Make-A-Story: Visual Memory Conditioned Consistent Story Generation

Tanzila Rahman[1,3] Hsin-Ying Lee[2] Jian Ren[2] Sergey Tulyakov[2]
Shweta Mahajan[1,3] Leonid Sigal[1,3,4]
[1]University of British Columbia [2]Snap Inc.
[3]Vector Institute for AI [4]Canada CIFAR AI Chair

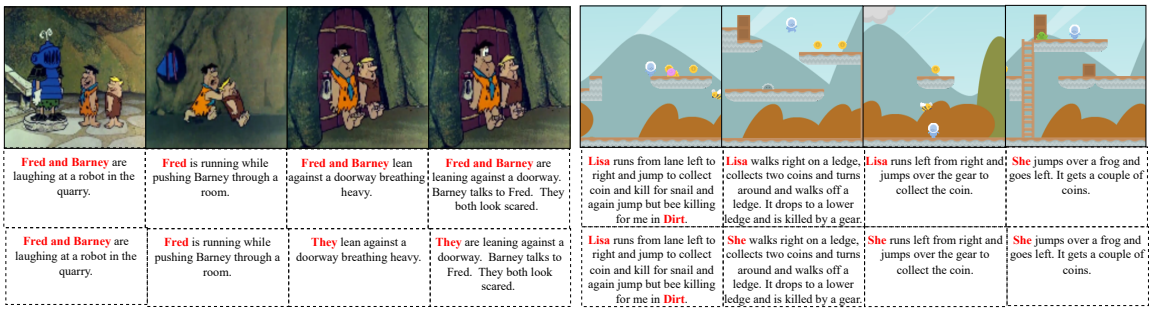


Figure 1. **Referential and consistent story visualization.** Examples of more *natural* stories with references for the FlintstonesSV [14] and MUGEN [16] datasets (bottom text) compared to more typical but less natural story text (top). We extend the MUGEN dataset by introducing additional two characters (*e.g.* Lisa and Jhon) and four backgrounds (*e.g.* Sand, Grass, Stone and Dirt).

Abstract

There has been a recent explosion of impressive generative models that can produce high quality images (or videos) conditioned on text descriptions. However, all such approaches rely on conditional sentences that contain unambiguous descriptions of scenes and main actors in them. Therefore employing such models for more complex task of story visualization, where naturally references and co-references exist, and one requires to reason about when to maintain consistency of actors and backgrounds across frames/scenes, and when not to, based on story progression, remains a challenge. In this work, we address the aforementioned challenges and propose a novel autoregressive diffusion-based framework with a visual memory module that implicitly captures the actor and background context across the generated frames. Sentence-conditioned soft attention over the memories enables effective reference resolution and learns to maintain scene and actor consistency when needed. To validate the effectiveness of our approach, we extend the MUGEN dataset [16] and introduce additional characters, backgrounds and referencing in multi-sentence storylines. Our experiments for story generation on the MUGEN, the PororoSV [27] and the FlintstonesSV [14] dataset show that our method not only outperforms prior state-of-the-art in generating frames with high visual quality, which are consistent with the story, but also models appropriate correspondences between the characters and the background.

1. Introduction

Multimodal deep learning approaches have pushed the quality and the breadth of conditional generation tasks such as image captioning [20, 30, 34, 50, 53] and text-to-image synthesis [22, 41, 55, 57–59]. Owing to the technical leaps made in generative models, such as generative adversarial networks (GANs) [12], variational autoencoders (VAEs) [24] and the more recent diffusion models [18], approaches for text-to-image synthesis can now generate images with high visual fidelity representative of the textual descriptions. The captions, however, in such cases, are generally short self-contained sentences representing the high-level semantics of a scene. This is rather restrictive in the real-world applications [6, 27, 28] where fine-grained understanding of object interactions, motion and background information described by multiple sentences becomes necessary. One such task is that of *story generation* or *visualization* – the goal of which is to generate a sequence of illustrative image frames with coherent semantics given a sequence of sentences [27, 32, 33, 56].

Characteristic features of a good visual story is high visual quality over multiple frames; this includes rendering of discernible objects, actors, poses and realistic interactions of those actors with objects and within the scene. Moreover, for text-based story generation it is crucial to maintain consistency between the generated frames and the multi-sentence descriptions. Not only the actor context, but also the background of the generated story should be in-line with the description demonstrating effortless transition and adaptation to the changing environments within the story [31].

Recent advances on the task of story generation have made significant advances along these lines, showing high visual fidelity and character consistency for story sentences that are self-contained and unambiguous (explicitly mentioning characters and the setting each time). While impressive, this setup is fundamentally unrealistic. Realistic story text is considerably more complex and referential in nature; requiring ability to resolve ambiguity and references (or co-references) through reasoning. As shown in Fig. 1, while the description corresponding to the first frame has an explicit reference to the character names, the (typical) subsequent frame descriptions, provided by human, contain references such as "*she, he, they*". Moreover, while maintaining character consistency, current approaches are limited in preserving, or transitioning through, the background information in agreement with the text (*cf*. Fig. 3) [27, 32, 33].

In natural language processing (NLP) co-reference resolution in text is an important and core task [3, 23, 35]. While it maybe possible to apply such methods to story text to first resolve ambiguous references and then generate corresponding images using existing story generation approaches, this is sub-optimal. The reason, is that co-reference resolution in the text domain, at best, would only allow to resolve references and maintain consistency across *identity* of the character. Appearance across frames would still lack consistency and require some form of visual reasoning. As also noted in [44], reference resolution in the visual domain, or visio-lingual domain, is more powerful.

