StoryDiffusion: Consistent Self-Attention for Long-Range Image and Video Generation

Yupeng Zhou[1*] **Daquan Zhou**[2§†] **Ming-Ming Cheng**[1,3] **Jiashi Feng**[2] **Qibin Hou**[1,3§†]

[1] VCIP & TMCC, CS, Nankai University [2] ByteDance Inc. [3] NKIARI, Futian, Shenzhen
https://StoryDiffusion.github.io

Abstract

For recent diffusion-based generative models, maintaining consistent content across a series of generated images, especially those containing subjects and complex details, presents a significant challenge. In this paper, we propose a simple but effective self-attention mechanism, termed Consistent Self-Attention, that boosts the consistency between the generated images. It can be used to augment pre-trained diffusion-based text-to-image models in a zero-shot manner. Based on the images with consistent content, we further show that our method can be extended to long-range video generation by introducing a semantic space temporal motion prediction module, named Semantic Motion Predictor. It is trained to estimate the motion conditions between two provided images in the semantic spaces. This module converts the generated sequence of images into videos with smooth transitions and consistent subjects that are more stable than the modules based on latent spaces only, especially in the context of long video generation. By merging these two novel components, our framework, referred to as StoryDiffusion, can describe a text-based story with consistent images or videos encompassing a rich variety of contents. The proposed StoryDiffusion encompasses pioneering explorations in visual story generation with the presentation of images and videos, which we hope could inspire more research from the aspect of architectural modifications.

1 Introduction

With extensive pre-training and advanced architectures, diffusion models have shown superior performance in generating very high-quality images and videos over previous generative-adversarial network (GAN) based methods [5]. However, generating subject-consistent (e.g. characters with consistent identity and attire) images and videos to describe a story is still challenging for existing models. The commonly used IP-Adapter [55] taking an image as a reference could be used to guide the diffusion process to generate images similar to it. However, due to the strong guidance, the controllability over the generated content of the text prompts is reduced. On the other hand, recent state-of-the-art identity preservation methods, such as InstantID [47] and PhotoMaker [26], focus on identity controllability but the consistency of the attires and the scenarios cannot be guaranteed. Hence, in this paper, we aim to find a method that can generate images and videos with consistent characters in terms of both identity and attire while maximizing the controllability of the user via text prompts.

A common approach to preserve the consistency between different images (or frames in the context of video generation) is to use a temporal module [15, 4]. However, this requires extensive computational resources and data. Differently, we target to explore a lightweight method with minimum data and computational cost, or even in a zero-shot manner.

[*]During the internship at ByteDance Inc. [§]Project lead. [†]Corresponding authors.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Figure 1: Images and videos generated by our StoryDiffusion. (a) Comic generated by StoryDiffusion telling the story of a man who discovers a treasure while exploring the jungle. (b) Comic generated by StoryDiffusion describing the expedition to the moon by Lecun, with a reference image control [26] same as Fig. 7(b). (c) Videos generated by our StoryDiffusion. Click the image to play the video. Best viewed with *Acrobat Reader*. More generated videos can be found in the uploaded supplementary file.

As evidenced by previous works [45, 19, 6], self-attention is one of the most important modules for modeling the overall structure of the generated visual content. Our main motivation is that we could use some shared reference image information to guide the self-attention calculation, the consistency between generated images is supposed to be improved clearly. As the self-attention weights are input-dependent, model training or fine-tuning might not be required. Following this idea, we propose Consistent Self-Attention, the core of our StoryDiffusion, which can be inserted into the diffusion backbone to replace the original self-attention in a zero-shot manner.

Different from the standard self-attention that operates on the tokens representing a single image, Consistent Self-Attention incorporates reference tokens sampled from reference images during the token similarity matrix calculation and token merging. The sampled tokens use the same set of Q-K-V weights and thus no extra training is required. As shown in Fig. 1, the generated images using Consistent Self-Attention successfully preserve the consistency in both identity and attire, which is vital for storytelling. Intuitively, Consistent Self-Attention builds correlations across images in the batch, generating consistent character images in terms of identity and attire, such as clothes. This enables us to generate subject-consistent images for storytelling.

For any given story text, we begin by dividing it into several prompts, with each prompt corresponding to an individual image. Then our method could generate highly consistent images that effectively narrate a story. To support long story generation, we also implement Consistent Self-Attention together with a sliding window across the generated consistent images. This removes the peak memory consumption's dependency on the input text length, making it possible to generate long stories. To stream the generated story frames into videos, we further propose Semantic Motion Predictor that can predict transitions between two images in the semantic spaces. We empirically found that predicting motions in the semantic space generates more stable results than the predictions in the image latent spaces. Combined with the pre-trained motion module [13], Semantic Motion Predictor can generate smooth video frames that are notably better than recent conditional video generation methods, such as SEINE [7] and SparseCtrl [12].

Our contributions are summarized below:

- We propose a training-free and hot-pluggable attention module, termed Consistent Self-Attention. It can maintain the consistency of characters in a sequence of generated images for storytelling with high text controllability.
- We propose a new motion prediction module that can predict transitions between two images in the semantic space, termed Semantic Motion Predictor. It can generate more stable long-range video frames that can be easily upscaled to minutes than recent popular image conditioning methods, such as SEINE [7] and SparseCtrl [12].
- We demonstrate that our approach could generate long image sequences or videos based on a pre-defined text-based story with the proposed Consistent Self-Attention and Semantic Motion Predictor with motions specified by text prompts. We term the new framework as StoryDiffusion.

2 Related work

2.1 Controllable text-to-image diffusion generation

As an important sub-field of diffusion model applications [40, 17, 36, 38, 27, 52, 41], text-to-image generation [37, 33, 34], has attracted considerable attention recently. In addition, to enhance the controllability of text-to-image generation, a multitude of methods emerged as well. Among them, ControlNet [57] and T2I-Adapter [31] introduce control conditions, such as depth maps, pose images, or sketches, to direct the generation of images. MaskDiffusion [61] and StructureDiffusion [9] focus on enhancing the text controllability. There are also some works [30, 28] controlling the layout of generated images.

ID-Preservation, which is expected to generate images with a specified ID, is also a hot topic. According to whether test-time fine-tuning is required, these works can be divided into two major categories. The first one only requires fine-tuning a part of the model with a given image, such as Textual Inversion [10], DreamBooth [39], and Custom Diffusion [25]. The other one, exemplified by IPAdapter [55] and PhotoMaker [26], leverages models that have undergone pre-training on large datasets, allowing the direct use of a given image to control image generation. Different from both of the two types, we focus on maintaining the subject consistency in multiple images, to narrate a story. Our Consistent Self-Attention is training-free and pluggable and can build connections across images within a batch to generate multiple subject-consistent images.

