

STORYDIFFUSION: CONSISTENT SELF-ATTENTION FOR LONG-RANGE IMAGE AND VIDEO GENERATION

Yupeng Zhou¹ Daquan Zhou^{2*†} Ming-Ming Cheng^{1,3} Jiashi Feng² Qibin Hou^{1,3*†}

¹ VCIP, School of Computer Science, Nankai University

² ByteDance Inc.

³ Nankai International Advanced Research Institute (Shenzhen Futian)

*Project lead. †Corresponding authors.

Homepage: <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 new way of self-attention calculation, termed Consistent Self-Attention, that significantly boosts the consistency between the generated images and augments prevalent pre-trained diffusion-based text-to-image models in a zero-shot manner. To extend our method to long-range video generation, we further introduce a novel 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 significantly 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

Recently, diffusion models have been developed rapidly and demonstrated extraordinary potential for content generation, such as images (Rombach et al., 2022; Peebles & Xie, 2022; Podell et al., 2023), 3D objects (Zeng et al., 2022; Zhou et al., 2021) and videos (Ho et al., 2022b; Wang et al., 2023b). With extensive pre-training and advanced architectures, diffusion models show superior performance in generating very high-quality images and videos over previous generative-adversarial network (GAN) based methods (Brock et al., 2018). 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 (Ye et al., 2023) 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 (Wang et al., 2024) and PhotoMaker (Li et al., 2023a), 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 (Ho et al., 2022a; Blattmann et al., 2023b). 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.

