Regressive Latent Diffusion Model

As we discussed in Sec. 1, different from single caption text-to-image task, synthesizing coherent stories requires the model to be aware of history descriptions and scenes. For instance, consider a story "*A red metallic cylinder cube is at the center. Then add a green rubber cube at the right.*" present in [16]. The second sentence alone cannot give enough guidance to generate a coherent image. It is crucial for the model to understand the history caption, the scene, and the appearance of the "*red metallic cylinder cube*" in the first generated image. The key point of designing a strong story synthesis model is to make it capable of incorporating history captions and scenes for current image generation.

In this work, we propose auto-regressive latent diffusion model (AR-LDM) to achieve better consistency across frames. As shown in Fig. 2a, AR-LDM leverages the history captions and images for future frame generation. For a certain story with a length of L, let $\mathbf{C} = [\mathbf{c}_1, \cdots, \mathbf{c}_j, \cdots, \mathbf{c}_L]$ be input captions and $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_j, \cdots, \mathbf{x}_L]$ be the image targets, each caption $\mathbf{c}_j$ is corresponding to an image $\mathbf{x}_j \in \mathbb{R}^{C \times H \times W}$. Existing works assume conditional independence between each frame and generate the whole visual story according to the captions. While AR-LDM gets rid of this assumption by additionally conditioned on history images $\hat{\mathbf{x}}_{<j}$ and directly estimating the posterior based on the chain rule, which has a form of

$$\begin{aligned} P_{\text{AR}}(\mathbf{X}|\mathbf{C}) &= \prod_{j=1}^{L} P(\mathbf{x}_j|\hat{\mathbf{x}}_{<j}, \mathbf{C}) \\ &\approx \prod_{j=1}^{L} P(\mathbf{x}_j|\hat{\mathbf{x}}_{<j}, \mathbf{c}_{\leq j}) \\ &= \prod_{j=1}^{L} P(\mathbf{x}_j|\tau_\theta(\hat{\mathbf{x}}_{<j}, \mathbf{c}_{\leq j})) \\ &= \prod_{j=1}^{L} p_\theta(\mathbf{z}_0^{[j]}|\tau_\theta(\mathcal{D}(\mathbf{z}_0^{[<j]}), \mathbf{c}_{\leq j})) \end{aligned} \tag{5}$$

where p_θ is the reverse diffusion process reparameterized by the generative network $\boldsymbol{\epsilon}_\theta$, and τ_θ denotes the history-aware conditioning network. $\mathcal{D}$ denotes the decoder of the perceptual compression model (i.e., VQ-VAE), which also contains an encoder $\mathcal{E}$. To avoid abuse of notations, we use $\mathbf{z}_t^{[j]}$ to denote the latent diffusion variable at j-th frame and t-th diffusion step. It should be noted that AR-LDM is based on a causal method that is only conditioned on current and previous captions instead of the entire captions. It offers greater flexibility and practicality in real-world applications such as comic drawing. AR-LDM allows users to generate images and iteratively refine prompts in a back-and-forth, dialogue-like manner, which enables them to add all desired details before moving on to the next image. The $\approx$ in Eq. (5) reflect the assumption that in a storytelling scenario, the details for image $\mathbf{x}_j$ only come from history and current captions rather than future captions. Fig. 2b shows the detailed architecture of AR-LDM.

Generative network Following [26], AR-LDM also performs the forward and reverse diffusion processes in an efficient, low-dimensional latent space. The latent space is approximately perceptually equivalent to high-dimensional RGB space, while the redundant semantically meaningless information in pixels is eliminated. Specifically, perceptual compression models consisting of $\mathcal{E}$ and $\mathcal{D}$ are trained to encode the real data into the latent space and reverse, such that $\mathcal{D}(\mathcal{E}(\mathbf{x})) \approx \mathbf{x}$. AR-LDM uses latent representations $\mathbf{z} = \mathcal{E}(\mathbf{x})$ instead of pixels during the diffusion process. The final output can be decoded back to pixel space with $D(\mathbf{z})$. The separate mild perceptual compression stage only eliminates imperceptible details, allowing the model to achieve competitive generation results at a much lower cost.