2.2 Video generation

Due to the success of diffusion models in the field of image generation [37, 17], the exploration in the domain of video generation [13, 23, 42, 49, 54, 56] is also becoming popular. VDM [15] is among the first that extends the 2D U-Net from image diffusion models to a 3D U-Net to achieve text-based generation. Later works, such as MagicVideo [60] and Mindscope [46], introduce 1D temporal attention mechanisms, reducing computations by building upon latent diffusion models. Following Imagen, Imagen Video [16] employs a cascaded sampling pipeline that generates videos through multiple stages.

In addition to traditional end-to-end text-to-video (T2V) generation, video generation using other conditions is also an important direction. This type of methods generates videos with other auxiliary

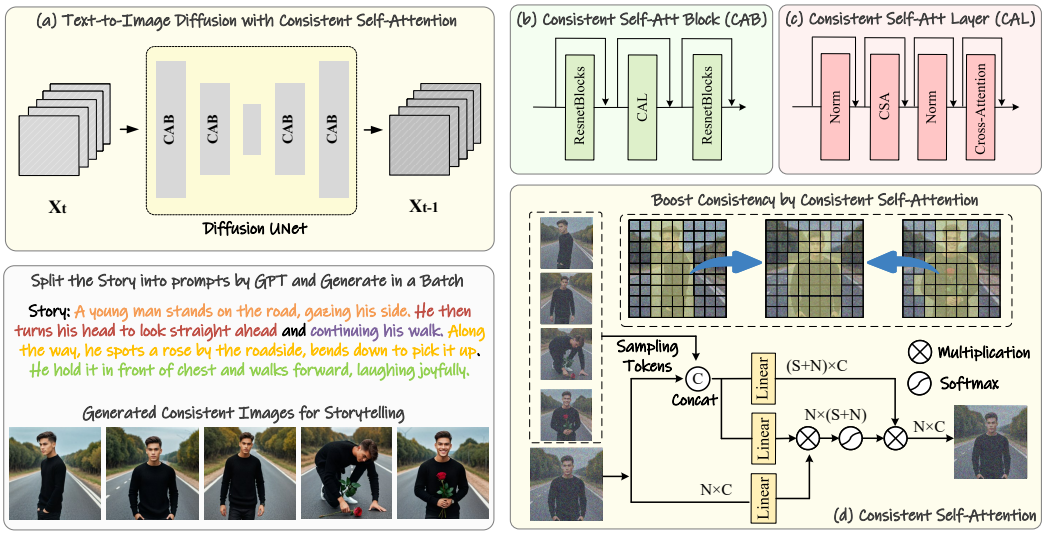

Figure 2: The Pipeline of StoryDiffusion to generating subject-consistent images. To create subject-consistent images to describe a story, we incorporate our Consistent Self-Attention into the pre-trained text-to-image diffusion model. We split a story text into several prompts and generate images using these prompts in a batch. Consistent Self-Attention builds connections among multiple images in a batch for subject consistency.

controls, such as depth maps [12, 14], pose maps [53, 21, 48, 29], RGB images [3, 7, 32], or other guided motion videos [59, 51].

Our video generation method focuses on transition video generation, which is expected to generate videos with a given start frame and an end frame. Typical related works are SEINE [7] and SparseCtrl [12]. SEINE randomly masks video sequences as the initial input of the video diffusion models in training to enable the predictions of the transition between two frames. SparseCtrl introduces a sparse control network to synthesize the corresponding control information for each frame using sparse control data, thereby directing the generation of videos. However, the aforementioned transition video generation methods rely solely on temporal networks in image latent space for the predictions of intermediate content. Thus, these methods often perform poorly on complex transitions, such as large-scale movements of characters. Our StoryDiffusion aims to perform predictions in image semantic spaces to achieve better performance and can handle larger movements, which we will show in our experiment section.

3 Method

Our method can be divided into two stages, as shown in Fig. 2 and Fig. 3. In the first stage, StoryDiffusion utilizes Consistent Self-Attention to generate subject-consistent images in a training-free manner. These consistent images can be directly applied to storytelling and can also serve as input for the second stage. In the second stage, our StoryDiffusion create consistent transition videos based on these consistent images.

3.1 Training-free consistent images generation

The key to addressing the above issues lies in how to maintain consistency of characters within a batch of images. This means we need to establish connections between images within a batch during generation. The previous image and video editing methods [6, 50] use DDIM inversion and insert additional keys and values in attention calculation [40] to keep similarity. Unlike existing methods [24, 20, 11] applied to single images or highly similar video clips, we aim to generate a set of images where the character remains consistent, yet each image portrays different scenes and actions, for use in anime production or story-boarding. Therefore, we aim to share intermediary tokens within

a batch of images to enable mutual interaction through self-attention computation, thereby preserving consistency. We obtain these intermediary tokens by randomly sampling some pixels from the image batch before attention calculation, enabling plug-and-play capability and eliminating the need for training. We name this operation as Consistent Self-Attention and insert it into the location of the original self-attention in the existing U-Net architecture of image generation models and reuse the original self-attention weights.

Formally, given a batch of image features $\mathcal{I} \in \mathbb{R}^{B \times N \times C}$, where B, N, and C are the batch size, number of tokens in each image, and channel number, respectively, we define a function $\mathrm{Attention}(X_k, X_q, X_v)$ to calculate self-attention. X_k, X_q, and X_v stand for the query, key, and value used in attention calculation, respectively. The original self-attention is performed within each image feature I_i in $\mathcal{I}$ independently. The feature I_i is projected to Q_i, K_i, V_i and sent into the attention function, yielding:

$$O_i = \mathrm{Attention}\left(Q_i, K_i, V_i\right). \tag{1}$$

To build interactions among the images within a batch to keep subject consistency, our Consistent Self-Attention samples some tokens S_i from other image features in the batch:

$$S_i = \mathrm{RandSample}\left(I_1, I_2, .., I_{i-1}, I_{i+1}, ..., I_{B-1}, I_B\right), \tag{2}$$

where $\mathrm{RandSample}$ denotes the random sampling function. After sampling, we pair the sampled tokens S_i and the image feature I_i to form a new set of tokens P_i. We then perform linear projections on P_i to generate the new key K_{Pi} and value V_{Pi} for Consistent Self-Attention. Here, the original query Q_i is not changed. Finally, we compute the self-attention as follows:

$$O_i = \mathrm{Attention}\left(Q_i, K_{Pi}, V_{Pi}\right). \tag{3}$$

Given the paired tokens, our method performs the self-attention across a batch of images, facilitating interactions among features of different images. This type of interaction promotes the model to the convergence of characters, faces, and attires during the generation process. Despite the simple and training-free manner, our Consistent Self-Attention can efficiently generate subject-consistent images, which we will demonstrate in detail in our experiments. These images serve as illustrations to narrate a complex story as shown in Fig. 2. To make it clearer, we also show the pseudo code in Algorithm **??** in the Appendix.