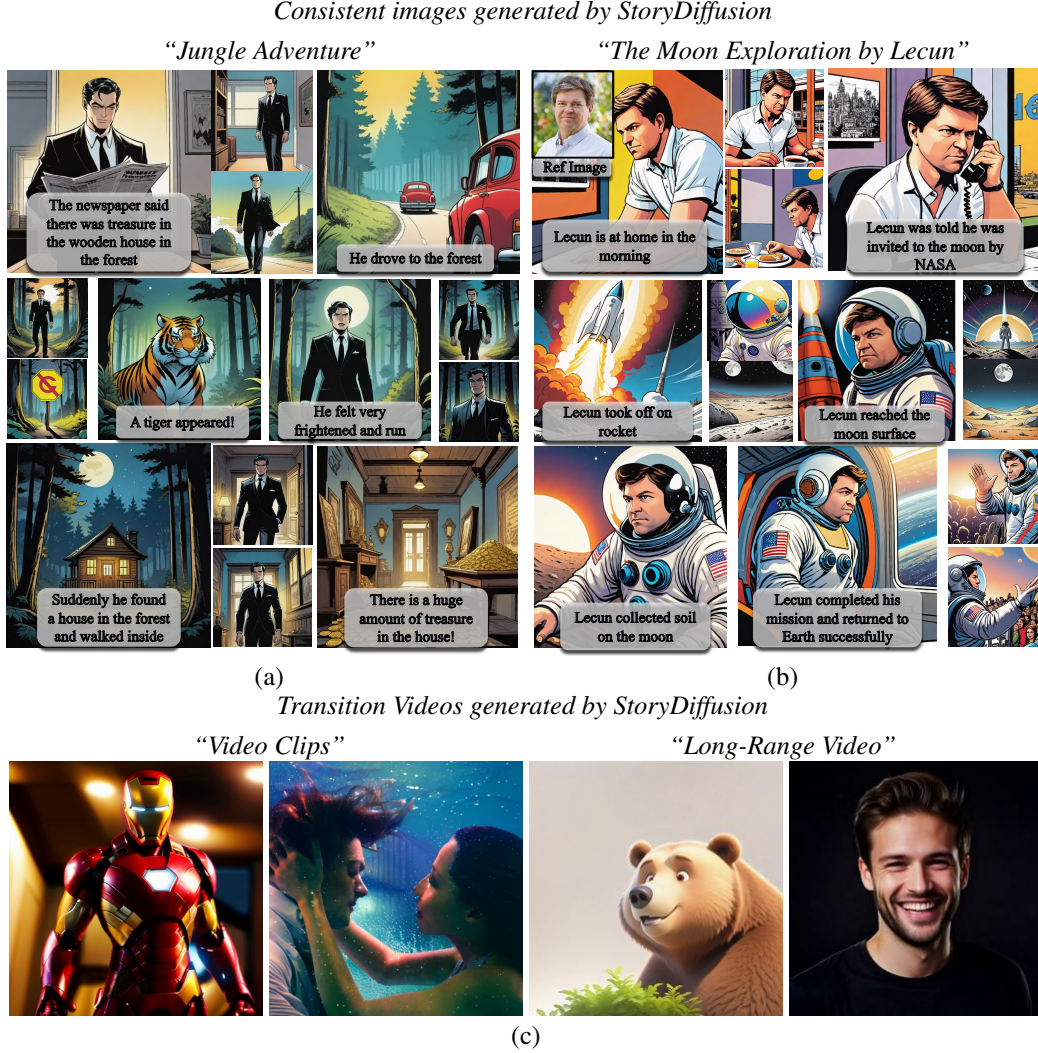


Figure 1: Images and videos generated by our StoryDiffusion. (a) Comic generated by StoryDiffusion telling the story of a man who discovers a treasure while exploring the jungle. (b) Comic generated by StoryDiffusion describing the expedition to the moon by Lecun, with a reference image control (Li et al., 2023a) same as Fig. 6(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 our home project: <https://StoryDiffusion.github.io>.

As evidenced by previous works (Tian et al., 2023; Hong et al., 2023), self-attention is one of the most important modules for modeling the overall structure of the generated visual content. Our main motivation is that if we could use a reference image to guide the self-attention calculation, the consistency between the two images is supposed to be improved significantly. 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 (as shown in Fig. 2(d)), Consistent Self-Attention incorporates sampled reference tokens from the reference images during the token similarity calculation and token merging. The sampled tokens share 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 along the temporal dimension. 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 (Guo et al., 2024), Semantic Motion Predictor can generate smooth video frames that are significantly better than recent conditional video generation methods, such as SEINE (Chen et al., 2023) and SparseCtrl (Guo et al., 2023).

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 significantly more stable long-range video frames that can be easily upscaled to minutes than recent popular image conditioning methods, such as SEINE (Chen et al., 2023) and SparseCtrl (Guo et al., 2023).
- 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 DIFFUSION MODELS

Diffusion models have rapidly demonstrated their stunning capabilities in generating realistic images and this also enables them to dominate the field of generative modeling in recent years (Rombach et al., 2022; Saharia et al., 2022; Ho et al., 2020; Ramesh et al., 2022; Peebles & Xie, 2022). By utilizing a deep denoising network (Ronneberger et al., 2015), diffusion models establish a connection between the noise distribution and the real image distribution through iterative noise addition and denoising. Early works (Ho et al., 2020; Song et al., 2021; Sohl-Dickstein et al., 2015) establish the theoretical foundation of diffusion models mainly focusing on unconditional image generation. Later, various efforts are made to enhance the efficiency and performance of diffusion models. Typical examples should be efficient sampling methods (Song et al., 2020; Zhang & Chen, 2023; Lu et al., 2022), denoising in the latent space (Rombach et al., 2022), controllability (Feng et al., 2023a; Zhou et al., 2023b), diffusion backbones (Ramesh et al., 2022; Peebles & Xie, 2022). Concurrently with the exploration of the foundational theory, diffusion models gradually gain popularity and demonstrate strong performance across various domains, such as image generation (Zhou et al., 2023b), video generation (Zhou et al., 2023a), 3D generation (Zeng et al., 2022; Zhou et al., 2021), image segmentation (Li et al., 2023b; Amit et al., 2021) and low-level vision tasks (Wang et al., 2023a; Lugmayr et al., 2022; Xie et al., 2022; Saharia et al., 2023).

2.2 CONTROLLABLE TEXT-TO-IMAGE GENERATION

As an important sub-field of diffusion model applications, text-to-image generation, represented by Latent Diffusion (Rombach et al., 2022), DiT (Peebles & Xie, 2022), and Stable XL (Podell et al., 2023), have 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 (Zhang et al., 2023b) and T2I-Adapter (Mou et al., 2023) introduce control conditions, such as depth maps,

pose images, or sketches, to direct the generation of images. MaskDiffusion (Zhou et al., 2023b) and StructureDiffusion (Feng et al., 2023b) focus on enhancing the text controllability. There are also some works (Mao & Wang, 2023; Ma et al., 2023a) 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 (Gal et al., 2022), DreamBooth (Ruiz et al., 2023), and Custom Diffusion (Kumari et al., 2023). The other one, exemplified by IPAdapter (Ye et al., 2023) and PhotoMaker (Li et al., 2023a), 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.

