ContextualStory: Consistent Visual Storytelling with Spatially-Enhanced and Storyline Context

Sixiao Zheng[1,2], **Yanwei Fu**[1,2]
[1]Fudan University
[2]Shanghai Innovation Institute
sxzheng18@fudan.edu.cn, yanweifu@fudan.edu.cn

Abstract

Visual storytelling involves generating a sequence of coherent frames from a textual storyline while maintaining consistency in characters and scenes. Existing autoregressive methods, which rely on previous frame-sentence pairs, struggle with high memory usage, slow generation speeds, and limited context integration. To address these issues, we propose ContextualStory, a novel framework designed to generate coherent story frames and extend frames for visual storytelling. ContextualStory utilizes Spatially-Enhanced Temporal Attention to capture spatial and temporal dependencies, handling significant character movements effectively. Additionally, we introduce a Storyline Contextualizer to enrich context in storyline embedding, and a StoryFlow Adapter to measure scene changes between frames for guiding the model. Extensive experiments on PororoSV and FlintstonesSV datasets demonstrate that ContextualStory significantly outperforms existing SOTA methods in both story visualization and continuation.

Introduction

Recent text-to-image (T2I) models, such as SD3 (Esser et al. 2024), excel at generating images from text but only produce individual images independently. Although text-to-video (T2V) models like SVD (Blattmann et al. 2023) and Sora (Brooks et al. 2024) generate coherent videos but often feature simple scene or motion changes. In contrast, this paper focuses on *visual storytelling*, which comprises generating a sequence of coherent story frames from a textual storyline in *story visualization* and extending an initial frame from a textual storyline in *story continuation*. This task has significant potential for educational applications, such as crafting vivid, coherent comics for storybooks. The key challenge is aligning generated frames with sentences while ensuring temporal consistency in characters and scenes. Providing sufficient context is essential due to the limited information in individual sentences.

Many diffusion-based visual storytelling methods use an autoregressive generative approach to capture temporal dependencies based on previous frame-sentence pairs, such as AR-LDM (Pan et al. 2024) and Story-LDM (Rahman et al. 2023). However, these methods face four key limitations:

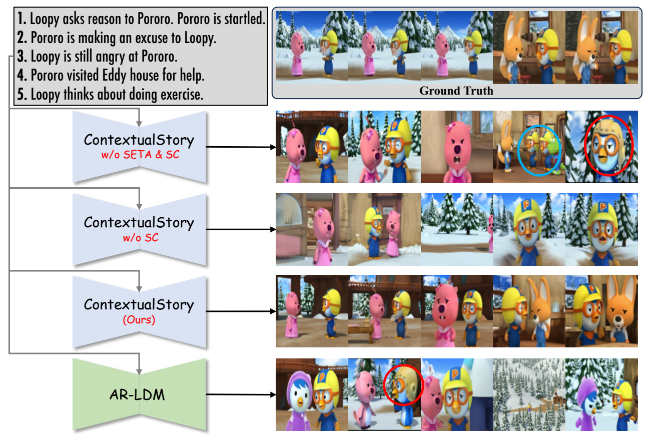

Figure 1: Story frames generated by our ContextualStory on PororoSV dataset. Red circles highlight character inconsistencies, and blue circles indicate repeated characters. SETA and SC enhance character consistency and scene coherence, achieving superior results compared to AR-LDM.

1) High memory usage due to storing all previous frame-sentence pairs, making longer storyline difficult to handle; 2) Limited context in early frame generation, which may impact frame quality; 3) Slow generation speed due to the sequential nature of the process; 4) Inconsistent frames arise from relying solely on past pairs and neglecting future context, missing the global story context. We address this by exploring how the model can 1) *access sufficient frame context* and 2) *obtain adequate context from the storyline.*

To access sufficient frame context, we integrate temporal convolutions and Spatially-Enhanced Temporal Attention (SETA) into the UNet, combining them with the spatial modeling layer. By alternating between spatial and temporal modeling, the model effectively captures spatial dependencies within individual frames and temporal dependencies across frames for comprehensive context. To obtain adequate context from the storyline, we propose the Storyline Contextualizer (SC), which processes the CLIP text embeddings to propagate the context information across sentences, providing sufficient context throughout.

Temporal attention is crucial for visual storytelling as it propagates context across frames along the temporal dimension. However, vanilla temporal attention struggles with significant *character movement* between frames, as shown in Figure 3(a). To overcome this, we propose SETA that employs a local window mechanism to allow queries to attend to features within local windows of other frames. This effectively captures moving characters and enhances spatial dependency modeling. As shown in Figure 1, SETA improves character consistency and reduces repeated characters, such as Pororo, compared to ContextualStory w/o SETA & SC.

The Storyline Contextualizer enhances the contextual information of storyline embeddings from the CLIP text encoder, which initially contain only sentence-level semantics. By integrating and propagating context across sentences, it generates context-enriched storyline embeddings. The Storyline Contextualizer, a transformer-based network, incorporates self-attention and temporal attention layers to capture both global and temporal dependencies. These enriched embeddings guide the model through a temporally-aligned cross-attention mechanism, ensuring consistent story frames. As shown in Figure 1, ContextualStory improves scene consistency and coherence over ContextualStory w/o SC. Furthermore, compared to AR-LDM, our ContextualStory significantly enhances both characters and scenes consistency.

Additionally, to leverage scene changes between story frames to guide the model, we proposed the StoryFlow Adapter to measure change between story frames. For story continuation, we simply add a convolution layer at the input end of the UNet block to match the size of the first frame latent with the noise latent and then concatenate them.