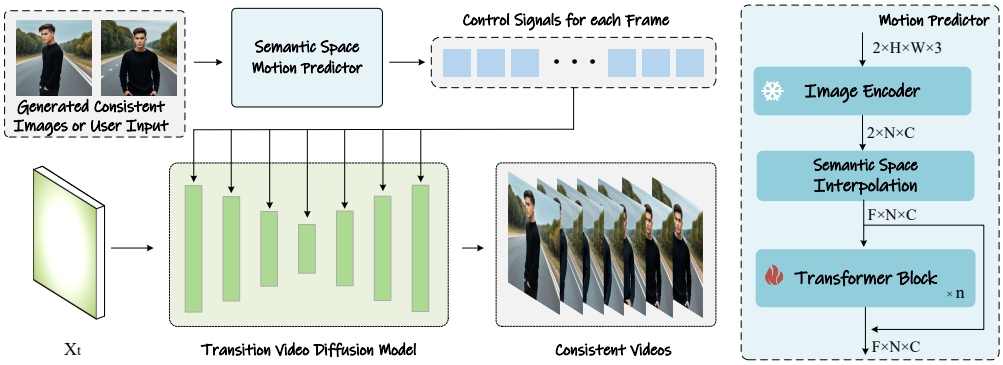

Figure 3: The pipeline of our method for generating transition videos for obtaining subject-consistent images, as described in Sec. 3.1. To effectively model the character's large motions, we encode the conditional images into the image semantic space for encoding spatial information and predict the transition embeddings. These predicted embeddings are then decoded using the video generation model, with the embeddings serving as control signals in cross-attention to guide the generation of each frame.

3.2 Semantic Motion Predictor for video generation

Illustrated in Fig. 3, the sequence of the generated character-consistent images can be further refined to videos by inserting frames between each pair of adjacent images. This can be regarded as a video generation task with known start and end frames as conditions. However, we empirically observed

that recent methods, such as SparseCtrl [12] and SEINE [7], cannot join two condition images stably when the difference between the two images is large. We argue that this limitation stems from their sole reliance on temporal modules to predict intermediate frames, which may be not enough to handle the large state gap between the image pair. The temporal module operates within pixels on each spatial location independently, therefore, there may be insufficient consideration of spatial information when inferring intermediate frames. This makes it difficult to model the long-distance and physically meaningful motion.

To address this issue, we propose Semantic Motion Predictor, which encodes the image into the image semantic space to capture the spatial information, achieving more accurate motion prediction from a given start frame and an end frame. More specifically, in our Semantic Motion Predictor, we first use a function E to establish a mapping from the RGB images to vectors in the image semantic space, encoding the spatial information. Instead of directly using linear layers as E, we utilize a pre-trained CLIP image encoder as E to leverage its zero-shot capabilities for enhancing performance. Using E, the given start frame F_s and end frame F_e are compressed to image semantic space vectors K_s, K_e.

$$K_s, K_e = E\left(F_s, F_e\right). \tag{4}$$

Subsequently, in the image semantic space, we train a transformer-based structure predictor to perform predictions of each intermediate frame. The predictor first performs linear interpolation to expand the two frames K_s and K_e into sequence $K_1, K_2, ..., K_L$, where L is the required video length. Then, the sequence $K_1, K_2, ..., K_L$ is sent into a series of transformer blocks B to predict the transition frames:

$$P_1, P_2, ..., P_l = B\left(K_1, K_2, ..., K_l\right). \tag{5}$$

Next, we need to decode these predicted frames in the image semantic space into the final transition video. Inspired by the image prompt methods [55], we position these image semantic embeddings $P_1, P_2, ..., P_L$ as control signals, and the video diffusion model as the decoder to leverage the generative ability of the video diffusion model. We also insert additional linear layers to project these embeddings into keys and values, involving into cross-attention of U-Net.

Formally, during the diffusion process, for each video frame feature V_i, we concatenate the text embeddings T and the predicted image semantic embeddings P_i. The cross-attention is computed as follows:

$$V_i = \text{CrossAttention}\left(V_i, \text{concat}(T, P_i), \text{concat}(T, P_i)\right). \tag{6}$$

Similar to previous video generation approaches, we optimize our model by calculating the MSE loss between L frames predicted transition video $O = (O_1, O_2, ..., O_L)$ and L frame ground truth $G = (G_1, G_2, ..., G_L)$:

$$Loss = \text{MSE}\left(G, O\right). \tag{7}$$

By encoding images into an image semantic space for integrating spatial positional relationships, our Semantic Motion Predictor could better model motion information, enabling the generation of smooth transition videos with large motion. The results and comparisons that showcase the notable improvements can be observed in Fig. 1 and Fig. 6.

4 Experiments

4.1 Implementation details

For the generation of subject-consistent images, due to the training-free and pluggable property, we implement our method on both Stable Diffusion XL [34] and Stable Diffusion 1.5 [37]. To align with the comparison models, we conduct comparisons on the Stable-XL model using the same pre-trained weights. All comparison models utilize 50-step DDIM sampling[43], and the classifier-free guidance score [18] is consistently set to 5.

For the generation of consistent videos, we implement our method based on the Stable Diffusion 1.5 pertained model and incorporate a pretrained temporal module [13] to enable video generation. All comparison models adopt a 7.5 classifier-free guidance score and 50-step DDIM sampling. Following the previous methods [12, 7], we use the Webvid10M [2] dataset to train our transition video model. For training our transition video model, we utilize the AnimateDiff V2 motion module [13] as our initial weights of the temporal module and and fine-tune the module. We then set our learning rate at 1e-4 and conduct training 100k iterations for our Semantic Motion Predictor on 8 A100

Figure 4: Comparison of consistent image generation with recent methods.