History-Aware Conditioning Network We use a history-aware conditioning network to encode the history caption-image pairs into a multimodal condition $\varphi_j = \tau_\theta(\hat{\mathbf{x}}_{<j}, \mathbf{c}_{\leq j})$ to guide denoising process $p_\theta(\mathbf{z}_t^{[j]}|\varphi_j)$ directly through cross attention, ensuring AR-LDM can effectively handle long stories generation in a complexity of $O(L)$. The estimated noise in Eq. (3) can be rewritten as $\boldsymbol{\epsilon}_\theta(\mathbf{z}_t^{[j]}, \varphi_j, t)$. Therefore, $P(\mathbf{x}_j|\hat{\mathbf{x}}_{<j}, \mathbf{C})$ in Eq. (5) can be simplified as $p_\theta(\mathbf{z}_0^{[j]}|\varphi_j)$. The conditioning network consists of CLIP [22] and BLIP [15], in charge of current caption encoding and previous caption-image encoding, respectively. BLIP is pre-trained using vision-language understanding and generation tasks with large-scale filtered clean web data. BLIP utilizes the cross-attention module to deeply integrate visual and language modalities. It is able to ground the entities generated in history frames, allowing the generative network to refer to history scenes.

In summary, AR-LDM can generate image $\hat{\mathbf{x}}_j$ through:

$$\begin{aligned} \overline{\mathbf{c}}_j &= \text{CLIP}(\mathbf{c}_j) \\ \overline{\mathbf{m}}_{<j} &= \left[\text{BLIP}(\mathbf{c}_1, \hat{\mathbf{x}}_1); \cdots; \text{BLIP}(\mathbf{c}_{j-1}, \hat{\mathbf{x}}_{j-1}))\right] \\ \varphi_j &= \left[\overline{\mathbf{c}}_j + \mathbf{c}^{type}; \overline{\mathbf{m}}_{<j} + \mathbf{m}^{type} + \mathbf{m}^{time}_{<j}\right] \\ \mathbf{z}_0^{[j]} &\sim p_\theta(\mathbf{z}_0^{[j]}|\varphi_j) \\ \hat{\mathbf{x}}_j &= \mathcal{D}(\mathbf{z}_0^{[j]}) \end{aligned} \tag{6}$$

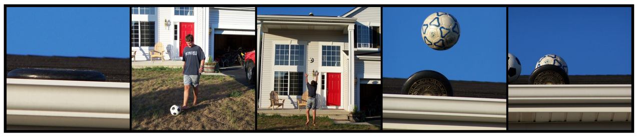
where $\overline{\mathbf{m}}_{<j}$ denotes encoded multimodal features from previous captions and generated images. $\mathbf{c}^{type}, \mathbf{m}^{type} \in \mathbb{R}^D$ are text and multimodal type embedding, respectively. $D = 768$ denotes the embedding dimension. $\mathbf{m}^{time} \in \mathbb{R}^{L \times D}$ is time embedding. Specifically, the first image $\mathbf{x}_1$ is provided as input for the story continuation setting.

3.3. Adaptive AR-LDM

For real-world applications like comic drawing, it's necessary to preserve consistency for the new (unseen) characters. As we discussed in Sec. 1, it is challenging because one cannot depict every single detail of the unseen character in captions, and the story synthesis model always suffers from the inconsistent descriptions of a certain unseen character like the generated results of AR-LDM shown in Fig. 7. Inspired by Textual Inversion [4] and DreamBooth [28], we add a new token `<char>` to represent the unseen character, and adapt the trained AR-LDM to generalize to the specific unseen character. Specifically, the embedding of the new token `<char>` is initialized by that of a similar existing word, like "*man*" or "*woman*". Then we finetune the whole parameters of AR-LDM on only a single story composed of 3-5 images of the character.

Models	# of Params	PororoSV	FlintstonesSV	VIST-SIS	VIST-DII
StoryGANc (2022) [20]	-	74.63	90.29	-	-
StoryDALL·E (prompt tuning, 2022) [20]	1.3B	61.23	53.71	-	-
StoryDALL·E (2022) [20]	1.3B	25.90	26.49	-	-
MEGA-StoryDALL·E (2022) [20]	2.8B	23.48	23.58	20.98*	24.61*
AR-LDM (Ours)	1.5B	**17.40**	**19.28**	**16.95**	**17.03**

Table 1. Story continuation FID scores (lower is better) of AR-LDM and several previous methods. * denotes experimental results reproduced by us, where we trained MEGA-StoryDALL·E for 50 epochs using the same training strategies as AR-LDM.



SIS

1. A discus got stuck up on the roof.
2. Why not try getting it down with a soccer ball?
3. Up the soccer ball goes.
4. It didn't work so we tried a volley ball.
5. Now the discus, soccer ball, and volleyball are all stuck on the roof.