In this work, for the first time (to our knowledge), we study co-reference resolution in story generation. Prior work [32] offers limited performance when faced with text containing references (see Sec. 5). We address this by proposing a new autoregressive diffusion-based framework with a visual memory module that implicitly captures the actor and background context across the generated frames. Sentence-conditioned soft attention over the memories enables effective visio-lingual co-reference resolution and learns to maintain scene and actor consistency when needed. Further, given the lack of datasets that contain references and more complex sentence structure, we extend the MUGEN dataset [16] and introduce additional characters, backgrounds and referencing in multi-sentence storylines.

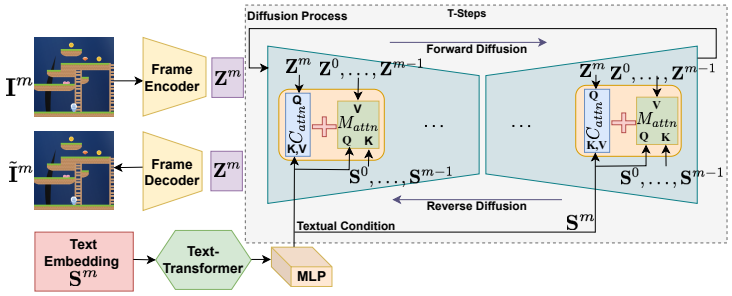
Contributions. Our contributions are three-fold: *(i)* First, we introduce a novel autoregressive deep generative framework, *Story-LDM*, that adopts and extends latent diffusion models for the task of story generation. As part of Story-LDM, we propose a meticulously designed memory-attention mechanism capable of encoding and leveraging contextual relevance between the part of the story-line that has already been generated, and the current frame being generated based on learned semantic similarity of corresponding sentences. Equipped with this, our sequential diffusion model can generate consistent stories by resolving and then capturing temporal character and background context. *(ii)* Second, to validate our approach for co-reference resolution, and character and background consistency in the visual domain, we extend existing datasets to include more complex scenarios and, importantly, referential text. Specifically, we extend the MUGEN dataset [16] to include multiple characters and diverse backgrounds. We also modify FlintstonesSV [14] and PororoSV [27] dataset to include character references. These enhancements allow us to increase the complexity of the aforementioned datasets by introducing co-references in the sentences of a story. *(iii)* Finally, to evaluate different approaches for foreground (character) as well as background consistency we propose novel evaluation metrics. Our results on the MUGEN [16], the PororoSV [27] and the FlintstonesSV [14] datasets show that we outperform the prior state-of-the-art on consistency metrics by a large margin.

2. Related work

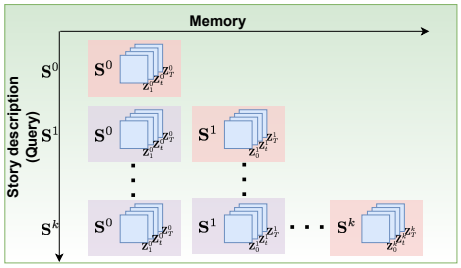
Text-to-image synthesis. Deep generative models, particularly, generative adversarial networks (GANs) [12], variational autoencoders (VAEs) [24] and normalizing flows [4, 9, 10] have been applied to multimodal tasks at the intersection of vision and language. Typical such tasks include image captioning [1, 30, 52] and text-to-image synthesis [29, 41, 55, 57, 59]. Early work on text conditioned image synthesis built upon the success of GANs [41]. More recent approaches have utilized multi-stage generators [57] and normalizing flow-based priors [29] in the latent space to model the distribution of images given text. Various approaches have found cross-domain contrastive loss to improve text-to-image generation models [22, 58]. DALL-E [39] and Cogview [8] harness the power of transformers [49] and discrete variational autoencoders (VQ-VAE) [40] yielding very high quality image samples.

More recent are the advances in diffusion models which have revolutionized the domain of image generation [18]. Diffusion models progressively add noise to the data and learn a reverse diffusion process to reconstruct it. Nichol *et al*. [36] adapted diffusion models for text-to-image generation and explore CLIP [37] guided generation as well as classifier-free modeling. Standard diffusion models are employed directly in the high-dimensional pixel space and therefore, cannot directly be used for the more complex task of story generation. Recent work [13,42] instead use encodings from pre-trained models as input to the diffusion models, thereby reducing the complexity of the task by working in a lower-dimensional space. In this work, we build upon this idea and extend it for sequential story generation.

Text-to-video synthesis. One of the challenges of text-to-video synthesis is the smoothness of motion in a video [27]. Early work focused on generating short clips [6, 28]. To



(a) Story-LDM architecture for conditional story generation.



(b) Memory state with query and values for our memory-attention module.

Figure 2. **Story-Latent Diffusion Models for consistent story generation.** *(a)* The autoregressive conditional generative Story-LDM model with the proposed memory-attention module. *(b)* A snapshot of the memory for k frames.

effectively learn the motion, various approaches disentangle the motion features from the background information [15,48,51]. Wu *et al*. [54] propose a novel two-dimensional VQ-VAE and sparse attention module for real-world text-to-video generation. Singer *et al*. [45] decompose the temporal U-Net [43] and the attention modules to approximate them in space and time to extend the text-to-image diffusion models to text-based video modeling.

Story Generation. Li *et al*. [27] proposed the initial idea and task of story generation. A two-level StoryGAN framework is applied to ensure image-level consistency between each sentence and image pair, and a global discriminator enforces global consistency between the entire image sequence and the story. Various approaches have proposed improvements to the StoryGAN architecture. Zeng *et al*. [56] introduce sentence-level alignment and word-based attention to improve relevance. Li *et al*. [26] further improve the performance with enhanced discriminators and dilated convolutions. In [46] foreground-background information is provided as additional supervision and [31] use video captioning for semantic alignment between text and frames.

