Make-A-Story

https://github.com/xichenpan/ARLDM

Datasets

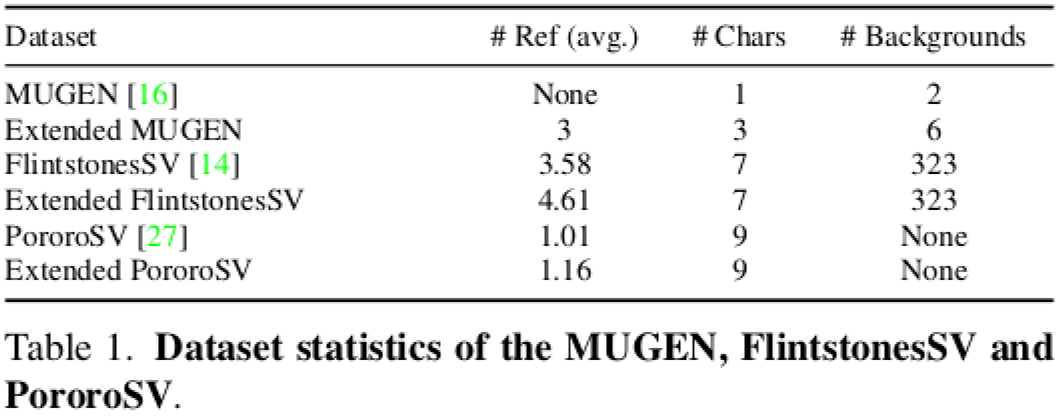
FlintstonesSV-

→ replaces named entities wherever possible

→ including pronouns

PororoSV

MUGEN



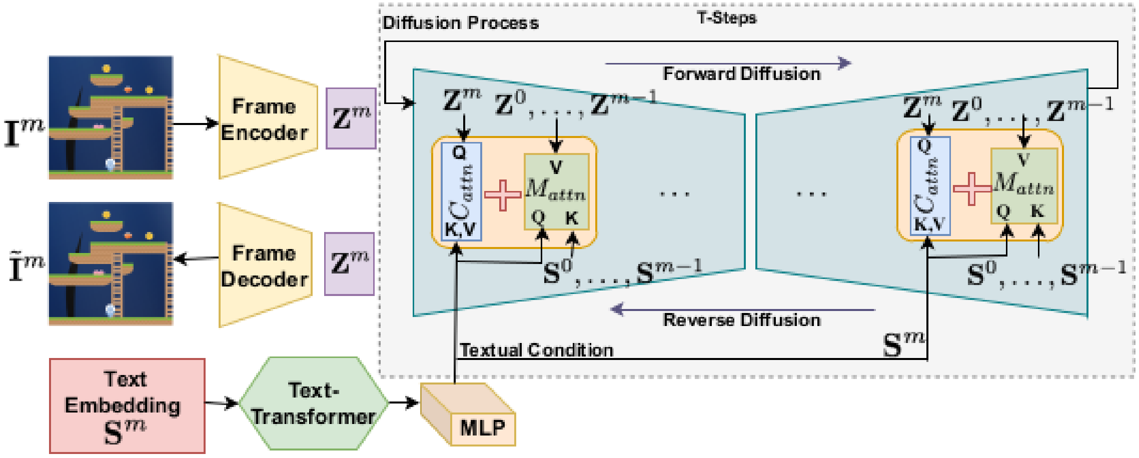
Dataset	# Ref (avg.)	# Chars	# Backgrounds
MUGEN [16]	None	1	2
Extended MUGEN	3	3	6
FlintstonesSV [14]	3.58	7	323
Extended FlintstonesSV	4.61	7	323
PororoSV [27]	1.01	9	None
Extended PororoSV	1.16	9	None

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

Model

(read after method)