2.3 VIDEO GENERATION

Due to the success of diffusion models in the field of image generation (Rombach et al., 2022; Ho et al., 2020), the exploration in the domain of video generation is also becoming popular. As text is the most intuitive descriptor users can specify, text-based video generation has attracted the most attention (Guo et al., 2024; Jiang et al., 2023; Singer et al., 2022; Wang et al., 2023d; Yang et al., 2023). VDM (Ho et al., 2022a) is among the first that extend the 2D U-Net from image diffusion models to a 3D U-Net to achieve video generation. Due to the significant increase of computational cost for video generation, later works, such as MagicVideo (Zhou et al., 2023a) and Mindscope (Wang et al., 2023b), introduce 1D temporal attention mechanisms, reducing computations by building upon latent diffusion models. Following Imagen, Imagen Video (Ho et al., 2022b) employs a cascaded sampling pipeline that generates videos through multiple stages. Show-1 (Zhang et al., 2023a) also proposes a multi-stage approach to balance the generation quality and efficiency.

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 controls, such as depth maps (Guo et al., 2023; He et al., 2023), pose maps (Xu et al., 2023; Hu et al., 2023; Wang et al., 2023c; Ma et al., 2023b), RGB images (Blattmann et al., 2023a; Chen et al., 2023; Ni et al., 2023), or other guided motion videos (Zhao et al., 2023; Wu et al., 2023). Different from the ambiguity of the text prompt, introducing this conditional information enhances the controllability of video generation.

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 (Chen et al., 2023) and SparseCtrl (Chen et al., 2023). 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.

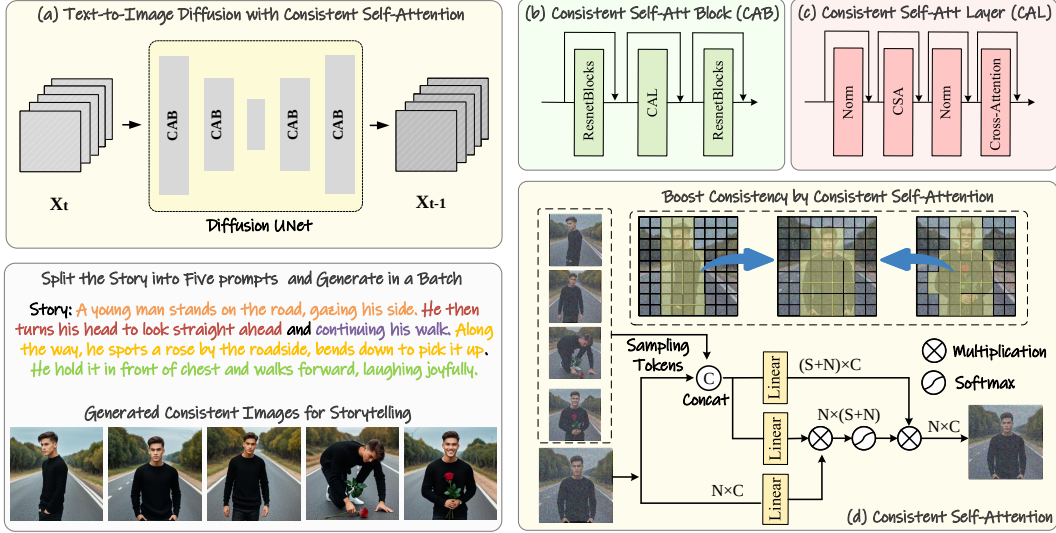


Figure 2: The Pipeline of StoryDiffusion to generating subject-consistent images. To create subject-consistent images to describe a story, we incorporate our Consistent Self-Attention into the pre-trained text-to-image diffusion model. We split a story text into several prompts and generate images using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

3.1 TRAINING-FREE CONSISTENT IMAGES GENERATION

In this subsection, we introduce how our method generates subject-consistent images in a training-free manner. 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. After revisiting the role of the different attention mechanisms within diffusion models, we get inspired to explore utilizing self-attention to serve the consistency of images within a batch and propose our Consistent Self-Attention. We insert Consistent Self-Attention 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 to remain training-free and pluggable.