Story Completion. Recently, another task for text-to-story synthesis referred to as story completion has been proposed [33]. In this task, in addition to sentences, the first frame of the story is provided as input. In effect, story completion is a simplified variant of story generation. StoryDALL-E [33] leverages models pre-trained for text-to-image synthesis to perform story completion. Datasets for this task include CLEVR-SV [27] and Pororo-SV [14] which are derived from the CLEVR dataset [19], and the Flintstones dataset for text-to-video synthesis has also been modified for the task of story visualization [31]. Additionally, to evaluate the generalization performance, popular DiDeMo dataset for video captioning [2] is adapted for the task in [33].

Reference Resolution. Co-reference resolution is an important and well-researched topic in NLP and focuses on resolving the pronouns and their associated entities. Classic methods in NLP to co-reference resolution employ decision trees [3,35], maximum-entropy modeling [23], cluster-ranking [38] and classification algorithms [47]. More recent approaches [11,21,25] leverage neural network architectures to obtain improved performance with Transformers [7]. Seo *et al*. [44] proposed visual co-reference resolution for the task of Visual Question-Answering (VQA) dialogs. We take inspiration from [44], but propose a much more sophisticated memory-attention module that allows us to perform visio-lingual co-reference resolution (and visual consistency modeling) for visual story generation.

3. Approach

To generate temporally consistent stories based solely on the linguistic story-line, we develop a deep generative approach with autoregressive structure. We build upon the success of diffusion models in modeling the underlying data distribution of images to produce high quality samples, and learn the generative conditional distribution of the visual story based on the textual descriptions. Given that the multi-frame stories involve high-dimensional data input, we employ Latent Diffusion Models [42], such that diffusion models can be applied in a computationally efficient manner. Besides, to ensure temporal consistency and smooth story progression, we propose a novel memory attention mechanism which not only attends to the multimodal representations of the current frame but also takes into account the already generated semantics of the previous frames. This module also allows us to resolve ambiguous references (*e.g*., he/she, they, *etc*.) using visual memory and is the core of our technical contribution. We first provide an overview of the Diffusion Models and the Latent Diffusion Models, following which we present our autoregressive latent diffusion model for stories called *Story-LDM*[1].

3.1. The Latent Diffusion Model Backbone

Diffusion Models. Diffusion models are a class of generative models that approximate the underlying data distri-

[1]We will make the code and datasets available upon publication.

bution $p(\mathbf{x})$ by denoising a base (Gaussian) distribution in multiple steps using a reverse process of a fixed Markov Chain of length T. To estimate $p(\mathbf{x})$, the forward diffusion process starts from the input data $\mathbf{x}_0 = \mathbf{x}$ and gradually adds noise to obtain a set of noisy samples $\mathbf{x}_1, \ldots, \mathbf{x}_T$ such that $\mathbf{x}_T \sim \mathcal{N}(0, 1)$ represents a sample from a Gaussian distribution. Under the Markov assumption, the probability of the forward process modeling the distribution $q(\mathbf{x}_{0:T} \mid \mathbf{x}_0)$ and the reverse diffusion process estimating probability at an earlier time-step are formulated as:

$$\begin{aligned} q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) &:= \prod_{i=1}^{T} \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \boldsymbol{I}) \\ p_\theta(\mathbf{x}_{0:T}) &= p_\theta(\mathbf{x}_T) \prod_{i=1}^{T} p(\mathbf{x}_{t-1} \mid \mathbf{x}_t). \end{aligned} \tag{1}$$

Here, $\{\beta_i\}_{i=1}^{T}$ is the variance schedule for each time-step such that $\mathbf{x}_T$ is nearly a Gaussian. The model parameters θ are learnt with the following objective,

$$\mathcal{L}_{DM} := \mathbb{E}_{t,\mathbf{x},\epsilon}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|_2^2\right], \tag{2}$$

where $\epsilon \sim \mathcal{N}(0, 1)$ and $\epsilon_\theta(\mathbf{x}_t, t)$, $t = 1, \ldots, T$ is a sequence of denoising autoencoders with noisy input $\mathbf{x}_t$ predicting the noise that was added to the original input $\mathbf{x}$.

Despite yielding state-of-the-art results in various image generation tasks, diffusion models directly operating in the high-dimensional pixel are computationally expensive and resource exhaustive. This limits their application to an even higher-dimensional data such as multi-frame stories or video datasets, which is the focus of this work.

Diffusion Models in the Latent Space. To broaden the applicability of the diffusion models to very high-dimensional data *e.g.* high-resolution images, Latent Diffusion Models (LDM) [42] first compress the original image to a lower-dimensional space using perceptual image compression. An auto-encoder approach is employed such that the original spatial structure of the input image is preserved in the latent space. That is, the encoder $E(\cdot)$ maps the input image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ to a latent representation $\mathbf{Z} \in \mathbb{R}^{h \times w \times c}$, downsampling the image to a lower spatial dimension. Following this, the diffusion model is applied to the latents $\mathbf{Z}$, where time-conditioned U-Net $\epsilon_\theta(\mathbf{Z}_t, t)$ is employed to model the diffusion process. The objective of the diffusion model from Eq. (2) becomes,

$$\mathcal{L}_{LDM} := \mathbb{E}_{t,E(\mathbf{Z}),\epsilon}\left[\|\epsilon - \epsilon_\theta(\mathbf{Z}_t, t)\|_2^2\right], \tag{3}$$

During training, a forward diffusion process is applied to generate $\mathbf{Z}$, which are mapped to the original image space using a decoder $D(\cdot)$.

Dataset	# Ref (avg.)	# Chars	# Backgrounds
MUGEN [16]	None	1	2
Extended MUGEN	3	3	6
FlintstonesSV [14]	3.58	7	323
Extended FlintstonesSV	4.61	7	323
PororoSV [27]	1.01	9	None
Extended PororoSV	1.16	9	None

Table 1. **Dataset statistics of the MUGEN, FlintstonesSV and PororoSV**.