GPUs. To encode the conditional images into the image semantic space, we utilize the OpenCLIP ViT-H-14 [35, 8] pre-trained model. Our Semantic Motion Predictor incorporates 8 transformer layers, with a hidden dimension of 1024 and 12 attention heads.

4.2 Comparisons of consistent image generation

We mainly evaluate our method for generating subject-consistent images by comparing it with the two most recent ID preservation methods, IP-Adapter [55] and Photo Maker [55]. To test the performance, we use GPT-4 to generate twenty character prompts and one hundred activity prompts to describe specific activities. The format of our character prompts is "[adjective] [group or profession] [wearing clothing]" and the format of activity prompts is “[action] [location or object]”. We combine character prompts with activity prompts to obtain groups of test prompts. For each test case, we use the three comparison methods to generate a group of images that depict a person engaging in different activities to test the model’s consistency. Since IP-Adapter and PhotoMaker require an additional image to control the ID of the generated images, we first generate an image of a character to serve as the control image. We conduct both qualitative and quantitative comparisons to comprehensively evaluate the performance of these methods on consistent image generation.

Figure 5: Additional comparison of our StoryDiffusion with recent storybook generation methods, The Chosen One [1] , ConsiStory [44] and Zero-shot coherent storybook [22].

Table 1: Quantitative comparisons of consistent image generation. Our StoryDiffusion achieves better text similarity and subject similarity even without any training.

Metric	IP-Adapter [55]	Photo Maker [26]	StoryDiffusion (ours)
Text-Image Similarity	0.6129	0.6541	**0.6586**
Character Similarity	0.8802	0.8924	**0.8950**

Qualitative comparisons. The qualitative result is shown in Fig. 4. Our StoryDiffusion can generate highly consistent images, whereas other methods, IP-Adapter and PhotoMaker, may produce images with inconsistent attire or diminished text controllability. For the first example, the IP-Adapter method generates an image lost "telescope" with the text prompt "Stargazing with a telescope". PhotoMaker generates images matching the text prompt, but there are notable discrepancies in the attire across the three generated images. The third-row images generated by our StoryDiffusion exhibit consistent faces and attire with better text controllability. For the last example "A focused gamer wearing oversized headphones", IP-Adapter loses the "dog" in the second image and the "cards" in the third image. The images generated by PhotoMaker could not maintain the attire. Our StoryDiffusion still generates subject-consistent images, with the same face, and same attire, and conforms to the description in the prompt. To further demonstrate the effectiveness of our method, we compare our method with concurrent or recent storybook generation works, including The Chosen One [1], ConsiStory [44], and Zero-shot Coherent Storybook [22] in Fig. 5. Our approach not only outperforms these methods but also offers greater flexibility and faster inference times. By contrast, The Chosen One requires time-consuming LoRA self-training for each sample; Zero-shot Coherent Storybook necessitates a two-step process, first generating images and then embedding with Iterative Coherent Identity Injection; and ConsiStory involves iterative segmentation mask calculations during diffusion to maintain consistency.

Quantitative comparisons. We evaluate the quantitative comparison and show the results in Tab. 1. We evaluate two metrics, the first one is text-image similarity, which calculates the CLIP Score between the text prompts and the corresponding images. The second aspect is character similarity, measured by the CLIP Scores of character images after using the background removal method RMBG-1.4. Our StoryDiffusion achieves the best performance on both quantitative metrics, which shows our method's robustness in maintaining character meanwhile conforming to prompt descriptions.

4.3 Comparisons of transition videos generation

In transition video generation, we conduct comparisons with the two state-of-the-art methods, SparseCtrl [12] and SEINE [7], to evaluate our performance. We randomly sample around 1000 videos as the test dataset. We employ the three comparison models to predict the intermediate frames of a transition video, given the start and end frames, in order to assess their performance.

Figure 6: Comparisons of transition video generation with the recent state-of-the-art methods.

Qualitative comparisons. We conduct the qualitative comparison of transition video generation and show the results in Fig. 6. Our StoryDiffusion notably outperforms SEINE [7] and SparseCtrl [12], generating transition videos that are smooth and physically plausible. For the first example, two

Table 2: Quantitative comparisons with state-of-the-art transition video generation models.

Methods	LPIPS-*f* (↓)	LPIPS-*a* (↓)	CLIPSIM-*f* (↑)	CLIPSIM-*a* (↑)	FVD (↓)	FID(↓)
SEINE	0.4332	0.2220	0.9259	0.9736	321	140
SparseCtrl	0.4913	0.1768	0.9032	0.9756	429	181
Ours	**0.3794**	**0.1635**	**0.9606**	**0.9870**	**271**	**109**

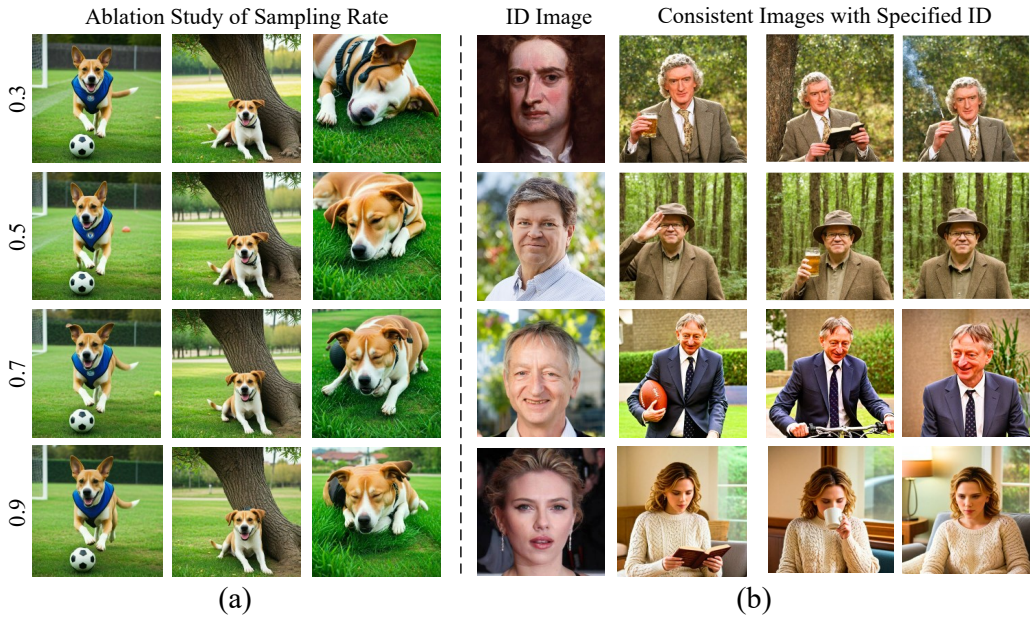

Figure 7: Ablation study. (a) Evaluations of the impact of different sampling rates in Consistent Self-Attention. (b) We explore the introduction of external control IDs to govern the generation of characters. Our StoryDiffusion can generate consistent images that conform to the ID images.