Story-LDM



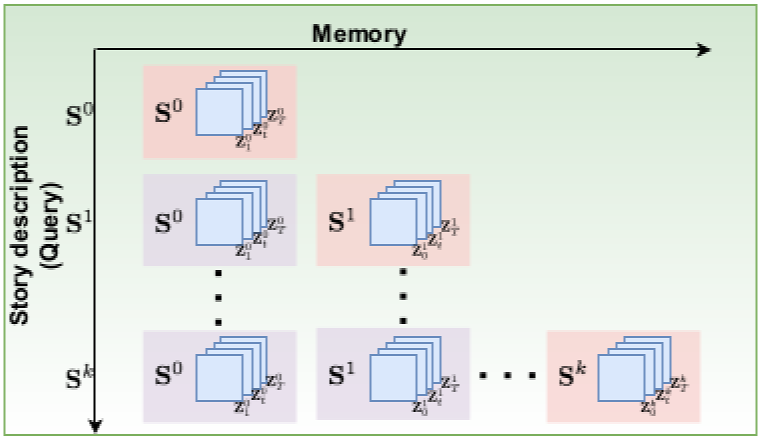
(a) Story-LDM architecture for conditional story generation.

Story-LDM

LDM extended to support story

Meticulously designed memory-attention mechanism capable of encoding and leveraging contextual reference between the part of the story-line being generated, and the current frame being generated based on learned semantic similarity of corresponding sentences. This sequential diffusion model can thus generate consistent stories by resolving and then capturing temporal character and background context.

Memory Attention Module-



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

Network Architecture

Introduce an autoregressive structure and modify the two-dimensional U-Net present in https://arxiv.org/abs/2112.10752 to process the temporal information in the storyline. The frame encoder, E equipped with the positional information of the frame in the sequence, is applied to get the low dimensional representation