3.2. Story-Latent Diffusion Models

Given a textural story, characterized by sequence of M sentences $\mathbf{S}_{txt} = \{\mathbf{S}^0, \ldots, \mathbf{S}^M\}$, the goal of story generation is to produce a sequence of corresponding frames $\mathbf{S}_{img} = \{\mathbf{I}^0, \ldots, \mathbf{I}^M\}$ that visualize the story. We note that this is a more difficult problem than one of story continuation [33], where in addition to the textual story $\mathbf{S}_{txt}$ approaches have access to a source frame $\mathbf{I}^0$ for additional context at inference time. During training it is assumed that we have access to paired dataset of N samples $\mathcal{D} = \{\mathbf{S}_{txt}^{(i)}, \mathbf{S}_{img}^{(i)}\}_{i=1}^{N}$. We extend latent diffusion models to this task, by allowing them to generate multi-frame stories autoregressively, and by introducing rich conditional structure that takes into account current sentence as well as context from earlier generated frames through visual memory module. This visual memory allows the model to incorporate character/background consistency and resolve text references when needed, resulting in improved performance.

Given a condition $\mathbf{y}$, LDM utilizes a cross-attention layer with key $(\mathbf{K})$, query $(\mathbf{Q})$ and value $(\mathbf{V})$ where,

$$\text{Attention}(\mathbf{K}, \mathbf{Q}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right).\mathbf{V}. \tag{4}$$

Here, $\mathbf{Q} = \mathbf{W}_Q.\hat{f}(\mathbf{Z})$, $\mathbf{K} = \mathbf{W}_K.f(\mathbf{y})$ and $\mathbf{V} = \mathbf{W}_V.f(\mathbf{y})$, and $\mathbf{W}_Q \in \mathbb{R}^{d \times d_q}$, $\mathbf{W}_K \in \mathbb{R}^{d \times d_k}$ and $\mathbf{W}_V \in \mathbb{R}^{d \times d_v}$ are learnable parameters, $\hat{f}(\mathbf{Z})$ an intermediate flattened feature representation of $\mathbf{Z}$ within the diffusion model and $f(\mathbf{y})$ the feature representation of the condition $\mathbf{y}$. The objective in Eq. (3) for conditional generation becomes,

$$\mathcal{L}_{LDM} := \mathbb{E}_{t,E(\mathbf{I}),\epsilon}\left[\|\epsilon - \epsilon_\theta(\mathbf{Z}_t, f(\mathbf{y}), t)\|_2^2\right]. \tag{5}$$

Note that the denoising autoencoders ϵ_θ now additionally depend on the condition encoding $f(\mathbf{y})$.

For a sample $\{\mathbf{S}_{txt}^{(i)}, \mathbf{S}_{img}^{(i)}\}$, we first project all the M input frames of the story, $\mathbf{I}^0, \ldots, \mathbf{I}^M$ onto a low-dimensional space using a frame-encoder E, and obtain the encoded frames $\mathbf{Z}^0, \ldots, \mathbf{Z}^M$ for a single story[2]. To effectively condition on the corresponding frame, we apply this cross-attention layer to the intermediate representations of the neural network for each frame $\mathbf{I}^m$ and its textual description

[2]We drop the superscript denoting sample i for ease of notation.

Dataset	Method	w/ ref. text	Char-acc (↑)	Char-F1 (↑)	BG-acc (↑)	BG-F1 (↑)	FID (↓)
Flintstones	VLCStoryGAN [31]	×	27.73	42.01	4.83	16.49	120.85
	LDM [42]	×	79.86	92.33	48.02	37.86	61.40
	LDM [42]	✓	57.38	78.68	44.19	28.25	87.39
	Story-LDM (Ours)	✓	69.19	86.59	35.21	28.80	69.49
PororoSV	DUCO-STORYGAN [32]	✓	13.97	38.01	-	-	96.51
	VLCStoryGAN [31]	✓	17.36	43.02	-	-	84.96
	LDM [42]	✓	16.59	56.30	-	-	60.23
	Story-LDM (Ours)	✓	**20.26**	**57.95**	-	-	**36.64**
MUGEN	LDM [42]	✓	31.39	21.28	15.74	18.66	120.99
	Story-LDM (Ours)	✓	**93.40**	**95.60**	**92.19**	**92.37**	**62.16**

Table 2. **Quantitative results.** Experimental results on the FlintstoneSV, PororoSV and the MUGEN datasets.

$\mathbf{S}^m$ using Eq. (4), where the relevance of the description $\mathbf{S}^m$ is weighted by the similarity between the textual representation and the encoded frame representation $\mathbf{Z}^m$ (*cf*. Fig. 2a).

For sequential generation, the model in addition to the current state, requires information from all the previous states. To enable this, the diffusion process for any frame representation $\mathbf{Z}^m$ is conditioned on the visual representations of the previous frames $\mathbf{Z}^0, \ldots, \mathbf{Z}^{m-1}$ as well as the sentence descriptions $\mathbf{S}^0, \ldots, \mathbf{S}^m$. This conditioning is realized through a novel *Memory-attention* module which forms the basis of our autoregressive approach.