Our contributions are as follows: (1) **ContextualStory Framework**: Our novel framework overcomes limitations of existing autoregressive methods, including high memory usage, limited context, slow generation speed, and image inconsistency. (2) **Spatially-Enhanced Temporal Attention (SETA)**: We present the SETA into the UNet model, combining temporal convolutions with spatial modeling to capture both spatial and temporal dependencies, addressing challenges of significant character movement and improving frame consistency. (3) **Storyline Contextualizer (SC)**: A transformer-based network enriches CLIP text embeddings by capturing global and temporal dependencies, ensuring consistent story frames. (4) **StoryFlow Adapter**: We repurpose this tool to measure scene changes between frames, guiding the model to handle scene transitions more effectively. (5) Extensive experiments on PororoSV and FlintstonesSV datasets demonstrate that our ContextualStory significantly outperforms previous SOTA in visual storytelling.

Related Works

Visual storytelling. Early methods for story visualization primarily relied on GANs (Goodfellow et al. 2020). StoryGAN (Li et al. 2019) pioneers story visualization using a sequential conditional GAN with a context encoder and dual discriminators to improve narrative and visual coherence. Subsequent works (Song et al. 2020; Li, Torr, and Lukasiewicz 2022; Maharana, Hannan, and Bansal 2021; Maharana and Bansal 2021; Li 2022) improve StoryGAN, while others (Ahn et al. 2023; Chen et al. 2022) adopt Transformer-based methods to enhance character consistency. StoryDALL-E (Maharana, Hannan, and Bansal 2022) extends the story visualization to story continuation with a given initial frame and pre-trained DALL-E (Ramesh et al. 2021). Recently, diffusion models (DM) (Ho, Jain, and Abbeel 2020) have shown success in image generation. Some works (Pan et al. 2024; Rahman et al. 2023; Feng et al. 2023; Song et al. 2024; Liu et al. 2024; Shen and Elhoseiny 2023; Wang et al. 2024) propose an autoregressive diffusion framework based on previous captions and generated frames for consistency. For example, Story-LDM (Rahman et al. 2023) incorporates a visual memory module to capture the context of previous generated images. However, these autoregressive methods are memory-intensive and often fail to capture the global context of the storyline. RCDMs (Shen et al. 2024) is a two-stage model that predicts the embedding of the unknown clip before generating the corresponding images. StoryImager (Tao et al. 2024) is a unified framework for story visualization, continuation, and completion. StoryGPT-V (Shen and Elhoseiny 2023) combines the image generation capability of LDM with the reasoning ability of Large Language Model (LLM) to ensure semantic consistency. TaleCrafter (Gong et al. 2023), Animate-A-Story (He et al. 2023), and AutoStory (Wang et al. 2023) focus on designing system pipelines for story visualization, all employing LLM to generate storylines. In contrast, our ContextualStory addresses consistency by leveraging SETA to capture complex spatial and temporal dependencies, departing from autoregressive methods.

Text-to-image generation. Recently, significant progress (Rombach et al. 2022; Saharia et al. 2022; Ramesh et al. 2022) has been achieved in T2I generation, primarily due to advancements in DM (Ho, Jain, and Abbeel 2020). Another line of work (Dhariwal and Nichol 2021; Ho and Salimans 2022; Ruiz et al. 2023; Kumari et al. 2023) focuses on flexible and controllable image generation, including ControlNet (Zhang, Rao, and Agrawala 2023), Composer (Huang et al. 2023), IP-Adapter (Ye et al. 2023), and T2I-Adapter (Mou et al. 2024). ControlNet provides a general pipeline for conditioning on both text and image data. The Diffusion Transformer (Peebles and Xie 2023) showcases scalability by replacing UNet with a Transformer, and Pixart-α (Chen et al. 2023) further reduces training costs while achieving superior image quality. However, these methods focus on generating individual images aligned with text and struggle to produce multiple coherent and consistent images in a sequence.

Method

Story visualization aims to generate a sequence of images $\tilde{\mathcal{I}} = \{\tilde{I}^1, \dots \tilde{I}^N\}$ that align with a multi-sentence storyline $\mathcal{S} = \{S^1, \dots S^N\}$, ensuring consistency in characters and scenes throughout. For the story continuation task, the first frame I^1 is provided as additional input, guiding the generation of subsequent images $\tilde{\mathcal{I}} = \{\tilde{I}^2, \dots \tilde{I}^N\}$ by extracting and maintaining characters and scenes, eliminating the need to generate them from scratch. During training, ground truth images are denoted as $\mathcal{I} = \{I^1, \dots I^N\}$.