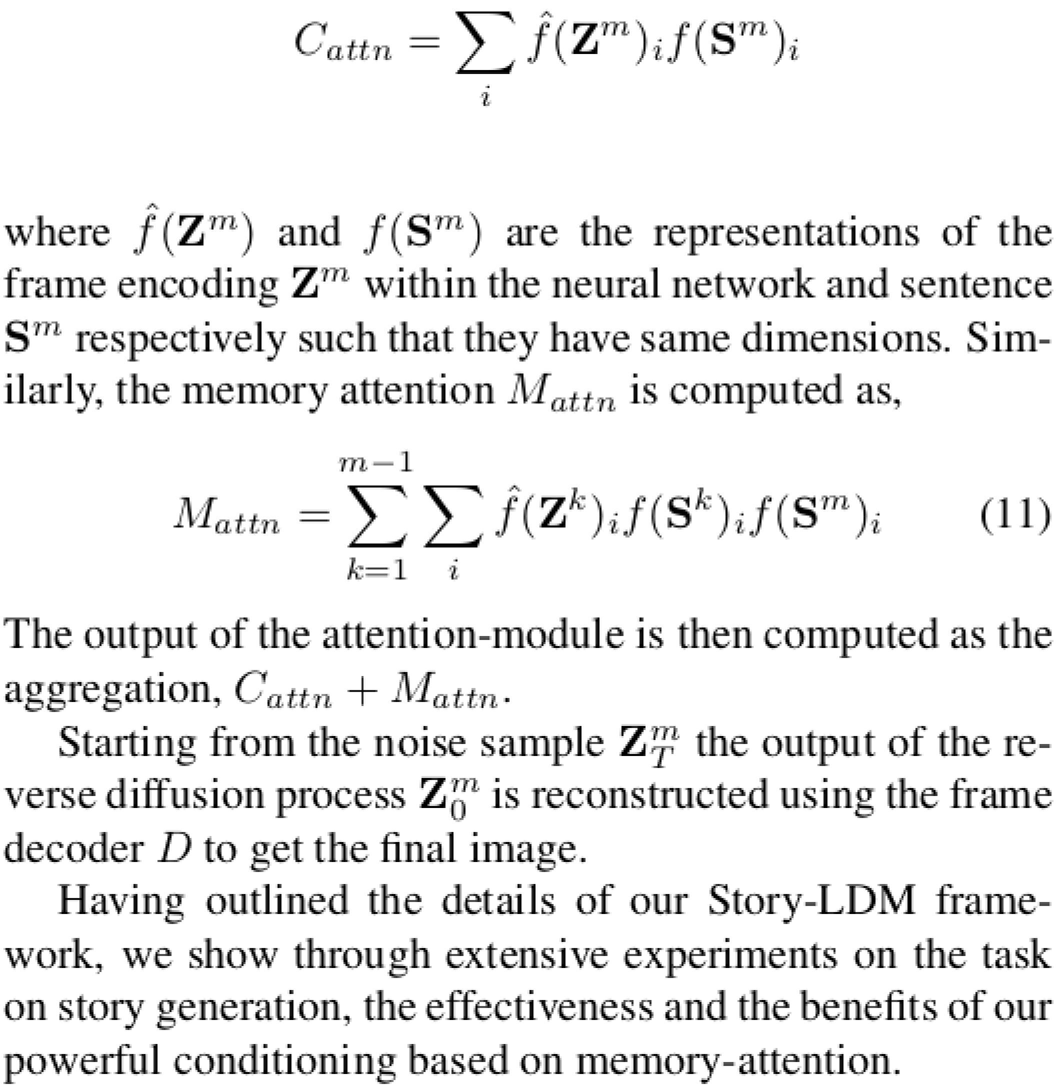
$$Z^m$$

for all the frames in a datapoint D. A text-based transformer is applied to get a suitable representation for the sentence

$$S^m$$

The U-net is 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 conditional on the already encoded(learned) story-line up the to the previous time-step. For any frame m, the cross-attention

$$C_attn$$



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

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.

Method

→ Employs Latent Diffusion model given that multi-frame stories involve high-dimensional Data

→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**

→**Diffusion Models in Latent Space**- To make it usable in real life and actually viable computation, first make low res and upscale.

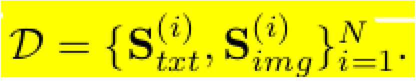
→**Story Generation**-

STORY LDM

Given a textural story, characterized by sequence of M sentences

Stxt = {S0 , . . . , SM }, the goal of story generation is to produce a sequence of corresponding frames Simg = {I0 , . . . , IM } that visualize the story.

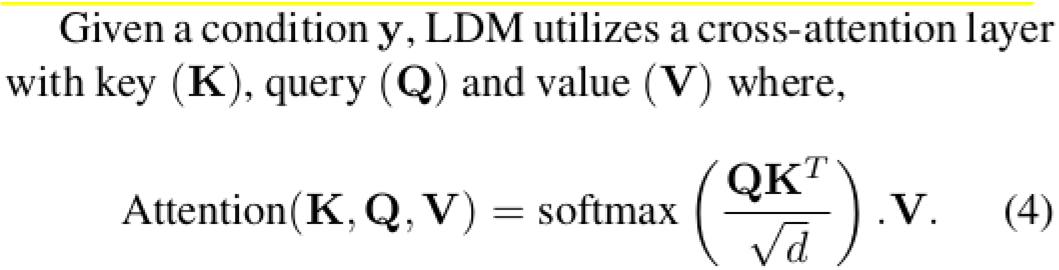
During training it is as-
sumed that we have access to paired dataset of N samples



$$\mathcal{D} = \{\mathbf{S}_{txt}^{(i)}, \mathbf{S}_{img}^{(i)}\}_{i=1}^{N}.$$

We extend latent diffusion models to this task, by allowing them to generate multi-frame stories autoregressively, and by introducing rich conditional struc-
ture 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.

We have an attention layer-



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

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

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,

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

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 $\mathbf{S}^m$ 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 Zm to account for all previous texts.