Memory-attention Module. To capture the spatio-temporal interactions across multiple frames and sentences for a story, in our conditional diffusion model, we condition the frame $\mathbf{Z}^m$ not only on the corresponding text $\mathbf{S}^m$ but also on the previous texts $\mathbf{S}^i$, for $i \in \{0, \ldots m-1\}$. This conditioning is applied throughout the T time-steps of the diffusion process for $\mathbf{Z}^m$. The conditional denoising autoencoder thus models the conditional distribution $p(\mathbf{Z}^m \mid \mathbf{Z}^{<m}, \mathbf{S}^{\leq m})$. The (conditional) generative process of our *Story-LDM* approach over the T steps of the diffusion process for a single frame is thus given by

$$p(\mathbf{Z}^m|\mathbf{Z}^{<m}, \mathbf{S}^{\leq m}) = p(\mathbf{Z}^m_T|\mathbf{Z}^{<m}, \mathbf{S}^{\leq m}) \prod_{i=1}^{T} p(\mathbf{Z}^m_{i-1}|\mathbf{Z}^m_i, \mathbf{Z}^{<m}, \mathbf{S}^{\leq m}). \tag{6}$$

The key motivation for this approach is to propagate the semantic (visual and textual) features from the already processed story-line based on the relevance of the current description to the previous frames as well as the previous descriptions. We achieve this by implementing a special attention layer, called *memory-attention module*. Similar to the cross-attention layer, we utilize the attention mechanism based on the key, query and value formulation. In this case Eq. (4) becomes,

$$\begin{aligned} \mathbf{Q} &= \mathbf{W}_Q.f(\mathbf{S}^m), \mathbf{K} = \mathbf{W}_K.f(\mathbf{S}^{<m}), \\ \mathbf{V} &= \mathbf{W}_V.\hat{f}(\mathbf{Z}^{<m}), \end{aligned} \tag{7}$$

where $\hat{f}(.)$ is applied to align the dimensions of the values, $\mathbf{V}$ with the keys, $\mathbf{K}$. In the memory attention module, the relevance the query $\mathbf{Q}$ which depends on the current sentence $\mathbf{S}^m$ and the keys $\mathbf{K}$ which represent the previous sentences $\mathbf{S}^{<m}$ is used to weight the feature representations $\mathbf{Z}^{<m}$. The aggregated representation now contains the information relevant for the current frame $\mathbf{Z}^m$ from the already generated story-line (see Fig. 2b). That is, our mechanism based on the similarity of the current sentence to the previous sentences in the story, identifies the features in the previous frames which are of importance to the context of the current frame. This may include recurrence of certain semantics with-in the story such as characters or backgrounds. In all, this formulation of the diffusion process allows us to maintain temporal consistency as we amplify the visual feature information from the sequence of story already generated. This allows the model to implicitly capture temporal dependencies in storylines for resolving ambiguities in character and background information.

Given the above conditioning, the objective for the story latent diffusion model for a single frame is formalized as

$$\mathcal{L}_{story-LDM} = \mathbb{E}_{\mathbf{Z},\epsilon,t}\left[\|\epsilon - \epsilon_{\theta^m}(\mathbf{Z}^m_t, \mathbf{S}^{\leq m}, \mathbf{Z}^{<m}, t)\|^2_2\right], \tag{8}$$

where ϵ_{θ^m} are the denoising autoencoders for the frame m.

Having formalized the diffusion process for single frame generation, the generative process for the entire story-line using Eq. (6) for autoregressive conditional frame generation, is given by

$$p(\mathbf{Z}^{0:M} \mid \mathbf{S}^{0:M}) = p(\mathbf{Z}^0 \mid \mathbf{S}^0) \prod_{i=m}^{M} p(\mathbf{Z}^m \mid \mathbf{Z}^{<m}, \mathbf{S}^{\leq i}). \tag{9}$$

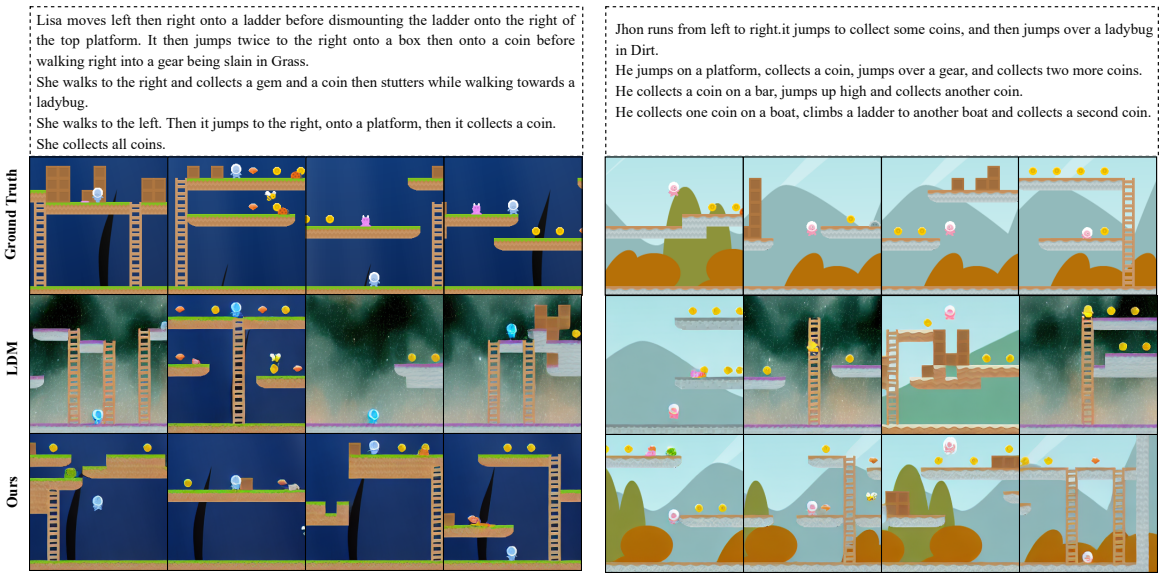


Figure 3. **Story generation result for MUGEN.** Here we can compare between our method and LDM [42]. See text for details.

Figure 4. **Story generation results for FlintstoneSV.** Our method is able to generate more consistent characters/backgrounds.