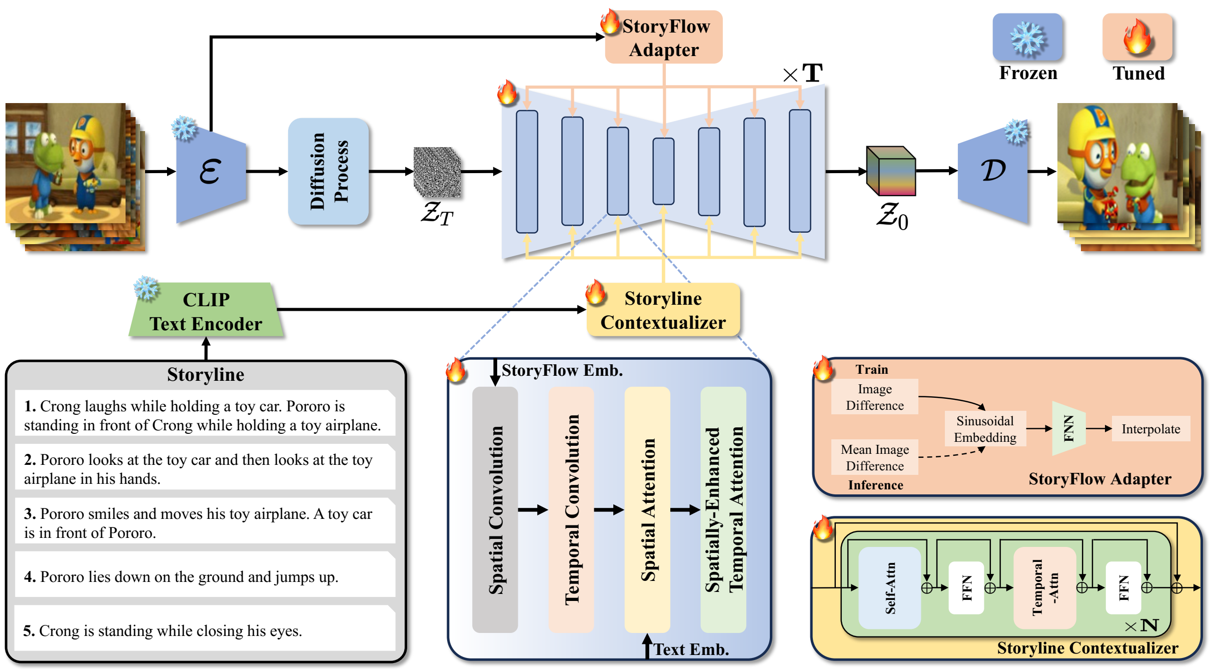

Figure 2: Architecture of ContextualStory for Story Visualization. Each UNet block includes temporal convolution and *Spatially-Enhanced Temporal Attention* to effectively capture complex spatial and temporal dependencies. The *Storyline Contextualizer* enriches the storyline embedding by integrating context information from all text embeddings, while the *StoryFlow Adapter* measures scene changes by computing differences between adjacent frames.

Preliminaries

Diffusion models (DM) (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020) are generative models that approximate data distributions by iteratively denoising a Gaussian distribution through a reverse process of a Markov Chain. Given a training sample $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ and add Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to the input in a forward process $q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I})$, where $\bar{\alpha}_t := \prod_{s=1}^{t}(1-\beta_s)$ and $\beta_1, \ldots, \beta_T$ is the variance schedule. The model is trained to approximate the backward process $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ by minimizing the mean squared error (MSE) between the predicted and target noise $L_{DM} := \mathbb{E}_{\mathbf{x},\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2\right]$.

Latent Diffusion Models (LDM) (Rombach et al. 2022) extend DM to high-dimensional data by compressing images into latent space. An encoder $\mathcal{E}$ maps the input $\mathbf{x}$ to a latent representation $\mathbf{z} = \mathcal{E}(\mathbf{x})$, , where the forward and backward processes are applied. The denoising network $\epsilon_\theta(\mathbf{z}_t, t, \mathbf{c})$ is trained by minimizing $L_{LDM} := \mathbb{E}_{\mathcal{E}(\mathbf{x}),\mathbf{c},\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{c})\|^2\right]$, where $\mathbf{c}$ denotes conditional signals, such as storyline embeddings. The generated image $\hat{\mathbf{x}}$ is obtained by decoding the denoised latent $\mathbf{z}$ with pre-trained decoder $\mathcal{D}(\mathbf{z})$.

Model Architecture

Previous methods based on T2I diffusion models typically use an autoregressive approach, generating story frames sequentially with each frame relying on the preceding frames and captions. However, these methods often fail to capture sufficient storyline context, leading to poor frame consistency. Moreover, the UNet struggles to capture temporal dependencies, while vanilla temporal attention layers are ineffective in addressing significant character movement across frames. To overcome these challenges, as shown in Figure 2, we introduce temporal convolution and SETA into the UNet. These components enable the model to capture contextual information across both spatial and temporal dimensions, allowing it to better handle complex spatial and temporal dependencies. We also propose the Storyline Contextualizer that ensures contextual information propagates to each sentence. Additionally, to address the significant changes in characters and scenes, we introduce the StoryFlow Adapter to quantify these changes and guide the model in generating more coherent visual stories.

Spatially-Enhanced Temporal Attention. In video diffusion models (Zhang et al. 2023), temporal attention layers are often employed to model temporal dependencies. However, unlike video frames with minimal changes and redundant pixels, story frames feature significant character and scene changes. As shown in Figure 3, significant character movement across story frames make it challenging for vanilla temporal attention to capture the same character.

To address this challenge, we propose Spatially-Enhanced Temporal Attention. Assuming the green block within the red-bordered area is the query, the query itself, along with the green blocks covered by the $k \times k$ local window at the same position across other frames (*i.e.*, all the green blocks), form the key and value. Formally, given a hidden state $\mathcal{Z}_t = \{\mathbf{z_t^1}, \ldots, \mathbf{z_t^N}\} \in \mathbb{R}^{n\times c\times h\times w}$, where $n = N$ is