Applied through all T time steps

Conditional Denoising Autoencoder models the conditional Distribution-

$$p(\mathbf{Z}^m \mid \mathbf{Z}^{<m}, \mathbf{S}^{\leq m}).$$

Conditional Generative Process of STORY-LDM, over T steps, process for generating a single frame is given by-

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

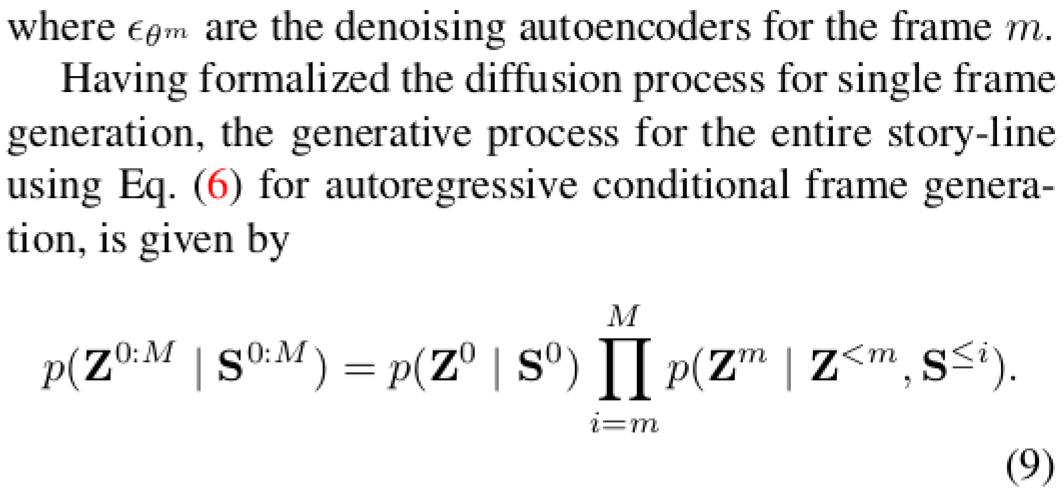
The main motivation for this approach is to propagate the semantic (visual and textual) features from the already processed storyline based on the relevance of the current description to the previous frames as well as the previous descriptions. We achieve this by implementing a special attention layer called a 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, Attention Layer Eq becomes-

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

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

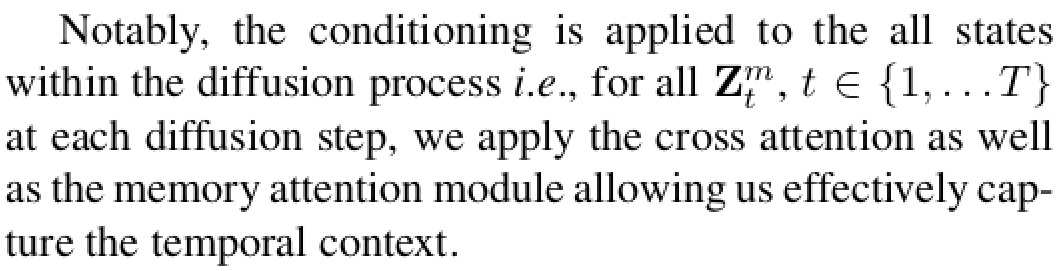
$$\mathcal{L}_{story-LDM} = \mathbb{E}_{\mathbf{Z},\epsilon,t}\left[\|\epsilon - \epsilon_{\theta^m}(\mathbf{Z}_t^m, \mathbf{S}^{\leq m}, \mathbf{Z}^{<m}, t)\|_2^2\right], \tag{8}$$



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

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



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

Evaluation

i) Character Classification - F1 Score comparison w/ ground truth

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 Classification - Compare the correspondence of background with ground-truth (F1 Score)

iii) Frechet Inception Distance(FID), To assess the quality of images, we consider FID score which is the distance between feature vectors from real and generated images.

Comparison w/ other models-

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.