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 $\text{Attention}(X_k, X_q, X_v)$ 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:

$$O_i = \text{Attention}(Q_i, K_i, V_i). \quad (1)$$

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:

$$S_i = \text{RandSample}(I_1, I_2, \dots, I_{i-1}, I_{i+1}, \dots, I_{B-1}, I_B), \quad (2)$$

where 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_{P_i} and value V_{P_i} for Consistent Self-Attention. Here, the original query Q_i is not changed. Finally, we compute the self-attention as follows:

$$O_i = \text{Attention}(Q_i, K_{P_i}, V_{P_i}). \quad (3)$$

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

Algorithm 1 Consistent Self-Attention

```

def ConsistentSelfAttention(images_features, sampling_rate, tile_size):
    """
    images_tokens: [B, C, N]
    sampling_rate: Float (0-1)
    tile_size: Int
    """
    output = zeros(B, N, C), count = zeros(B, N, C), W = tile_size
    for t in range(0, N - tile_size + 1):
        # Use tile to override out of GPU memory
        tile_features = images_tokens[t:t + W, :, :]
        reshape_featrue = tile_feature.reshape(1, W*N, C).repeat(W, 1, 1)
        sampled_tokens = RandSample(reshape_featrue, rate=sampling_rate, dim=1)

        # Concat the tokens from other images with the original tokens
        token_KV = concat([sampled_tokens, tile_features], dim=1)
        token_Q = tile_features

        # perform attention calculation:
        X_q, X_k, X_v = Linear_q(token_Q), Linear_k(token_KV), Linear_v(token_KV)
        output[t:t+W, :, :] += Attention(X_q, X_k, X_v)
        count[t:t+W, :, :] += 1

    output = output/count
    return output

```

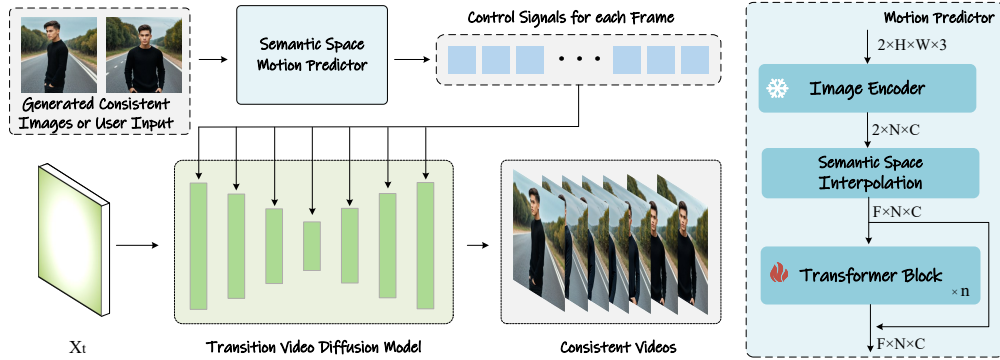


Figure 3: The pipeline of our method for generating transition videos for obtaining subject-consistent images, as described in Sec. 3.1. To effectively model the character’s large motions, we encode the conditional images into the image semantic space for encoding spatial information and predict the 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.

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 1.

3.2 SEMANTIC MOTION PREDICTOR FOR VIDEO GENERATION

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 (Guo et al., 2023) and SEINE (Chen et al., 2023), 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 .

$$K_s, K_e = E(F_s, F_e). \quad (4)$$

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 $K_1, K_2, \dots, K_L$, where L is the required video length. Then, the sequence $K_1, K_2, \dots, K_L$ is sent into a series of transformer blocks B to predict the transition frames:

$$P_1, P_2, \dots, P_L = B(K_1, K_2, \dots, K_L). \quad (5)$$

Next, we need to decode these predicted frames in the image semantic space into the final transition video. Inspired by the image prompt methods (Ye et al., 2023), we position these image semantic embeddings $P_1, P_2, \dots, P_L$ 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:

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

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

$$\text{Loss} = \text{MSE}(G, O). \quad (7)$$

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 significant improvements can be observed in Fig. 1 and Fig. 5.