DII

1. A black frisbee is sitting on top of a roof.
2. A man playing soccer outside of a white house with a red door.
3. The boy is throwing a soccer ball by the red door.
4. A soccer ball is over a roof by a frisbee in a rain gutter.
5. Two balls and a frisbee are on top of a roof.

Figure 3. A data sample from VIST with description-in-isolation (DII), and story-in-sequence (SIS) captions.

4. Experiments

4.1. Datasets

We use three datasets as our testbed, PororoSV [16], FlintstonesSV [18], and VIST [10]. Each story in these datasets contains 5 consecutive frames. For story visualization, we predict all 5 frames from captions. For story continuation, the first frame is assigned as a source frame, and we generate the rest 4 frames with reference to the source frame. We will briefly go through these three datasets, a more detailed introduction can be found in Appendix A.

PororoSV and FlintstonesSV The PororoSV [16] and FlintstonesSV [18] datasets are adapted from Pororo video question answering dataset [11] and Flintstones text-to-video synthesis dataset [6], respectively. Both two datasets contain several recurring characters. While FlintstonesSV is relatively harder than PororoSV, for there are many unseen characters within the stories.

VIST However, there are two major limitations of existing story synthesis datasets: (1) current datasets are all cartoon ones; (2) sentences are isolated descriptions rather than sequential stories. We propose to use the Visual Story Telling (VIST) dataset [10] for real-world story synthesizing. VIST provides two kinds of captions: description-in-isolation (DII) and story-in-sequence (SIS). As shown in Fig. 3, DII captions are more like the ones in PororoSV and FlintstonesSV, every single caption contains detailed information about the image. In contrast, SIS captions describe the five images like a story and merely repeat the content mentioned before. The story-style captions are closer to real-world use cases and require the model to have a better contextual understanding ability.

Models	FID
StoryGAN (2019) [16]	158.06
CP-CSV (2020) [33]	149.29
DUCO-StoryGAN (2021) [19]	96.51
VLC-StoryGAN (2021) [18]	84.96
VP-CSV (2022) [1]	65.51
Word-Level SV (2022) [14]	56.08
Make-A-Story (2022) [23]	36.64
AR-LDM (Ours)	**16.59**

Table 2. Story visualization FID score results on PororoSV. We use the results reported by [1, 14, 23].

4.2. Experimental Settings

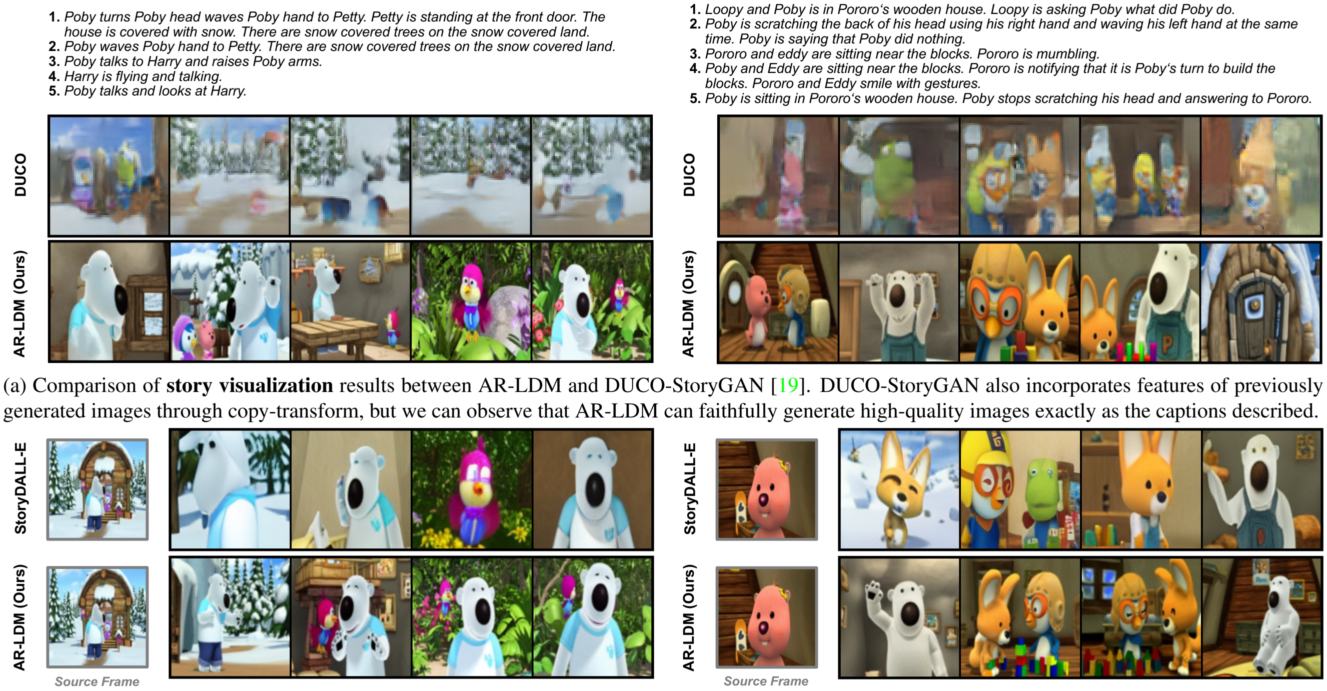
Our model is initialized by the weight of Stable Diffusion [26], a publicly available text-to-image LDM trained on LAION-5B [30]. We trained AR-LDM for 50 epochs on 8 NVIDIA A100-80GB GPUs for two days. We only freeze the encoder $\mathcal{E}$ and decoder $\mathcal{D}$, and finetune the rest parameters using the AdamW optimizer [17] with an initial learning rate of 1×10^{-5} and a weight decay of 10^{-4}. During training, we randomly drop the condition φ at a probability of 0.1 for each frame. A cosine scheduler and 8000 steps learning rate warm-up are used during training. During inference, we sample images using the DDIM scheduler [32] for 250 inference steps with guidance scale w set to 6.0.