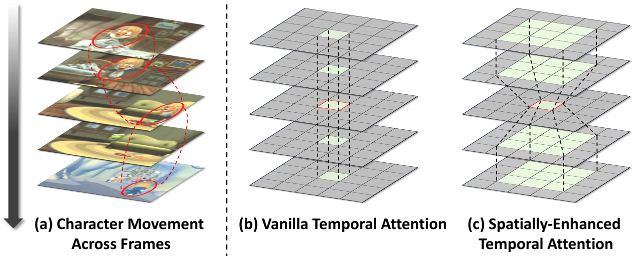

Figure 3: Spatially-Enhanced Temporal Attention leverages a local window mechanism across frames to capture both spatial and temporal dependencies, effectively handling significant character movements.

the number of story frames, and c, h, w represent the channel, height, and width dimensions of the hidden state, respectively. We first reshape $\mathcal{Z}_t$ to $\mathcal{Z}_t' \in \mathbb{R}^{hw \times n \times c}$, then extract the local window feature $\mathcal{Z}_t^{lw'} \in \mathbb{R}^{hw \times n^{lw} \times c}$ at each spatial position, where $n^{lw} = (n-1)k^2 + 1$. Subsequently, we compute the query, key, and value and then perform the $\mathrm{Attention}(Q_T, K_T, V_T)$ through Eq. (2).

$$Q_S = \mathcal{Z}_t' W_S^Q, K_S = \mathcal{Z}_t^{lw'} W_S^K, V_S = \mathcal{Z}_t^{lw'} W_S^V, \quad (1)$$

$$\mathrm{Attention}(Q, K, V) = \mathrm{Softmax}(\frac{QK^T}{\sqrt{d}})V, \quad (2)$$

where W_T^Q, W_T^K, W_T^V are learnable projection matrices, and d is the feature dimensionality. To ensure a complete local window at the boundaries, we pad $\mathcal{Z}_t'$ by replicating the boundary features. We utilize rotary positional embedding (RoPE) (Su et al. 2024) as the temporal positional embedding to enable the model understand temporal relationships between frames efficiently.

Storyline Contextualizer. We use the pre-trained CLIP text encoder to independently extract text embeddings for each sentence in the storyline. These embeddings contain the semantic information of the corresponding sentences but lack the global contextual information of the storyline. Directly using these text embeddings to guide the model may result in inconsistent story frames. To address this challenge, we propose the Storyline Contextualizer, which propagates and integrates the contextual information from all text embeddings to generate a context-enriched storyline embedding. As shown in Figure 2, the Storyline Contextualizer is a transformer-based network, each layer contains a self-attention layer, a temporal attention layer, and two Feed-Forward Networks (FFNs).

Given the storyline embedding $\mathcal{C} = \{\mathbf{c^1}, \dots, \mathbf{c^N}\} \in \mathbb{R}^{n \times l \times c_T}$ from the CLIP text encoder, where $n = N$ is the number of story sentences, l is the feature sequence length, and c_T is the feature dimension. In Storyline Contextualizer, we first reshape the storyline embedding to $1 \times nl \times c_T$ and apply self-attention, then reshape it to $l \times n \times c_T$ for temporal attention. To minimize any adverse effects from additional modules, we zero-initialize the weights of the second FFN in the final layer and incorporate a residual connection, ensuring that the Storyline Contextualizer functions as an identity mapping at the beginning of training. The context-enriched storyline embedding $\mathcal{C}'$ guides the model through a cross-attention layer. Unlike T2I and T2V models that compute cross-attention between images/frames and a single text, we adopt a temporally-aligned approach, computing cross-attention between each individual context-enriched text embedding $\mathbf{c^i}$ and hidden state $\mathbf{z_t^i}$ of the corresponding frame.

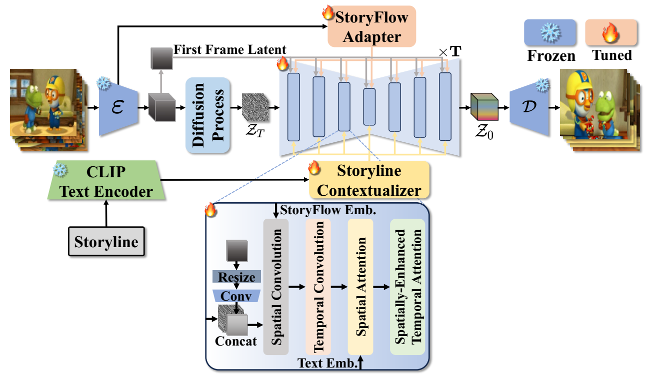

Figure 4: Architecture of ContextualStory for Story Continuation. The first frame latent is used as additional input for all UNet blocks, resized and adjusted with a 1×1 convolution layer before concatenation with the hidden state.

StoryFlow adapter. To leverage scene changes between story frames to guide the model, we propose the StoryFlow Adapter, inspired by (Jeong, Park, and Ye 2023; Wu et al. 2023; Qing et al. 2023). As shown in Figure 2, the storyflow is computed as the L2 norm of differences between adjacent images, $\delta_i = \left\|\mathbf{z_0^i} - \mathbf{z_0^{i+1}}\right\|$, to quantify the image difference. Given N ground truth frames, we calculate the storyflow $\Delta = \{\delta^1, \dots, \delta^{N-1}\}$. We then encode the storyflow Δ into a c-dimensional embedding using sinusoidal embedding and a zero-initialized FFN. Through linear interpolation, we obtain the storyflow embedding $\Delta' \in \mathbb{R}^{N \times c}$. Finally, we add storyflow embedding to timestep embedding and feed them into the spatial convolution of the UNet block. During inference, we use the average of the storyflows computed from the training set as the storyflow Δ.