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 and Stable Diffusion 1.5. 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, and the classifier-free guidance score is consistently set to 5.0.

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 (Guo et al., 2024) to enable video generation. All comparison models adopt a 7.5 classifier-free guidance score and 50-step DDIM sampling. Following the previous methods (Guo et al., 2023; Chen et al., 2023), we use the Webvid10M (Bain et al., 2021) dataset to train our transition video model. More details can be found in the supplementary materials.

4.2 COMPARISONS OF CONSISTENT IMAGES GENERATION

We evaluate our method for generating subject-consistent images by comparing it with the two most recent ID preservation methods, IP-Adapter (Ye et al., 2023) and Photo Maker (Ye et al., 2023). To test the performance, we use GPT-4 to generate twenty character prompts and one hundred activity

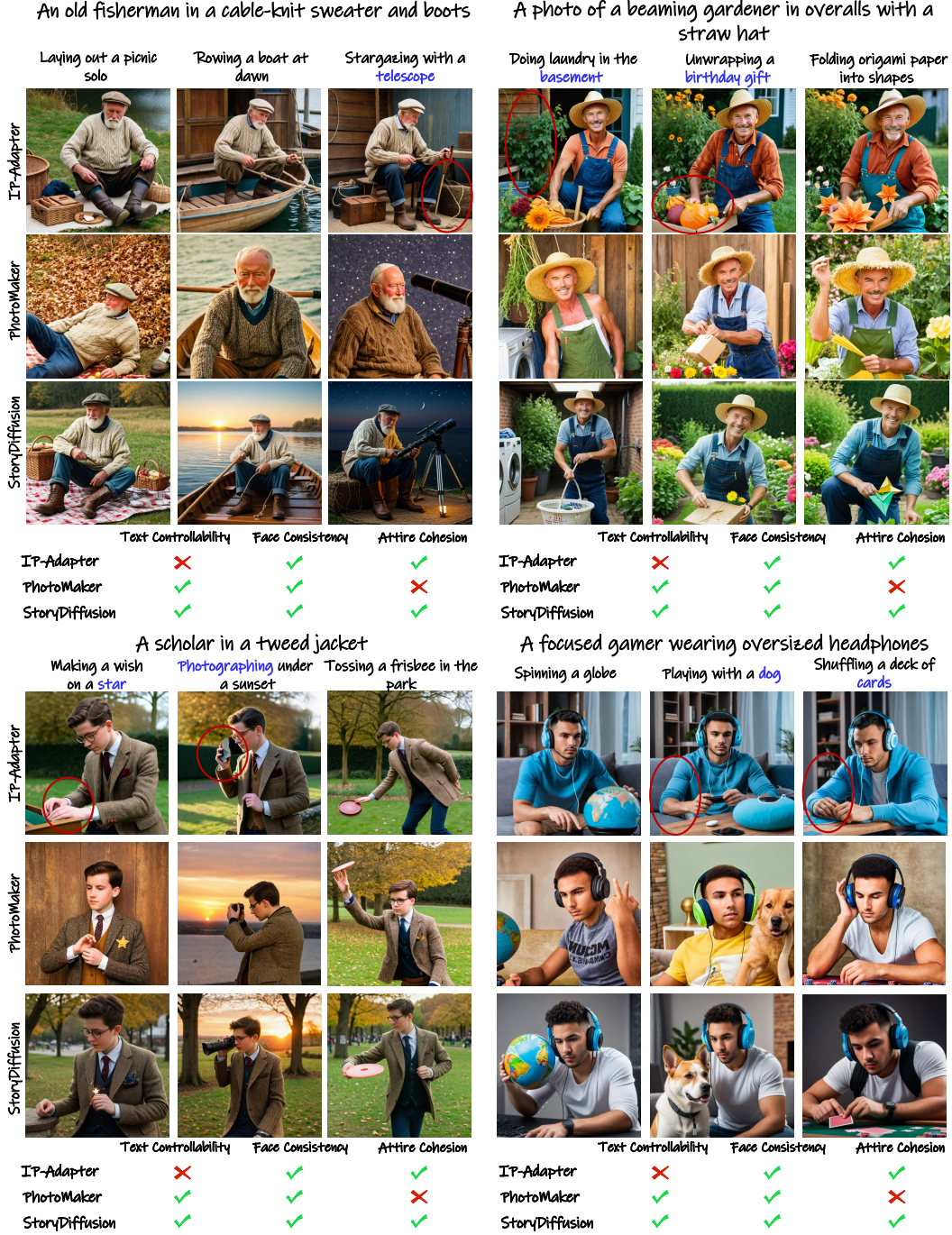


Figure 4: Comparison of consistent image generation with recent methods.

prompts to describe specific activities. 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.