5. Results

We evaluate AR-LDM using two settings: (1) quantitative evaluation using automatic metric FID score [7]; (2) large-scale human evaluations regarding visual quality, relevance, and consistency. Note that we report FID scores at a resolution of 64×64 following [20] for a fair comparison.

5.1. Quantitative Results

Story Visualization In Tab. 2, we present our story visualization results on PororoSV, which is the most commonly



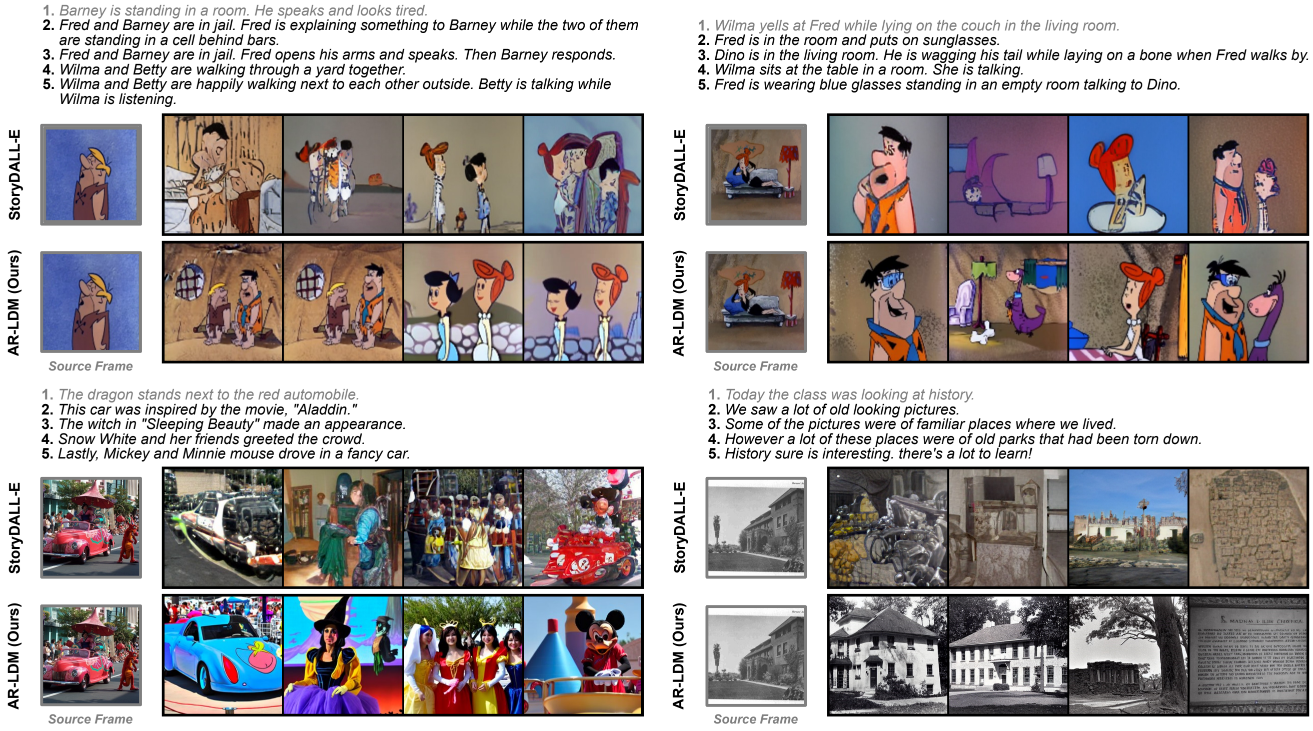
1. Poby turns Poby head waves Poby hand to Petty. Petty is standing at the front door. The house is covered with snow. There are snow covered trees on the snow covered land.
2. Poby waves Poby hand to Petty. There are snow covered trees on the snow covered land.
3. Poby talks to Harry and raises Poby arms.
4. Harry is flying and talking.
5. Poby talks and looks at Harry.