Solving the Story Continuation Task

For story continuation tasks, besides the storyline, the first frame is provided as an additional input. As shown in Figure 4, we extract the first frame latent $\mathbf{z^1}$ and use it as an additional input to all UNet blocks. Within each UNet block, $\mathbf{z^1}$ is resized to match the spatial dimensions of the hidden state, then a 1×1 convolution layer adjusts the channels to match the hidden state. Finally, it is concatenated with the hidden state before inputting into the spatial convolution.

Experiment

Experimental Setup

Datasets. We employ two popular benchmark datasets, PororoSV (Li et al. 2019) and FlintstonesSV (Gupta et al. 2018), to evaluate the performance of our model on story

Model	PororoSV			FlintstonesSV		
	FID ↓	Char. F1 ↑	Frm. Acc. ↑	FID ↓	Char. F1 ↑	Frm. Acc. ↑
StoryGANc	74.63	39.68	16.57	90.29	72.80	58.39
StoryDALL-E	25.90	36.97	17.26	26.49	73.43	55.19
MEGA-StoryDALL-E	23.48	39.91	18.01	23.58	74.26	54.68
Story-LDM	26.64	47.56	29.19	24.24	76.59	57.19
AR-LDM	17.40	-	-	19.28	-	-
Causal-Story	16.98	-	-	19.03	-	-
StoryImager	15.45	-	-	18.32	-	-
RCDMs	16.25	59.03	41.48	14.96	85.51	78.44
ContextualStory	**13.86**	**76.25**	**50.72**	**13.27**	**91.29**	**81.91**

Table 1: Quantitative comparison with SOTA methods of story continuation on PororoSV and FlintstonesSV.

Model	FID ↓	Char. F1 ↑	Frm. Acc. ↑
PororoSV			
StoryGAN	158.06	18.59	9.34
CP-CSV	149.29	21.78	10.03
DUCO	96.51	38.01	13.97
VLC	84.96	43.02	17.36
VP-CSV	65.51	56.84	25.87
Word-Level SV	56.08	-	-
Story-LDM	27.33	-	-
AR-LDM	16.59	-	-
Causal-Story	16.28	-	-
StoryImager	15.63	-	-
ContextualStory	**13.61**	**77.24**	**51.59**
FlintstonesSV			
StoryGAN	127.19	46.20	32.96
DUCO	78.02	54.92	36.34
VLC	72.87	58.81	39.18
Story-LDM	36.55	-	-
AR-LDM	23.59	-	-
StoryImager	22.27	-	-
ContextualStory	**20.15**	**91.70**	**83.08**

Table 2: Quantitative comparison with SOTA methods of story visualization on PororoSV and FlintstonesSV.

Model	Memory (GB)↓	Inference Speed (s)↓
StoryDALL-E	20	347
Story-LDM	11	18.5
AR-LDM	40	40.4
StoryGen	29	31.7
StoryGPT-V	25	14.1
RCDMs	22	30.4
ContextualStory	**5**	**11.8**

Table 3: Comparison of GPU memory usage and inference speed across SOTA models.

Dataset	Attribute	Ours	Tie	AR-LDM
PororoSV	Visual Quality	**81.0%**	6.9%	12.1%
	Semantic Relevance	**85.6%**	9.2%	5.2%
	Temproal Consistency	**84.1%**	8.8%	7.1%
FlintstonesSV	Visual Quality	**80.4%**	6.2%	13.4%
	Semantic Relevance	**82.6%**	6.3%	11.1%
	Temproal Consistency	**84.8%**	5.4%	9.8%

Table 4: Human evaluations of story visualization task. Ours (%) means our ContextualStory is preferred over AR-LDM. AR-LDM (%) means AR-LDM is preferred over our ContextualStory. Tie (%) means the annotator believes that the two image sequences are similar.

visualization and story continuation tasks. PororoSV contains 10,191, 2,334, and 2,208 stories within the train, validation, and test splits, respectively, featuring 9 main characters. FlintstonesSV contains 20,132, 2,071, and 2,309 stories within the train, validation, and test splits, respectively, featuring 7 main characters and 323 backgrounds. Each story in both datasets comprises 5 consecutive story images.

Automatic metrics. To evaluate the quality of generated images, we employ the following three evaluation metrics following previous works (Maharana, Hannan, and Bansal 2022; Pan et al. 2024) in story visualization: (1) Frechet Inception Distance (FID) (Heusel et al. 2017), which measures the distance between feature vectors of ground truth and generated frames; (2) Frame accuracy (Frm. Acc.), which evaluates character matching to ground truth using a fine-tuned Inception-v3 model; (3) Character F1-score (Char. F1), which assesses the quality of generated characters using the same Inception-v3 model as Frm. Acc.

Implementation details. We initialize ContextualStory with the pre-trained Stable Diffusion 2.1-base and fine-tune only the UNet parameters with the AdamW optimizer. Training is performed on 4 NVIDIA A800 GPUs with a batch size of 12, a learning rate of 5×10^{-5} and 40,000 iterations for PororoSV and 80,000 iterations for FlintstonesSV. The SETA window size is $k = 3$, and the SC layer count is 4. During training, we apply classifier-free guidance by randomly dropping input storylines with a 0.1 probability and use the PYoCo mixed noise prior for noise initialization. For inference, we use the DDIM sampler with 50 steps and a guidance scale of 7.5 to generate 256×256 images.