Notably, the conditioning is applied to the all states within the diffusion process *i.e.*, for all $\mathbf{Z}_t^m$, $t \in \{1, \dots T\}$ at each diffusion step, we apply the cross attention as well as the memory attention module allowing us effectively capture the temporal context.

Network Architecture. To generate visual storylines, in our *Story-LDM*, we first introduce an autoregressive structure and modify the two-dimensional U-Net in [42] to so as to process the temporal information in the storyline. As shown in Fig. 2a, the frame encoder, E equipped with the positional information of the frame in the sequence, is applied to get the low-dimensional representation $\mathbf{Z}^m$ for all the frames in a datapoint from $\mathcal{D}$. A text-based transformer is applied to get a suitable representation for the sentence $\mathbf{S}^m$. The U-Net is then applied to model the diffusion process over T time-steps. The layers within the U-Net are augmented with the cross-attention layer and our memory attention layer. After each downsampling or upsampling operation we apply the attention mechanism to reinforce the conditioning on the already encoded (learned) story-line up to the previous time-step. For any frame m, the cross-attention C_{attn} is given by

$$C_{attn} = \sum_i \hat{f}(\mathbf{Z}^m)_i f(\mathbf{S}^m)_i \tag{10}$$

where $\hat{f}(\mathbf{Z}^m)$ and $f(\mathbf{S}^m)$ are the representations of the frame encoding $\mathbf{Z}^m$ within the neural network and sentence $\mathbf{S}^m$ respectively such that they have same dimensions. Similarly, the memory attention M_{attn} is computed as,

$$M_{attn} = \sum_{k=1}^{m-1} \sum_{i} \hat{f}(\mathbf{Z}^k)_i f(\mathbf{S}^k)_i f(\mathbf{S}^m)_i \qquad (11)$$

The output of the attention-module is then computed as the aggregation, $C_{attn} + M_{attn}$.

Starting from the noise sample $\mathbf{Z}_T^m$ the output of the reverse diffusion process $\mathbf{Z}_0^m$ is reconstructed using the frame decoder D to get the final image.

Having outlined the details of our Story-LDM framework, we show through extensive experiments on the task on story generation, the effectiveness and the benefits of our powerful conditioning based on memory-attention.

4. Datasets and Evaluation Metrics

In this paper, we formulate story generation with co-references to actors and backgrounds across frames.

Datasets. Since reference resolution has not been studied in story generation, to validate our approach on this much harder task, we construct the following datasets: *(i)* We take an existing story-generation dataset – FlintstonesSV [14], and modify the sentences by replacing the named entities (characters) with references where possible; including pronouns such as *he*, *she*, or *they* (*cf*. Fig. 1). This dataset contains 20132-training, 2071-validation and 2309-test stories with 7 main characters and 323 backgrounds. *(ii)* MUGEN [16] is a video dataset collected from the open–sourced platform game CoinRun [5]. The dataset is divided into $104,796$-train and $11,802$ test stories with 96 to 602 frames. We extend the MUGEN dataset by introducing two additional characters *Lisa* and *Jhon* (we rename *Mugen* to *Tony*). We construct stories of four frames and corresponding text, ensuring consistent co-referencing in the story; each story has 3 such references. Moreover, we augment the existing two backgrounds (*Planet* and *Snow*) with four additional backgrounds: *Sand*, *Dirt*, *Grass* and *Stone*. *(iii)* We also modify existing PororoSV [27] dataset which contains 10191/2334/2208 train/val/test set. Similarly, we reference characters by pro-nouns to generate more natural story. We show in Fig. 1, example stories from the two modified datasets and enlist the complete statistics in Tab. 1.

Evaluation Metrics. To measure the consistency of the characters as well as the backgrounds in the generated stories, we consider following evaluation metrics: *(i) Character Classification*: Following [31], we consider fine-tuned Inception-v3 to measure the classification accuracy and F1-score. Frame accuracy evaluates the character match to the ground-truth and F1-score measures the quality of generated characters in the predicted images. *(ii) Background*

Figure 5. **Qualitative Comparison on Story Generation.** Comparison on the FlintstonesSV dataset visual story generation.

Classification: Similar to character classification, we use fine-tuned Inception-v3 to measure the correspondence of the background to the ground-truth and consider F1-score as a measure of quality. *(iii) Frechet Inception Distance (FID)*: To assess the quality of images,we consider FID score [17] which is the distance between feature vectors from real and generated images.

Figure 6. **Comparison to Story Continuation.** Comparison on the FlintstonesSV dataset for story continuation. Prior work uses first frame as additional input to the model; our model does not.

5. Experiments

In this section, we evaluate our Story-LDM approach for consistent story generation with reference resolution.

Baselines. We construct a strong baseline with the LDM[3] [42] which contains a cross-attention layer to generate text-to-image based story, without using our proposed autoregressive memory modules as our baselines for MUGEN, PororoSV and FlintstonesSV datasets. The parameters of the diffusion model within the Story-LDM are initialized with the pre-trained LDM [42]. Similarly, for the textual embedding, we use BERT-tokenizer [7] and use the pre-trained text-transformer from LDM.

Quantitative Results. Table 2 shows quantitative results for consistent story generation on the FlintstoneSV dataset. We compare the performance of our approach (row 4) to the LDM [42] which we train/test with both original (row 2) as well as the co-referenced (row 3) descriptions. Furthermore, we include the results of the state-of-art VLCStoryGAN [31] (row 1) with the original text of the dataset[4] (*i.e.* without co-references). We note that VLCStoryGAN was shown to be better than Duco-StoryGAN [32], CP-CSV [46] and original StoryGAN [27] (see [31]).