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 im-

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

Metric	IP-Adapter (Ye et al., 2023)	Photo Maker (Li et al., 2023a)	StoryDiffusion (ours)
Text-Image Similarity	0.6129	0.6541	0.6586
Character Similarity	0.8802	0.8924	0.8950

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

Methods	LPIPS- <i>first</i> ($\downarrow$)	LPIPS- <i>frames</i> ($\downarrow$)	CLIPSIM- <i>first</i> ($\uparrow$)	CLIPSIM- <i>frames</i> ($\uparrow$)
SEINE	0.4332	0.2220	0.9259	0.9736
SparseCtrl	0.4913	0.1768	0.9032	0.9756
Ours	0.3794	0.1635	0.9606	0.9870

ages 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 significant 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 lose 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.

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 is character similarity, which measures the CLIP Scores of the character images. 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 (Guo et al., 2023) and SEINE (Chen et al., 2023), 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.

Qualitative Comparisons. We conduct the qualitative comparison of transition video generation and show the results in Fig. 5. Our StoryDiffusion significantly outperforms SEINE (Chen et al., 2023) and SparseCtrl (Guo et al., 2023), generating transition videos that are smooth and physically plausible. For the first example, two 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 supplementary material.

Quantitative Comparisons. Following previous works (Guo et al., 2023; Zhang et al., 2018), we compare our method with SEINE and SparseCtrl with four quantitative metrics, including LPIPS-*first*, LPIPS-*frames*, CLIPSIM-*first*, and CLIPSIM-*frames*, as shown in Tab. 2. LPIPS-*first* and CLIPSIM-*first* measure the similarities between the first frame and other frames, which reflect the