people kissing underwater, the intermediate frames generated by SEINE are corrupted, and there is a direct jump to the final frame. SparseCtrl generates results with slightly better continuity, but the intermediate frames still contain corrupted images, with numerous hands appearing. However, our StoryDiffusion succeeds in generating videos with very smooth motion without corrupted intermediate frames. For the second example, the intermediate frames generated by SEINE have corrupted arms. SparseCtrl, on the other hand, fails to maintain consistency in appearance. Our StoryDiffusion generates consistent videos with excellent continuity. For the last example, the video we generate adheres to physical spatial relationships, unlike SEINE and SparseCtrl, which only change the appearance in the transition. More visual examples can be found in the Sec. A.

Quantitative comparisons. Following previous works [12, 58], we compare our method with SEINE and SparseCtrl with four quantitative metrics, including LPIPS-*f*, LPIPS-*a*, CLIPSIM-*f*, and CLIPSIM-*a*, as shown in Tab. 2. LPIPS-*f* and CLIPSIM-*f* measure the similarities between the first frame and other frames, which reflect the overall continuity of the video. LPIPS-*a* and CLIPSIM-*a* measure the average similarities between consecutive frames, which reflect the continuity between frames. FVD and FID are also computed to evaluate generation quality. Our model outperforms the other two methods across all four quantitative metrics. These quantitative experimental results demonstrate the strong performance of our method in generating consistent and seamless transition videos.

4.4 Ablation study

User-specified ID generation. We conduct an ablation study to test the performance of generating consistent images with a user-specified ID. Since our Consistent Self-Attention is pluggable and training-free, we combine our Consistent Self-Attention with PhotoMaker, giving images to control

Table 3: Ablation study on different random sampling ratios for both random sampling and grid sampling.

Sampling Method	Rand 0.3	Rand 0.5	Rand 0.7	Grid 0.5
Character Similarity	86.39%	88.37%	89.26%	**89.29%**
CLIP Score	**57.14%**	57.11%	56.96%	56.53%

Table 4: **User study** on subject-consistent image generation and transition video generation.

Consistent Images Generation	IP-Adapter	PhotoMaker	StoryDiffusion (ours)
User Preference	20.8 %	10.9 %	**68.3 %**
Transition Video Generation	SEINE	SparseCtrl	StoryDiffusion (ours)
User Preference	5.9 %	9.6 %	**84.5 %**

the characters for consistent image generation. The results are shown in Fig. 7. With the control of the ID image, our StoryDiffusion can still generate consistent images conformed to the given control ID, which strongly indicates the scalability and plug-and-play capability of our method.

Sampling Rate of Consistent Self-Attention. Our Consistent Self-Attention approach samples tokens from other images within a batch, incorporating them into keys and values during self-attention. The original intent of random sampling is to maintain consistency while avoiding excessive structural information, thereby preventing the weakening of text control and maintaining diversity in poses. We conducted an ablation study to find the optimal sampling rate (results in Fig. 7). A sampling rate of 0.3 does not maintain subject consistency, as seen in the third column of images on the left in Fig. 7, whereas higher rates do. Quantitatively, higher sampling rates can over-correlate images and reduce text control, while lower rates weaken character consistency. We also added a quantitative comparison with the grid sampling method (Tab. 3). While grid sampling better preserves character consistency, it sacrifices text prompt controllability, making our adjustable sampling ratio ideal for balancing both factors. In practice, we set the sampling rate to 0.5, balancing consistency with minimal diffusion process impact.

4.5 User study

We conduct a user study with 79 people. Each user is assigned 50 questions to evaluate the effectiveness of our subject-consistent image generation method and transition video generation method. For subject-consistent image generation, we compare with the recent state-of-the-art methods IP-Adapter and PhotoMaker. In transition video generation, we compare with recent state-of-the-art methods SparseCtrl and SEINE. For fairness, the order of the results is randomized, and users are not informed about which generation model corresponds to each result. As shown in Tab. 4, whether for subject-consistent image generation or transition video generation, our model demonstrates an overwhelming advantage.

5 Conclusions

In this paper, we propose StoryDiffusion, a novel method that can generate consistent images in a training-free manner for storytelling and transition these consistent images into videos. Our Consistent Self-Attention builds connections among multiple images to efficiently generate images with consistent faces and clothing. We further propose the Semantic Motion Predictor to transition these images into videos and better narrate the story. We hope that our StoryDiffusion can inspire future controllable image and video generation endeavors.

Broader impact. Our StoryDiffusion can generate high-quality character-consistent pictures and videos. Certainly, similar to the previous image and video generation methods, our method may encounter some ethical issues. The generated portraits and videos may be used improperly, such as for fabricating false information. We strongly hope that the use of relevant technologies has clear responsibilities and strengthens legal and technical supervision to promote proper use.

Acknowledgement. This work is supported by the National Science Fund of China (No. 62225604) and the Fundamental Research Funds for the Central Universities (Nankai University, 070-63223049). We also acknowledge "Science and Technology Yongjiang 2035" key technology breakthrough plan project (2024Z120).