Based on Table 2 we make three observations: (1) Our LDM baseline is better than VLCStoryGAN on the original reference-free text (*cf*. Tab. 2, rows 1 & 2). (2) Reference resolution makes the task considerably harder. With the reference text in our modified dataset, we observe a drop in performance in terms of character and background classification scores (*cf*. Tab. 2, rows 2 & 3). (3) Our model, with memory-attention module, significantly outperforms the baseline (*cf*. Tab. 2, rows 3 & 4) both in terms of generative image quality and character consistency; and outperforms SoTA of VLCStoryGAN by $\sim$ 41% percentage points on character accurary (while performing a more difficult version of the task). Further, our model, that is required to conduct reference resolution, comes close to the LDM trained with original, reference-free, text (*cf*. Tab. 2, rows 3 & 4), which can be viewed as a sort of an upper bound.

On the PororoSV dataset, our method outperforms previous baseline models (including LDM baseline) in character evaluation metrics. To be noted, PororoSV [27] dataset has no background information, therefore we only perform character level evaluation on this dataset.

On the MUGEN dataset, our method outperforms the strong LDM baseline with gains of $\sim$ 62% on character accuracy and $\sim$ 76% on the background accuracy, thereby showing the advantages of the memory-attention mechanism for consistent story generation. We note that MUGEN dataset has more references across story scenes. Flintstones while contains more references per story overall, many of those references are within scenes as opposed to across scenes. Meaning that in terms of reference impact on consistency, MUGEN dataset is actually harder.

[3] https://github.com/CompVis/latent-diffusion

[4] Results for [31] were obtained using pretrained model provided by original authors in private communication.

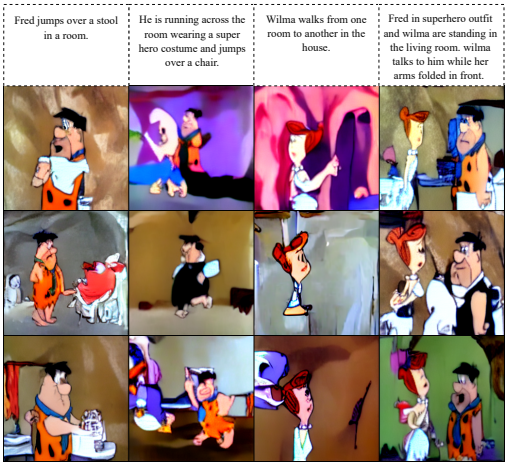


Figure 7. **Story Diversity.** Diverse outputs for a single storyline obtained with our Story-LDM.

Qualitative Results. Figure 3 illustrates qualitative results on the MUGEN dataset. Rows 1, 2 & 3 show ground truth, LDM [42] and our Story-LDM approach, respectively. Here, we see that our method is able to maintain consistency in terms of both character and background. Similarly, in Figure 4 we can show the results on FlintstoneSV dataset which further validates the strong performance of our method when generating high-quality, consistent story. Compared to the LDM, our approach is able to adapt to the diverse backgrounds in the story descriptions.

Additional Results. We compare the qualitative results of our method to both story generation [31] and story continuation [33] in Figs. 5 and 6 respectively. The comparative images are taken directly from respective papers. We note that story continuation Fig. 6 is solving a different (easier) problem and with text that contains no-references. This makes the comparison to our method, which receives fewer inputs, not very meaningful. Nether-the-less, our approach, that can resolve references and is solving a harder story generation task, obtains highly competitive results. Furthermore, to show that our autoregressive visual memory module can generate diverse stories conditioned on the current and previous condition, we create different story-lines starting for a single sentence. In Fig. 8, we can see for reference 'they', the model can generate both the characters according to the storyline already parsed. Moreover, in Fig. 7 we show that our approach can not only generate consistent visual stories, but also diverse frames for the same text (*cf*. Fig. 7). Additional results are provided in the Supplemental.

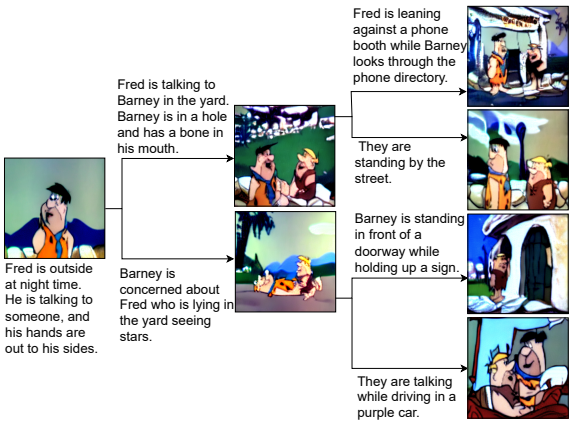


Figure 8. **Branching Storyline.** Generating different yet consistent stories by branching the storyline. Frames in the later columns are generated based on earlier ones and corresponding text.

6. Conclusion

In this paper, we formulate consistent story generation in a more realistic way by co-referencing actors/backgrounds in the story descriptions. We develop an autoregressive Story-LDM approach with memory attention capable of maintaining consistency across the frames based on the previously generated frames and their corresponding descriptions. We introduced modified datasets to evaluate the performance for reference resolution. We expect our proposed formulation and models to be conductive to the real-world use cases and further the research.