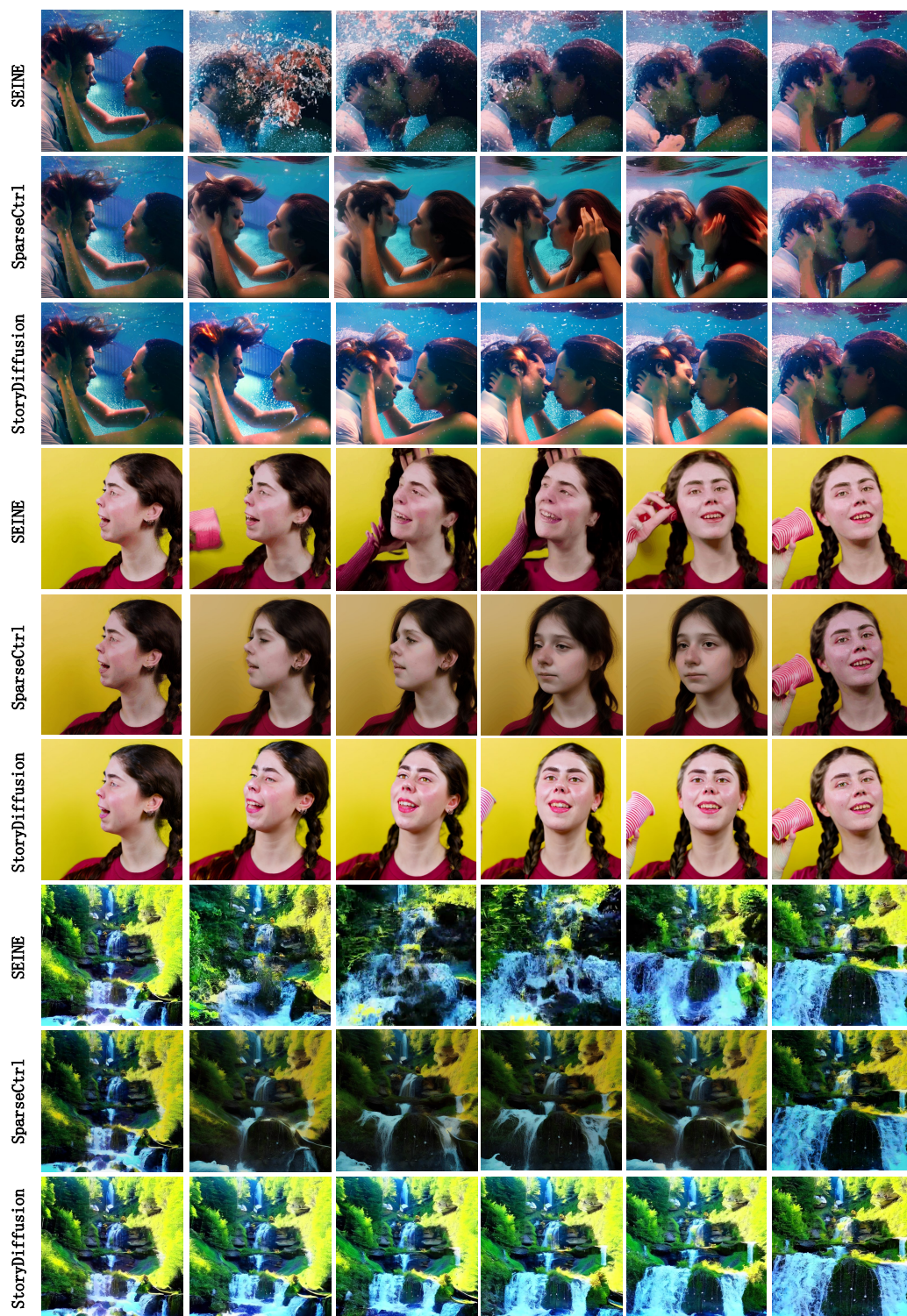


Figure 5: Comparisons of transition video generation with the recent state-of-the-art methods.

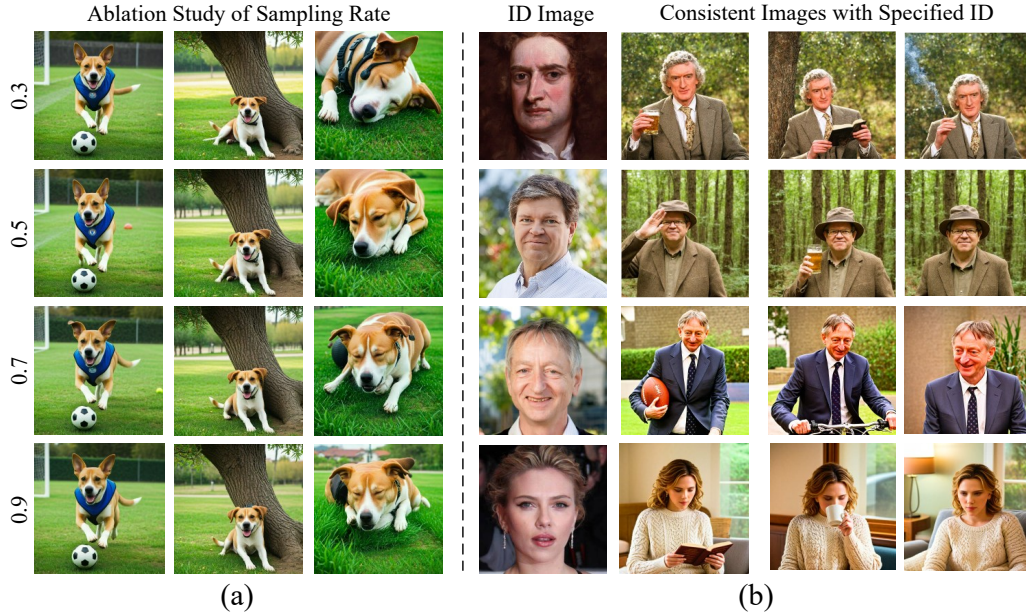


Figure 6: Ablation study. (a) Evaluations of the impact of different sampling rates in Consistent Self-Attention. (b) We explore the introduction of external control IDs to govern the generation of characters. Our StoryDiffusion can generate consistent images that conform to the ID images.

Table 3: **User Study** on subject-consistent image generation and transition video generation.