References

[1] Omri Avrahami, Amir Hertz, Yael Vinker, Moab Arar, Shlomi Fruchter, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. The chosen one: Consistent characters in text-to-image diffusion models. In *ACM SIGGRAPH 2024 conference papers*, 2024. 7, 8

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 6, 19

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint*, 2023. 4

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023. 1

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 1

[6] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *ICCV*, 2023. 2, 4

[7] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. *arXiv preprint*, 2023. 3, 4, 6, 8, 17, 22

[8] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, 2023. 7, 19

[9] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *ICLR*, 2023. 3

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint*, 2022. 3

[11] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint*, 2023. 4

[12] Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. *arXiv preprint*, 2023. 3, 4, 6, 8, 9, 17, 22

[13] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *ICLR*, 2024. 3, 6

[14] Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, et al. Animate-a-story: Storytelling with retrieval-augmented video generation. *arXiv preprint*, 2023. 4

[15] J Ho, T Salimans, A Gritsenko, W Chan, M Norouzi, and DJ Fleet. Video diffusion models. arxiv 2022. *arXiv preprint*, 2022. 1, 3

[16] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint*, 2022. 3

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3

[18] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint*, 2022. 6

[19] Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *ICCV*, 2023. 2

[20] Susung Hong, Junyoung Seo, Heeseong Shin, Sunghwan Hong, and Seungryong Kim. Direct2v: Large language models are frame-level directors for zero-shot text-to-video generation. *arXiv preprint*, 2023. 4

[21] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint*, 2023. 4

[22] Hyeonho Jeong, Gihyun Kwon, and Jong Chul Ye. Zero-shot generation of coherent storybook from plain text story using diffusion models. *arXiv preprint*, 2023. 7, 8

[23] Yuming Jiang, Shuai Yang, Tong Liang Koh, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2performer: Text-driven human video generation. *arXiv preprint*, 2023. 3

[24] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 4

[25] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 3

[26] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. *arXiv preprint*, 2023. 1, 2, 3, 8

[27] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022. 3

[28] Wan-Duo Kurt Ma, JP Lewis, W Bastiaan Kleijn, and Thomas Leung. Directed diffusion: Direct control of object placement through attention guidance. *arXiv preprint*, 2023. 3

[29] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. *arXiv preprint*, 2023. 4

[30] Jiafeng Mao and Xueting Wang. Training-free location-aware text-to-image synthesis. *arXiv preprint*, 2023. 3

[31] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint*, 2023. 3

[32] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *CVPR*, 2023. 4

[33] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint*, 2022. 3

[34] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint*, 2023. 3, 6, 19

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 7, 19

[36] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint*, 2022. 3

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3, 6

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI 2015*, 2015. 3

[39] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 3

[40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022. 3, 4

[41] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE TPAMI*, 2023. 3

[42] 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*, 2022. 3

[43] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint*, 2020. 6

[44] Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)*, 2024. 7, 8

[45] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse, attend, and segment: Unsupervised zero-shot segmentation using stable diffusion. *arXiv preprint*, 2023. 2

[46] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint*, 2023. 3

[47] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint*, 2024. 1

[48] Tan Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for referring human dance generation in real world. *arXiv preprint*, 2023. 4

[49] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint*, 2023. 3

[50] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*, 2023. 4

[51] Ruiqi Wu, Liangyu Chen, Tong Yang, Chunle Guo, Chongyi Li, and Xiangyu Zhang. Lamp: Learn a motion pattern for few-shot-based video generation. *arXiv preprint*, 2023. 4

[52] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. *arXiv preprint*, 2022. 3

[53] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. *arXiv preprint*, 2023. 4

[54] Mengjiao Yang, Yilun Du, Bo Dai, Dale Schuurmans, Joshua B Tenenbaum, and Pieter Abbeel. Probabilistic adaptation of text-to-video models. *arXiv preprint*, 2023. 3

[55] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint*, 2023. 1, 3, 6, 7, 8, 19

[56] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint*, 2023. 3

[57] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 3, 17

[58] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 9

[59] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jiawei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. *arXiv preprint*, 2023. 4

[60] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint*, 2023. 3

[61] Yupeng Zhou, Daquan Zhou, Zuo-Liang Zhu, Yaxing Wang, Qibin Hou, and Jiashi Feng. Maskdiffusion: Boosting text-to-image consistency with conditional mask. *arXiv preprint*, 2023. 3

Appendix overview

We provide an overview to clearly display the contents of the appendix, **we have also uploaded a video folder.**

- In Sec. A, we provide additional experimental results, including generated video and comics, and additional comparisons of image and video generation.
- In Sec. B, we list the limitations and the future work.
- In Sec. C, we provide the license of dataset and pre-trained model we use.

A Additional experimental results

We show video results in the upload files and more comic results in Fig. 8. We present additional comparison results here as a supplement to those discussed in the main text, as shown in Fig. 9 and Fig. 10.

A.1 Video results

We have also uploaded video files in the supplementary materials, which contain videos generated by our method. The files we have uploaded include 20 videos, which, based on our Semantic Motion Predictor, turn keyframes generated by Consistent Self-Attention or from the video into videos. The video we uploaded contains two types: longer ones with a lot of movement and the second with shorter, less active ones. Together, they show the model can handle different styles. Due to the upload size limit of 100MB, we compressed some of the video files.

A.2 Comic results

We show more comic examples in Fig. 8, which narrate two compelling stories, serving as a complement to Fig. 1. These examples collectively demonstrate the practical value of our method in artistic creation.

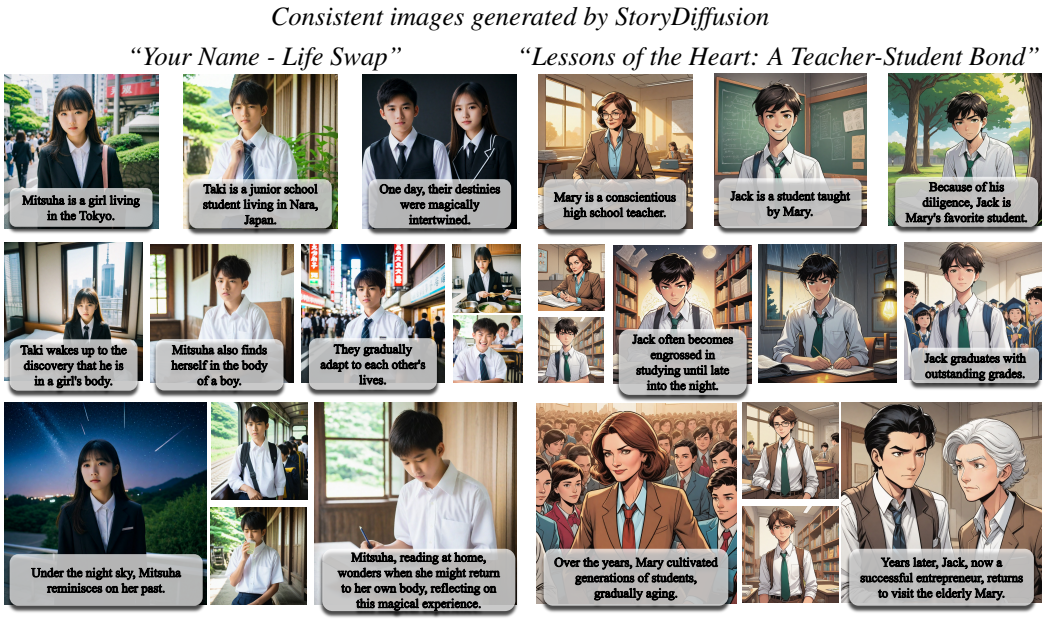

Figure 8: More comics results generated by our StoryDiffusion.