1. Loopy and Poby is in Pororo's wooden house. Loopy is asking Poby what did Poby do.
2. Poby is scratching the back of his head using his right hand and waving his left hand at the same time. Poby is saying that Poby did nothing.
3. Pororo and eddy are sitting near the blocks. Pororo is mumbling.
4. Poby and Eddy are sitting near the blocks. Pororo is notifying that it is Poby's turn to build the blocks. Pororo and Eddy smile with gestures.
5. Poby is sitting in Pororo's wooden house. Poby stops scratching his head and answering to Pororo.

(a) Comparison of **story visualization** results between AR-LDM and DUCO-StoryGAN [19]. DUCO-StoryGAN also incorporates features of previously generated images through copy-transform, but we can observe that AR-LDM can faithfully generate high-quality images exactly as the captions described.

(b) Comparison of **story continuation** results between AR-LDM and StoryDALL·E. [20]

Figure 4. Visual story synthesis results on PororoSV. Note the case in Fig. 4a and Fig. 4b is the same one.



1. Barney is standing in a room. He speaks and looks tired.
2. Fred and Barney are in jail. Fred is explaining something to Barney while the two of them are standing in a cell behind bars.
3. Fred and Barney are in jail. Fred opens his arms and speaks. Then Barney responds.
4. Wilma and Betty are walking through a yard together.
5. Wilma and Betty are happily walking next to each other outside. Betty is talking while Wilma is listening.

1. Wilma yells at Fred while lying on the couch in the living room.
2. Fred is in the room and puts on sunglasses.
3. Dino is in the living room. He is wagging his tail while laying on a bone when Fred walks by.
4. Wilma sits at the table in a room. She is talking.
5. Fred is wearing blue glasses standing in an empty room talking to Dino.

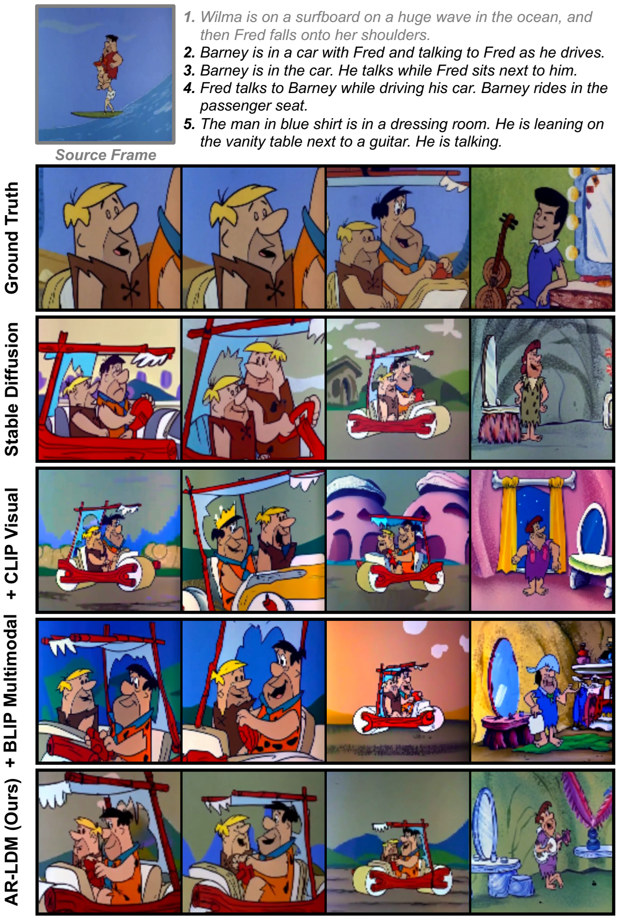
1. The dragon stands next to the red automobile.
2. This car was inspired by the movie, "Aladdin."
3. The witch in "Sleeping Beauty" made an appearance.
4. Snow White and her friends greeted the crowd.
5. Lastly, Mickey and Minnie mouse drove in a fancy car.