Quantitative Results

Story Visualization. Table 2 presents quantitative results for story visualization on both PororoSV and FlintstonesSV, comparing ContextualStory to several SOTA methods, including StoryGAN, CP-CSV (Song et al. 2020), DUCO (Maharana, Hannan, and Bansal 2021), VLC (Maharana and Bansal 2021), VP-CSV (Chen et al. 2022), Word-Level SV (Li 2022), Story-LDM, AR-LDM, Causal-Story (Song et al. 2024), and StoryImager. The results clearly demonstrate that ContextualStory significantly outperforms existing SOTA methods across all metrics on both datasets. This superior performance is primarily due to SETA, SC, and StoryFlow Adapter, which effectively utilize context information to generate coherent story frames.

Story Continuation. Table 1 presents the quantitative results for story continuation on both PororoSV and FlintstonesSV. We evaluate the effectiveness of ContextualStory model against several SOTA methods, such as StoryDALL-

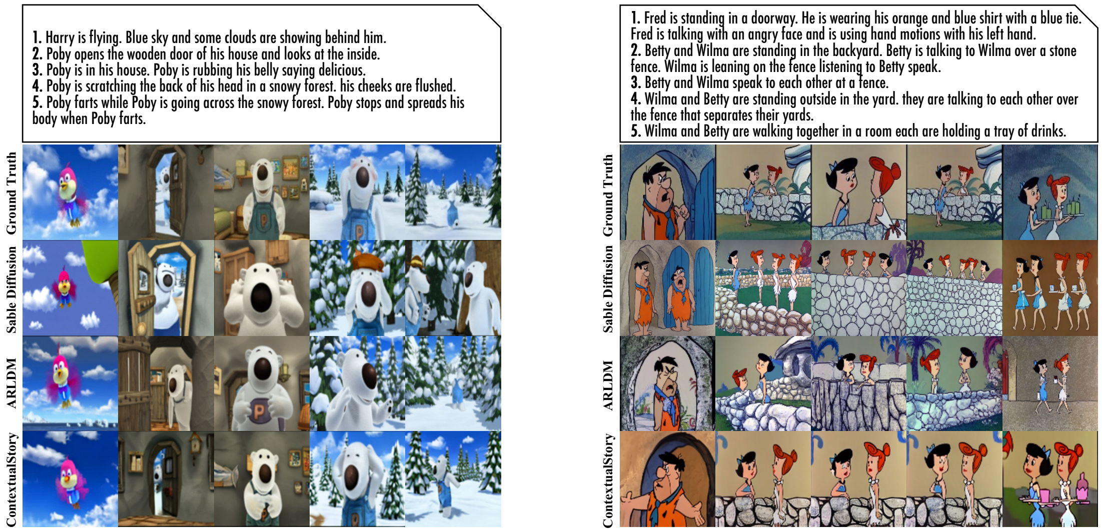

Figure 5: Qualitative comparison of story visualization on PororoSV (left) and FlintstonesSV (right).

E, MEGA-StoryDALL-E (Maharana, Hannan, and Bansal 2022), StoryImager, and RCDMs. The results demonstrate that ContextualStory outperforms existing methods by a large margin across all metrics for the story continuation on both datasets. This indicates that ContextualStory better utilize contextual information to generate coherent story frames based on the storyline and the first image.

Inference Speed. We compare the GPU memory usage and inference speed of recent open-source SOTA models (e.g., AR-LDM and StoryGen) in Table 3. The experiment is conducted on an A800 GPU with 50 DDIM steps to ensure a fair comparison. Autoregressive methods like Story-LDM, AR-LDM, StoryGen, and StoryGPT-V suffer from high memory usage and slow inference speeds. In contrast, ContextualStory, a non-autoregressive model, not only overcomes the bottleneck of autoregressive methods by achieving the lowest memory usage and fastest inference speed, but also outperforms SOTA methods in overall performance.

Qualitative Results

Story Visualization. Figure 5 shows a qualitative comparison of story visualization on PororoSV and FlintstonesSV. Stable Diffusion (SD) generates high-quality images independently from individual sentences, but its lack of contextual awareness leads to inconsistent character appearances and character duplication. AR-LDM avoids character duplication but still struggles with inconsistent character appearances. In contrast, ContextualStory produces high-quality images with coherent and consistent characters and scenes across both datasets.

Story Continuation. Figure 6 demonstrates a qualitative comparison of story continuation on PororoSV and FlintstonesSV datasets. StoryDALL-E produces low-quality characters with inconsistent backgrounds. AR-LDM generates higher-quality characters, but the backgrounds lack consistency and deviate significantly from the ground truth. In contrast, ContextualStory generates high-quality images with consistent characters and backgrounds that closely match the ground truth. More results are provided in the supplementary material.

Human Evaluation

Due to the limitations of metrics such as FID, Char. F1, and Frm. Acc. in accurately reflecting the quality of generated story frames, we conducted human evaluations for the story visualization task on PororoSV and FlintstonesSV, focusing on *Visual Quality*, *Semantic Relevance*, and *Temporal Consistency*. We randomly selected 300 pairs of story frame sequences generated from AR-LDM (Pan et al. 2024) and our ContextualStory. Annotators were tasked to select the better sequence for the three attributes: Visual Quality, Semantic Relevance, and Temporal Consistency. Each pair of story frame sequences was evaluated by 10 annotators. As shown in Table 4, the results indicate that our ContextualStory outperforms AR-LDM significantly across all three attributes.