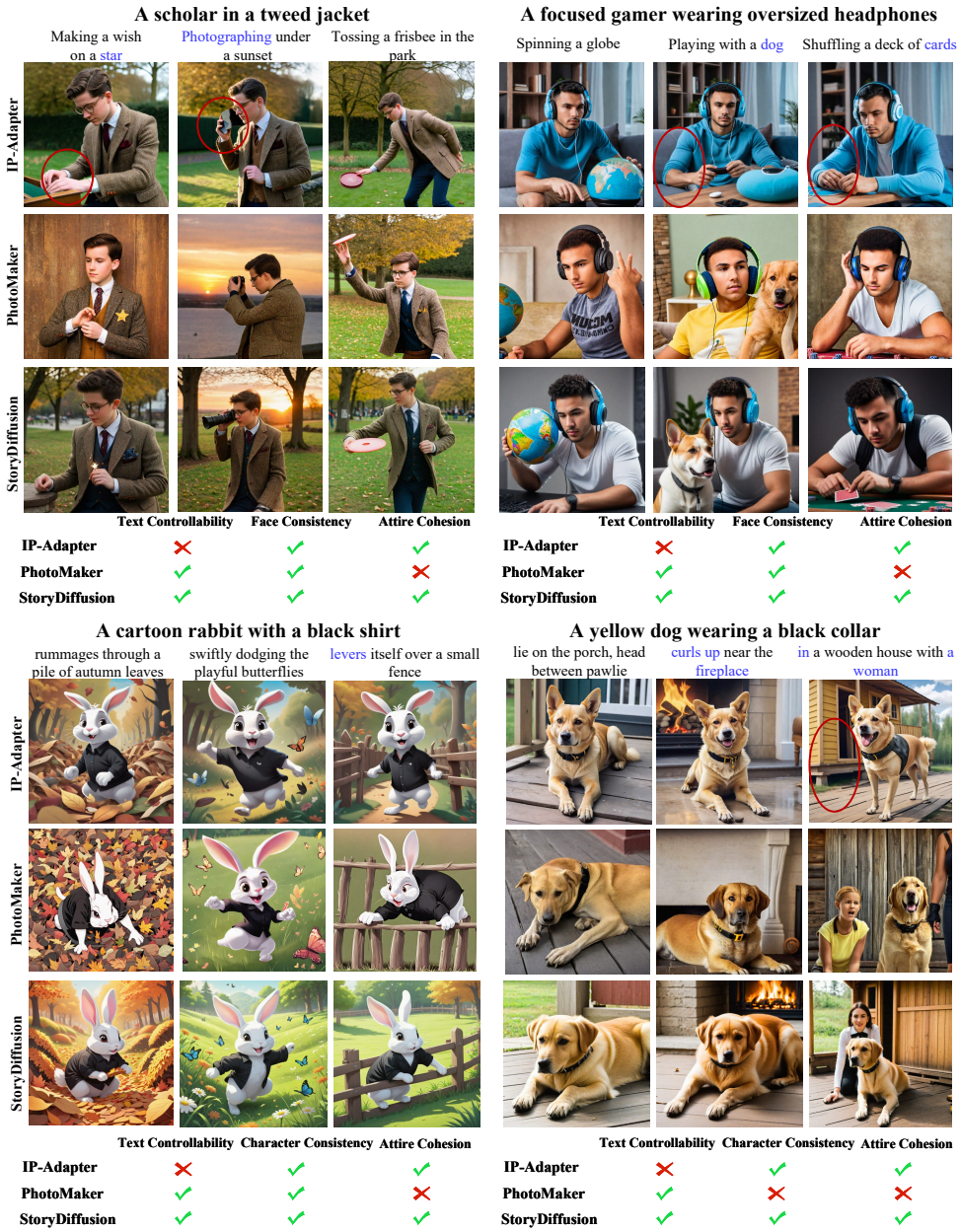

Figure 9: Additional visual comparison of consistent image generation.

A.3 Consistent image generation

We showcase additional results of consistent character image generation in Fig. 9. The results reveal better text consistency of Consistent Self-Attention than IP-Adapter. For example, in the case of 'A scholar in a tweed jacket,' IP-Adapter failed to generate a star. In 'A cartoon rabbit with a black shirt,' the action of flipping through the fence is not correctly generated in the result generated by the IP-Adapter. In 'A yellow dog wearing a black collar,' it fails to generate a woman and has incorrect positional relationships. PhotoMaker does not diminish text consistency but is unable to preserve the clothing of generated characters, as seen in the first two examples. Moreover, its performance drops with non-human characters, as illustrated by the latter two examples. Unlike them, our Consistent Self-Attention generates results with high text consistency and enhanced character coherence. The experimental results further attest to the efficacy of our Consistent Self-Attention.

A.4 Transition video generation

We present an additional comparison with SparseCtrl [12] and SEINE [7] in transition video generation. As illustrated in Fig. 10, our Semantic Motion Predictor can produce more coherent and smoother intermediate frames compared to SparseCtrl and SEINE, thus further demonstrating the advantages of Semantic Motion Predictor.