1. Today the class was looking at history.
2. We saw a lot of old looking pictures.
3. Some of the pictures were of familiar places where we lived.
4. However a lot of these places were of old parks that had been torn down.
5. History sure is interesting. there's a lot to learn!

Figure 5. Comparison of story continuation results between AR-LDM and StoryDALL·E on FlintstonesSV (upper) and VIST-SIS (lower). Better visual quality, relevance, and consistency can be observed in the visual stories synthesized by AR-LDM.

Models	FID	Δ
Stable Diffusion [26]	22.10	-
+ CLIP Visual Conditioning	21.01	-1.09
Stable Diffusion [26]	22.10	-
+ BLIP Multimodal Conditioning	19.94	-2.16
+ Auto Regression	**19.28**	-2.82

Table 3. Ablation study results for story continuation task on FlintstonesSV. All models are finetuned using the same training data and strategies as AR-LDM.



1. Wilma is on a surfboard on a huge wave in the ocean, and then Fred falls onto her shoulders.
2. Barney is in a car with Fred and talking to Fred as he drives.
3. Barney is in the car. He talks while Fred sits next to him.
4. Fred talks to Barney while driving his car. Barney rides in the passenger seat.
5. The man in blue shirt is in a dressing room. He is leaning on the vanity table next to a guitar. He is talking.

Figure 6. Synthesized visual story examples for different ablation models mentioned in Tab. 3

Dataset	Criterion	Win (%)	Tie (%)	Lose (%)
PororoSV	Visual Quality	**90.6**	2.2	7.2
	Relevance	**88.6**	1.0	10.4
	Consistency	**70.6**	9.4	20.0
FlintstonesSV	Visual Quality	**99.4**	0.2	0.4
	Relevance	**96.2**	1.0	2.8
	Consistency	**93.2**	6.0	0.8
VIST-SIS	Visual Quality	**86.6**	6.6	6.8
	Relevance	**69.8**	6.8	23.4
	Consistency	**81.6**	8.2	10.2

Table 4. Human evaluation results of story continuation task on PororoSV, FlintstonesSV, and VIST-SIS datasets. Win means AR-LDM is preferred over StoryDALL·E; Lose for vice-versa; Tie denotes the samples that human annotators can hardly choose.

used benchmark by previous works. AR-LDM makes significant progress and achieves a SoTA FID score of 16.59, surpassing previous methods and even concurrent diffusion model-based method [23] by a large margin. As shown in Fig. 4a, AR-LDM is able to generate high-quality, coherent visual stories while faithfully reproducing character details and scenes. More cases can be found in Appendix D.1.

Story Continuation We test the story continuation performance of AR-LDM and present the results in Tab. 1. AR-LDM achieves a series of new SoTA FID scores on all four datasets. Notably, AR-LDM outperforms MEGA-StoryDALL·E with around half parameters. As shown in Fig. 4b, AR-LDM can preserve consistency through auto-regression generation, like the background of the last two frames in the left case, as well as the blocks in the third and fourth frames in the right case. We further demonstrate this consistency across frames on FlintstonesSV and VIST-SIS datasets in Fig. 5. For example, the jail in the upper left case and the sunglasses in the upper right case. Moreover, we show that AR-LDM can infer from previous captions and scenes. In the lower right case, AR-LDM generates a black-and-white story from the black-and-white source image and caption containing ”old looking” and ”history”, without relying on captions describing the color of photos, which is more reasonable than the results of StoryDALL·E. More cases can be found in Appendix D.

Ablation Studies We conduct ablation studies on the proposed auto-regressive multimodal module. Particularly, we use the story continuation task on FlintstonesSV as our test bed, for it requires higher consistency across frames and contains many challenging unseen characters. Starting from a finetuned Stable Diffusion baseline, we utilize a source frame image to guide the generation through CLIP visual encoder and obtain an FID score improvement of 1.09. If we further leverage the source frame image caption pair and use BLIP to encode it into a multimodal embedding to condition the diffusion process, we can observe an improvement of 2.16 compared to the baseline model. Finally, we employ an auto-regressive generation manner and achieve a further improvement of 0.66. However, the FID score is only related to visual quality. Apart from visual quality, AR-LDM also performs better in terms of relevance and consistency, we provide a case in FlintstonesSV to illustrate it. As shown in Fig. 6, compared to other methods, AR-LDM with auto-regressive generation manner better preserves backgrounds and view of the scene across frames.