Ablation Studies

Ablation study of the proposed components. To evaluate the benefit of each proposed component, we conduct an ablation study on the story visualization task using PororoSV. As shown in Table 5, progressively removing components from ContextualStory results in a consistent decline across all three metrics. The removal of SETA has the most significant effect, increasing FID by 16.0%, and reducing Char. F1 and Frm. Acc. by 4.9% and 7.4%, respectively. The qualitative comparison in Figure 7 shows the following: 1) Removing the StoryFlow Adapter slightly reduces background consistency. 2) Further removing SC leads to duplicated characters, like Loopy. 3) Removing SETA reduces background consistency, introduces duplicated characters (*e.g.*, Pororo), and incorrect characters (*e.g.*, Petty and Poby), making images less accurate. 4) Removing Temporal Convolution further decreases character and scene consistency. These results

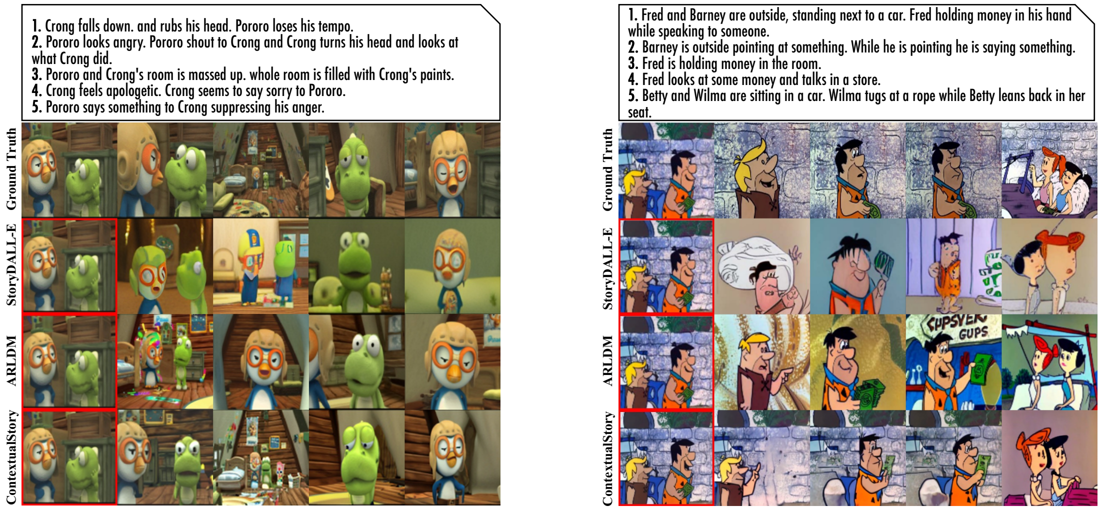

Figure 6: Qualitative comparison of story continuation on PororoSV (left) and FlintstonesSV (right). The image marked with a red box is the first frame additionally input to the model.

Figure 7: Qualitative results of the ablation study on the proposed components for story visualization on PororoSV.

indicate that all proposed components contribute to the performance of ContextualStory, with SETA having the most significant impact.

Ablation study of temporal attention. Table 5 presents the ablation study results of comparing Vanilla Temporal Attention and our proposed SETA for the story visualization task on PororoSV. The results clearly show that SETA outperforms Vanilla Temporal Attention across all metrics. Specifically, SETA achieves a lower FID score, indicating better alignment with ground truth images, and higher Char. F1 and Frm. Acc., demonstrating improved character consistency and accuracy. These improvements highlight the effectiveness of the local window mechanism of SETA, which allows the model to better capture both spatial and temporal dependencies, leading to more coherent and consistent story frames. The significant gains in performance suggest that incorporating spatial context within temporal attention is crucial for enhancing visual storytelling models. More ablation studies are provided in the supplementary material.

Model	FID ↓	Char. F1 ↑	Frm. Acc. ↑
ContextualStory	**13.61**	**77.24**	**51.59**
−StoryFlow Adapter	14.84	77.09	50.48
−Storyline Contextualizer	15.02	75.42	48.39
−SETA	17.42	71.70	44.83
−Temporal Convolution	19.69	68.12	39.60
Vanilla Temporal Attention	14.78	75.94	48.79
SETA (Ours)	**13.61**	**77.24**	**51.59**

Table 5: Ablation study of the proposed components and temporal attention for story visualization on PororoSV.

Conclusion

In this paper, we propose ContextualStory, a novel framework that overcomes the limitations of traditional autoregressive methods in visual storytelling. By incorporating Spatially-Enhanced Temporal Attention, we effectively capture spatial and temporal dependencies, ensuring consistency in characters and scenes across frames. Additionally, the Storyline Contextualizer enriches the global context from storyline, while the StoryFlow Adapter enhances the ability of model to handle scene changes. Extensive experiments on PororoSV and FlintstonesSV show that ContextualStory achieves superior performance over state-of-the-art methods in both story visualization and continuation tasks.

Acknowledgments

This paper is supported by the Doubao Fund.

References

Ahn, D.; Kim, D.; Song, G.; Kim, S. H.; Lee, H.; Kang, D.; and Choi, J. 2023. Story visualization by online text augmentation with context memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3125–3135.

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Brooks, T.; Peebles, B.; Holmes, C.; DePue, W.; Guo, Y.; Jing, L.; Schnurr, D.; Taylor, J.; Luhman, T.; Luhman, E.; Ng, C.; Wang, R.; and Ramesh, A. 2024. Video generation models as world simulators.

Chen, H.; Han, R.; Wu, T.-L.; Nakayama, H.; and Peng, N. 2022. Character-centric story visualization via visual planning and token alignment. *arXiv preprint arXiv:2210.08465*.

Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wu, Y.; Wang, Z.; Kwok, J.; Luo, P.; Lu, H.; and Li, Z. 2023. PixArt-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. arXiv:2310.00426.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*.