References

[1] Jyoti Aneja, Harsh Agrawal, Dhruv Batra, and Alexander G. Schwing. Sequential latent spaces for modeling the intention during diverse image captioning. *ICCV*, 2019. 2

[2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, pages 5803–5812, 2017. 3

[3] Chinatsu Aone and Scott William. Evaluating automated and manual acquisition of anaphora resolution strategies. In *Annual Meeting of the Association for Computational Linguistics*, pages 122–129, 1995. 2, 3

[4] Apratim Bhattacharyya, Shweta Mahajan, Mario Fritz, Bernt Schiele, and Stefan Roth. Normalizing flows with multi-scale autoregressive priors. In *CVPR*, pages 8412–8421, 2020. 2

[5] Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. In *ICML*, pages 1282–1289. PMLR, 2019. 7

[6] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *ICML*, pages 1174–1183. PMLR, 2018. 1, 2

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018. 3, 8

[8] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *NeurIPS*, 34:19822–19835, 2021. 2

[9] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: Non-linear independent components estimation. In *ICLR Workshop*, 2015. 2

[10] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *ICLR*, 2017. 2

[11] Hongliang Fei, Xu Li, Dingcheng Li, and Ping Li. End-to-end deep reinforcement learning based coreference resolution. In *Annual Meeting of the Association for Computational Linguistics*, pages 660–665, 2019. 3

[12] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014. 1, 2

[13] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, pages 10696–10706, 2022. 2

[14] Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. Imagine this! scripts to compositions to videos. In *ECCV*, pages 598–613, 2018. 1, 2, 3, 4, 7

[15] Zekun Hao, Xun Huang, and Serge Belongie. Controllable video generation with sparse trajectories. In *CVPR*, pages 7854–7863, 2018. 3

[16] Thomas Hayes, Songyang Zhang, Xi Yin, Guan Pang, Sasha Sheng, Harry Yang, Songwei Ge, Isabelle Hu, and Devi Parikh. Mugen: A playground for video-audio-text multimodal understanding and generation. *arXiv preprint arXiv:2204.08058*, 2022. 1, 2, 4, 7

[17] 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. *NIPS*, 30, 2017. 7

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 1, 2

[19] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, pages 2901–2910, 2017. 3

[20] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, 2016. 1

[21] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020. 3

[22] Minguk Kang and Jaesik Park. Contragan: Contrastive learning for conditional image generation. *NeurIPS*, 33:21357–21369, 2020. 1, 2

[23] Andrew Kehler. Probabilistic coreference in information extraction. 1997. 2, 3

[24] Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *ICLR*, 2014. 1, 2

[25] Hideo Kobayashi, Yufang Hou, and Vincent Ng. End-to-end neural bridging resolution. In *International Conference on Computational Linguistics*, pages 766–778, 2022. 3

[26] Chunye Li, Liya Kong, and Zhiping Zhou. Improved-storygan for sequential images visualization. *Journal of Visual Communication and Image Representation*, 73:102956, 2020. 3

[27] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization. In *CVPR*, pages 6329–6338, 2019. 1, 2, 3, 4, 7, 8

[28] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. In *AAAI*, volume 32, 2018. 1, 2

[29] Shweta Mahajan, Iryna Gurevych, and Stefan Roth. Latent normalizing flows for many-to-many cross-domain mappings. In *ICLR*, 2020. 2

[30] Shweta Mahajan and Stefan Roth. Diverse image captioning with context-object split latent spaces. *NeurIPS*, 33:3613–3624, 2020. 1, 2

[31] Adyasha Maharana and Mohit Bansal. Integrating visuospatial, linguistic and commonsense structure into story visualization. *arXiv preprint arXiv:2110.10834*, 2021. 1, 3, 5, 7, 8

[32] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. Improving generation and evaluation of visual stories via semantic consistency. 2021. 1, 2, 5, 8

[33] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. In *ECCV*, pages 70–87. Springer, 2022. 1, 2, 3, 4, 8

[34] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. Deep captioning with multimodal recurrent neural networks (m-RNN). In *ICLR*, 2015. 1

[35] Joseph F McCarthy and Wendy G Lehnert. Using decision trees for coreference resolution. pages 1050–1055, 1995. 2, 3

[36] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *ICML*, 162:16784–16804, 2022. 2

[37] 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 *ICML*, volume 139, pages 8748–8763. PMLR, 2021. 2

[38] Altaf Rahman and Vincent Ng. Narrowing the modeling gap: A cluster-ranking approach to coreference resolution. *Journal of Artificial Intelligence Research*, 40:469–521, 2011. 3

[39] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pages 8821–8831. PMLR, 2021. 2

[40] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *NeurIPS*, 32, 2019. 2

[41] Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, 2016. 1, 2

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 3, 4, 5, 6, 8

[43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015. 3

[44] Paul Hongsuck Seo, Andreas Lehrmann, Bohyung Han, and Leonid Sigal. Visual reference resolution using attention memory for visual dialog. *NIPS*, 30, 2017. 2, 3

[45] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 3

[46] Yun-Zhu Song, Zhi Rui Tam, Hung-Jen Chen, Huiao-Han Lu, and Hong-Han Shuai. Character-preserving coherent story visualization. In *ECCV*, pages 18–33. Springer, 2020. 3, 8

[47] Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544, 2001. 3

[48] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *CVPR*, pages 1526–1535, 2018. 3

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 2

[50] Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. Diverse beam search for improved description of complex scenes. In *AAAI*, 2018. 1

[51] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *NIPS*, 29, 2016. 3

[52] Liwei Wang, Alexander G. Schwing, and Svetlana Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive Gaussian encoding space. In *NIPS*, 2017. 2

[53] Wei Wang, Xavier Alameda-Pineda, Dan Xu, Pascal Fua, Elisa Ricci, and Nicu Sebe. Every smile is unique: Landmark-guided diverse smile generation. In *CVPR*, pages 7083–7092, 2018. 1

[54] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021. 3

[55] Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao. Semantics disentangling for text-to-image generation. *CVPR*, 2019. 1, 2

[56] Gangyan Zeng, Zhaohui Li, and Yuan Zhang. Pororogan: an improved story visualization model on pororo-sv dataset. In *Proceedings of the 3rd International Conference on Computer Science and Artificial Intelligence*, pages 155–159, 2019. 1, 3

[57] Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *ICML*, 2019. 1, 2

[58] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *CVPR*, pages 833–842, 2021. 1, 2

[59] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *CVPR*, pages 5802–5810, 2019. 1, 2