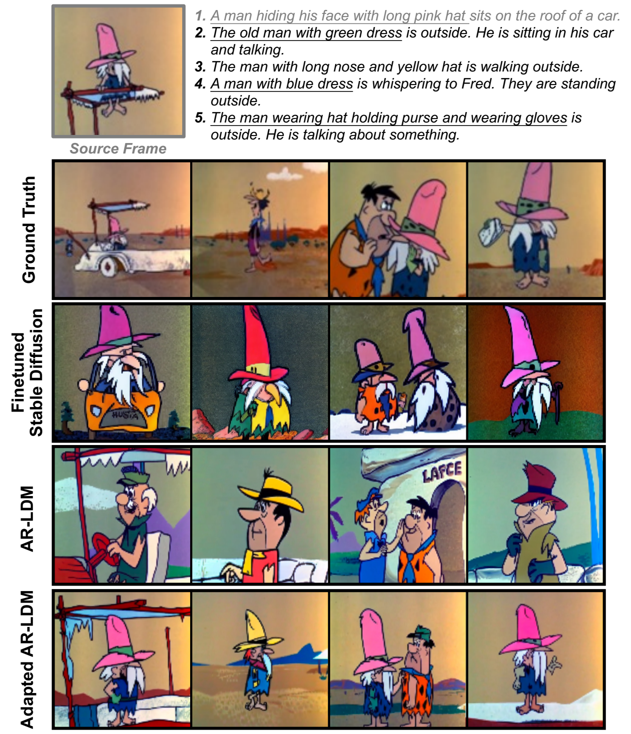

1. A man hiding his face with long pink hat sits on the roof of a car.
2. The old man with green dress is outside. He is sitting in his car and talking.
3. The man with long nose and yellow hat is walking outside.
4. A man with blue dress is whispering to Fred. They are standing outside.
5. The man wearing hat holding purse and wearing gloves is outside. He is talking about something.

Figure 7. Adaptation results for a case AR-LDM failed to properly generate on FlintstonesSV. The underlined texts refer to one specific person and can be replaced by `<char>` in adapted AR-LDM.

5.2. Large-Scale Human Evaluation

We also carry out large-scale human evaluations for the story continuation task on PororoSV, FlintstonesSV, and VIST-SIS datasets in terms of visual quality, relevance, and consistency. It is necessary because the FID score only measures visual quality. We enlist the services of a third-party annotation team to ensure fairness in comparisons. The team consists of five skilled, full-time annotators, whose involvement helps to minimize human error and maintain the quality of evaluations. For detailed human evaluation settings, see Appendix B. The annotation team conducted a comparison between the synthesized stories of AR-LDM and StoryDALL·E. We randomly select 500 samples of each dataset to be evaluated for each criterion. Our evaluation scale is 10 times larger than that of StoryDALL·E, providing a more comprehensive and precise result. The evaluation results are shown in Tab. 4. Owing to the powerful diffusion model, AR-LDM significantly outperforms StoryDALL·E in visual quality. The history-aware conditioning network also largely boosts AR-LDM's relevance and consistency. Cases in Appendix C can showcase the annotation quality.

5.3. Adapting to Unseen Characters

As shown in Fig. 7, all underlined texts are referring the same character (i.e., the man with a pink hat in the source frame), while the description is inconsistent. As a result, AR-LDM generates three different characters according to every single description. After being finetuned on 3-5 images, adapted AR-LDM can generate consistent characters as well as faithfully synthesize scenes and characters as captions describe. In contrast, simply finetuning Stable Diffusion using the same data cannot obtain satisfying results, because it confuses other characters with `<char>` and fails to generate them. More cases can be found in Appendix E.

Dataset	Criterion	Win (%)	Tie (%)	Lose (%)
PororoSV	Visual Quality	41.8	17.4	40.8
	Relevance	18.0	28.6	53.4
	Consistency	3.8	3.2	93.0
FlintstonesSV	Visual Quality	42.2	20.0	37.8
	Relevance	24.6	26.4	49.0
	Consistency	2.6	13.2	84.2
VIST-SIS	Visual Quality	14.6	20.6	64.8
	Relevance	19.2	48.6	32.2
	Consistency	3.0	46.2	50.8

Table 5. Human evaluation results of story continuation on PororoSV, FlintstonesSV, and VIST-SIS datasets. Comparison between AR-LDM synthesized visual stories and ground truth references.