Feng, Z.; Ren, Y.; Yu, X.; Feng, X.; Tang, D.; Shi, S.; and Qin, B. 2023. Improved visual story generation with adaptive context modeling. *arXiv preprint arXiv:2305.16811*.

Gong, Y.; Pang, Y.; Cun, X.; Xia, M.; Chen, H.; Wang, L.; Zhang, Y.; Wang, X.; Shan, Y.; and Yang, Y. 2023. Talecrafter: Interactive story visualization with multiple characters. *arXiv preprint arXiv:2305.18247*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Gupta, T.; Schwenk, D.; Farhadi, A.; Hoiem, D.; and Kembhavi, A. 2018. Imagine this! scripts to compositions to videos. In *Proceedings of the European conference on computer vision (ECCV)*, 598–613.

He, Y.; Xia, M.; Chen, H.; Cun, X.; Gong, Y.; Xing, J.; Zhang, Y.; Wang, X.; Weng, C.; Shan, Y.; et al. 2023. Animate-a-story: Storytelling with retrieval-augmented video generation. *arXiv preprint arXiv:2307.06940*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Huang, L.; Chen, D.; Liu, Y.; Shen, Y.; Zhao, D.; and Zhou, J. 2023. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*.

Jeong, H.; Park, G. Y.; and Ye, J. C. 2023. VMC: Video Motion Customization using Temporal Attention Adaption for Text-to-Video Diffusion Models. *arXiv preprint arXiv:2312.00845*.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1931–1941.

Li, B. 2022. Word-level fine-grained story visualization. In *European Conference on Computer Vision*, 347–362. Springer.

Li, B.; Torr, P. H.; and Lukasiewicz, T. 2022. Clustering generative adversarial networks for story visualization. In *Proceedings of the 30th ACM International Conference on Multimedia*, 769–778.

Li, Y.; Gan, Z.; Shen, Y.; Liu, J.; Cheng, Y.; Wu, Y.; Carin, L.; Carlson, D.; and Gao, J. 2019. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6329–6338.

Liu, C.; Wu, H.; Zhong, Y.; Zhang, X.; Wang, Y.; and Xie, W. 2024. Intelligent Grimm-Open-ended Visual Storytelling via Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6190–6200.

Maharana, A.; and Bansal, M. 2021. Integrating visuospatial, linguistic and commonsense structure into story visualization. *arXiv preprint arXiv:2110.10834*.

Maharana, A.; Hannan, D.; and Bansal, M. 2021. Improving generation and evaluation of visual stories via semantic consistency. *arXiv preprint arXiv:2105.10026*.

Maharana, A.; Hannan, D.; and Bansal, M. 2022. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. In *European Conference on Computer Vision*, 70–87. Springer.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; and Shan, Y. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4296–4304.

Pan, X.; Qin, P.; Li, Y.; Xue, H.; and Chen, W. 2024. Synthesizing coherent story with auto-regressive latent diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2920–2930.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4195–4205.

Qing, Z.; Zhang, S.; Wang, J.; Wang, X.; Wei, Y.; Zhang, Y.; Gao, C.; and Sang, N. 2023. Hierarchical spatio-temporal decoupling for text-to-video generation. *arXiv preprint arXiv:2312.04483*.

Rahman, T.; Lee, H.-Y.; Ren, J.; Tulyakov, S.; Mahajan, S.; and Sigal, L. 2023. Make-a-story: Visual memory conditioned consistent story generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2493–2502.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. Pmlr.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Shen, F.; Ye, H.; Liu, S.; Zhang, J.; Wang, C.; Han, X.; and Yang, W. 2024. Boosting Consistency in Story Visualization with Rich-Contextual Conditional Diffusion Models. *arXiv preprint arXiv:2407.02482*.

Shen, X.; and Elhoseiny, M. 2023. Large Language Models as Consistent Story Visualizers. *arXiv preprint arXiv:2312.02252*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, T.; Cao, J.; Wang, K.; Liu, B.; and Zhang, X. 2024. Causal-story: Local causal attention utilizing parameter-efficient tuning for visual story synthesis. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3350–3354. IEEE.

Song, Y.-Z.; Rui Tam, Z.; Chen, H.-J.; Lu, H.-H.; and Shuai, H.-H. 2020. Character-preserving coherent story visualization. In *European Conference on Computer Vision*, 18–33. Springer.

Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; and Liu, Y. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063.

Tao, M.; Bao, B.-K.; Tang, H.; Wang, Y.; and Xu, C. 2024. StoryImager: A Unified and Efficient Framework for Coherent Story Visualization and Completion. *arXiv preprint arXiv:2404.05979*.

Wang, W.; Zhao, C.; Chen, H.; Chen, Z.; Zheng, K.; and Shen, C. 2023. AutoStory: Generating Diverse Storytelling Images with Minimal Human Effort. *arXiv preprint arXiv:2311.11243*.

Wang, X.; Wang, Y.; Tsutsui, S.; Lin, W.; Wen, B.; and Kot, A. C. 2024. Evolving Storytelling: Benchmarks and Methods for New Character Customization with Diffusion Models. *arXiv preprint arXiv:2405.11852*.

Wu, R.; Chen, L.; Yang, T.; Guo, C.; Li, C.; and Zhang, X. 2023. Lamp: Learn a motion pattern for few-shot-based video generation. *arXiv preprint arXiv:2310.10769*.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Zhang, D. J.; Wu, J. Z.; Liu, J.-W.; Zhao, R.; Ran, L.; Gu, Y.; Gao, D.; and Shou, M. Z. 2023. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.