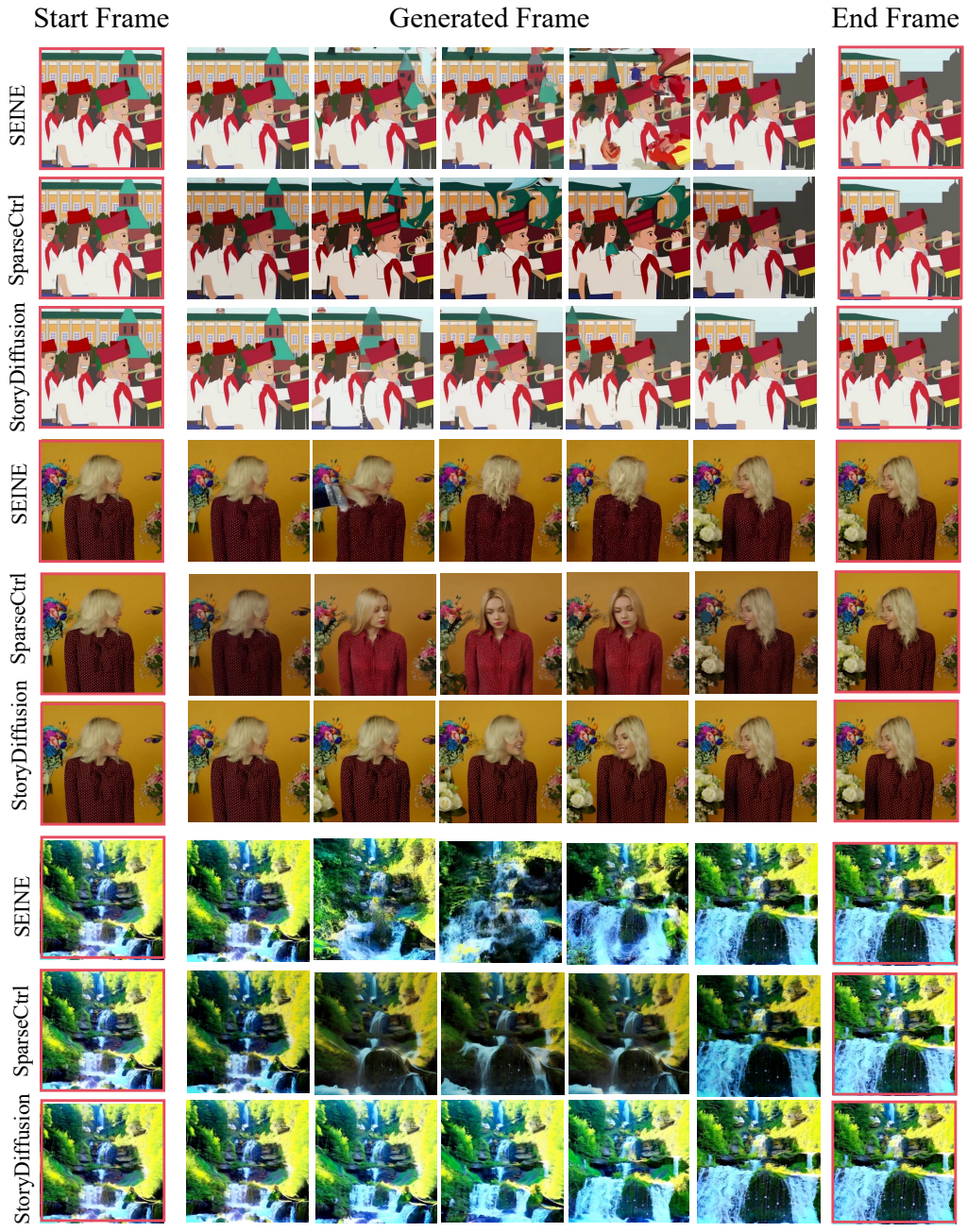

Figure 10: Additional visual comparison of transition video generation. The red box represents the frames input into the model.

A.5 Consistent images generation with ControlNet

Given that our Consistent Self-Attention is training-free and pluggable, we further explore integration with ControlNet [57] to introduce pose control in the generation of subject-consistent images. The

Figure 11: Generation results of our Consistent Self-Attention combined with ControlNet.

results of combining our Consistent Self-Attention with ControlNet are displayed in Fig. 11. Our approach is also capable of generating subject-consistent images under the guidance of ControlNet.

A.6 Experiments on Plug-and-Play capability

We have additionally implemented our StoryDiffusion on SD 1.5 and SD 2.1, and put the result compared SDXL into Fig. 12. Our method maintains good performance when integrated into different models, which demonstrates the "plug-and-play" characteristics.

StoryDiffusion based SD 1.5 StoryDiffusion based SD 2.1 StoryDiffusion based SDXL

Figure 12: The pluggable capability of our StoryDiffusion on several popular diffusion models, our Consistent Self-Attention works well across multiple models.

B Limitations and future work

The first limitation arises in our subject-consistent image generation. Similar to current state-of-the-art methods [55], there may exist inconsistencies in some minor clothing details, such as ties. In this case, our Consistent Self-Attention may require more detailed prompts to maintain consistency across images. The second limitation is in our transition video generation. Although one can utilize StoryDiffusion to generate longer videos by sequentially connecting the generated consistent images, it becomes challenging to stitch two images when there is a significant difference between them. Consequently, our method is not yet perfect for generating very long videos due to the absence of global information exchange. We will further explore long video generation in our future work.

C Dataset and model licenses

Webvid-10M: Webvid-10M [2] is a large-scale video dataset featuring 10 million video clips with associated textual descriptions, designed for training and evaluating machine learning models on video understanding and generation tasks.

URL: www.robots.ox.ac.uk/~vgg/research/frozen-in-time/

Stable XL: Stable XL [34] is a diffusion-based text-to-image generative model provided by Stability AI, which is capable of generating high-quality images based on the given text.

Licenses: CreativeML Open RAIL++-M License. URL: stability.ai/stable-image

OpenCLIP: OpenCLIP [8] is an open source implementation of OpenAI's CLIP [35] (Contrastive Language-Image Pre-training).

URL: github.com/mlfoundations/open_clip

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We confirm the main claims reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitation is discussed in Sec. B.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: We did not include theoretical results in our paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have described the model details in the implementation details in Sec. 4.1 and carefully described the experimental evaluation in the experimental chapter. The information we provide is sufficient and detailed for replication purposes.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We have made a detailed description of the implementation details to ensure that they are repeatable and use publicly available datasets. However, we intend to make our code publicly available following the paper's acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We have carried out a detailed narration in the implementation details in Sec. 4.1

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: Based on our experimental experience, the reproducibility of the experiments involved in this work is high, with results that are replicable and stable, rather than simply reporting the highest outcomes. Additionally, previous related work [7, 12] has also not reported error bars. We thus do not run the statistical significance test.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We state this detailed information of computer resources in Sec. 4.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We confirm that the research involved in the article complies with the NeurIPS Code of Ethics in all respects.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper have conducted a discussion of broader impacts at Sec. 5.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: At present, there is no relevant description we will set up safeguards when we release the model of StoryDiffusion.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We state dataset and model license in Sec. C.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not introduce new assets in the paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.