6. Limitations

Though AR-LDM achieves an unprecedented synthesis capability, largely outperforms StoryDALL·E in the human evaluation as discussed in Sec. 5.2, we find that AR-LDM is still far behind ground truth visual stories in terms of consistency. As the human evaluation results shown in Tab. 5, AR-LDM is comparable to ground truth visual stories regarding visual quality and relevance; We can also observe that 49.2% of generated stories on VIST are as consistent as ground truth. However, as for more challenging PororoSV and FlintstonesSV datasets whose frames are sampled from videos, we find that few synthesized visual stories are as consistent as ground truth references. This indicates consistency is a short board of current models, and it still needs to be improved in the future. We believe the gap with ground truth can be narrowed with a stronger conditioning module, additional discussion can be found in Appendix F.

7. Conclusion

We present AR-LDM, the first work employing diffusion models for story visualization and continuation, which demonstrates unprecedented effectiveness in generating coherent visual stories in high quality. AR-LDM incorporates captions and previously generated images into current frame generation through an auto-regressive manner to preserve consistency. The multimodal conditioning also extends the use of text conditioned diffusion model to a wider range as a general-purpose image decoder.

References

[1] 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 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8259–8272, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. 3, 5

[2] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *ArXiv preprint*, abs/2209.14491, 2022. 1

[3] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. *ArXiv preprint*, abs/2203.13131, 2022. 2

[4] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *ArXiv preprint*, abs/2208.01618, 2022. 1, 4

[5] 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*, 63(11):139–144, 2020. 2

[6] 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 (ECCV)*, pages 598–613, 2018. 5

[7] 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 Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6626–6637, 2017. 5

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 3

[9] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *ArXiv preprint*, abs/2207.12598, 2022. 3

[10] Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, San Diego, California, 2016. Association for Computational Linguistics. 2, 5

[11] Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. Deepstory: Video story QA by deep embedded memory networks. In Carles Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 2016–2022. ijcai.org, 2017. 5

[12] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR, 2021. 3

[13] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 3

[14] Bowen Li. Word-level fine-grained story visualization. In *European Conference on Computer Vision*, pages 347–362. Springer, 2022. 3, 5

[15] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR, 2022. 2, 4

[16] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David E. Carlson, and Jianfeng Gao. Storygan: A sequential conditional GAN for story visualization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6329–6338. Computer Vision Foundation / IEEE, 2019. 1, 2, 3, 5

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 5

[18] Adyasha Maharana and Mohit Bansal. Integrating visuospatial, linguistic, and commonsense structure into story visualization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6772–6786, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 3, 5

[19] 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*, pages 2427–2442, Online, 2021. Association for Computational Linguistics. 2, 5, 6

[20] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. *ArXiv preprint*, abs/2209.06192, 2022. 1, 2, 3, 5, 6

[21] Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1505–1514. Computer Vision Foundation / IEEE, 2019. 2

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 2, 4

[23] Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. Make-a-story: Visual memory conditioned consistent story generation. *ArXiv preprint*, abs/2211.13319, 2022. 3, 5, 7

[24] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv preprint*, abs/2204.06125, 2022. 1

[25] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR, 2021. 2, 3

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 4, 5, 7, 27

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[28] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *ArXiv preprint*, abs/2208.12242, 2022. 1, 4

[29] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *ArXiv preprint*, abs/2205.11487, 2022. 1

[30] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *ArXiv preprint*, abs/2210.08402, 2022. 5

[31] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2256–2265. JMLR.org, 2015. 1, 3

[32] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 5

[33] Yun-Zhu Song, Zhi Rui Tam, Hung-Jen Chen, Huiao-Han Lu, and Hong-Han Shuai. Character-preserving coherent story visualization. In *European Conference on Computer Vision*, pages 18–33. Springer, 2020. 5

[34] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6306–6315, 2017. 3

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. 3

[36] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *ArXiv preprint*, abs/2208.10442, 2022. 27

[37] Samuel Weinbach, Marco Bellagente, Constantin Eichenberg, Andrew Dai, Robert Baldock, Souradeep Nanda, Björn Deiseroth, Koen Oostermeijer, Hannah Teufel, and Andres Felipe Cruz-Salinas. M-vader: A model for diffusion with multimodal context. *ArXiv preprint*, abs/2212.02936, 2022. 2

[38] 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 *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1316–1324. IEEE Computer Society, 2018. 2

[39] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *ArXiv preprint*, abs/2206.10789, 2022. 2

[40] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-

image generation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 833–842. Computer Vision Foundation / IEEE, 2021. 2

[41] Han Zhang, Tao Xu, and Hongsheng Li. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5908–5916. IEEE Computer Society, 2017. 2