Consistent Images Generation	IP-Adapter	PhotoMaker	StoryDiffusion (ours)
User Preference	10.4 %	16.8 %	72.8 %
Transition Video Generation	SEINE	SparseCtrl	StoryDiffusion (ours)
User Preference	11.6 %	6.4 %	82 %

overall continuity of the video. *LPIPS-frames* and *CLIPSIM-frames* measure the average similarities between consecutive frames, which reflect the continuity between frames. 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 the characters for consistent image generation. The results are shown in Fig. 6. 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 sample tokens from other images within a batch and incorporate them into the keys and values during self-attention computation. To determine the optimal sampling rate, we conduct an ablation study on the sampling rate of Consistent Self-Attention. The results are also shown in Fig. 6. We found a sampling rate of 0.3 could not maintain subject consistency as seen in the third column of images in the left part of Fig. 6, whereas higher sampling rates successfully preserve it. In practice, we set the sampling rate to 0.5 by default to make a minimal impact on the diffusion process and maintain consistency.

4.5 USER STUDY

We conduct a user study with 30 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. The experimental results of our user study are shown in Tab. 3. Whether for subject-consistent image generation or transition video generation, our model demonstrates an overwhelming advantage. The user study further confirms the superior performance of our StoryDiffusion.

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.

ACKNOWLEDGMENTS

This research was supported by NSFC (NO. 62225604, No. 62276145), the Fundamental Research Funds for the Central Universities (Nankai University, 070-63223049), CAST through Young Elite Scientist Sponsorship Program (No. YESS20210377). Computations were supported by the Supercomputing Center of Nankai University (NKSC).

REFERENCES

- Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint*, 2021. 3
- 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. 7
- 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*, 2023a. 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*, 2023b. 1
- Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 1
- 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, 7, 9
- 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023. 16
- Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *ICLR*, 2023a. 3

-
- 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*, 2023b. 4
- Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint*, 2022. 4
- 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, 7, 9
- 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, 4, 7, 16
- 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
- J Ho, T Salimans, A Gritsenko, W Chan, M Norouzi, and DJ Fleet. Video diffusion models. *arxiv* 2022. *arXiv preprint*, 2022a. 1, 4
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3, 4
- 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*, 2022b. 1, 4
- Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *ICCV*, 2023. 2
- 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
- Yuming Jiang, Shuai Yang, Tong Liang Koh, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2performer: Text-driven human video generation. *arXiv preprint*, 2023. 4
- Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 4
- 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*, 2023a. 1, 2, 4, 9
- Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Open-vocabulary object segmentation with diffusion models. In *ICCV*, 2023b. 3
- Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *NeurIPS*, 2022. 3
- 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
- Wan-Duo Kurt Ma, JP Lewis, W Bastiaan Kleijn, and Thomas Leung. Directed diffusion: Direct control of object placement through attention guidance. *arXiv preprint*, 2023a. 4
- 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*, 2023b. 4
- Jiafeng Mao and Xueting Wang. Training-free location-aware text-to-image synthesis. *arXiv preprint*, 2023. 4

-
- 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
- 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
- William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint*, 2022. 1, 3
- 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. 1, 3
- 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. 16
- Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint*, 2022. 3
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 3, 4
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI 2015*, 2015. 3
- 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. 4
- 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
- Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE TPAMI*, 2023. 3
- 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. 4
- Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 3
- Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint*, 2020. 3
- Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *ICLR*, 2021. 3
- 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
- Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *arXiv preprint*, 2023a. 3
- Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint*, 2023b. 1, 4
- Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint*, 2024. 1

-
- 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*, 2023c. 4
- 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*, 2023d. 4
- 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
- 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
- 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
- Mengjiao Yang, Yilun Du, Bo Dai, Dale Schuurmans, Joshua B Tenenbaum, and Pieter Abbeel. Probabilistic adaptation of text-to-video models. *arXiv preprint*, 2023. 4
- Hu Ye, Jun Zhang, Sibio Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint*, 2023. 1, 4, 7, 9, 16
- Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. In *NeurIPS*, 2022. 1, 3
- 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*, 2023a. 4
- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023b. 3, 16
- Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. *ICLR*, 2023. 3
- 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
- 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
- Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint*, 2023a. 3, 4
- Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *ICCV*, 2021. 1, 3
- 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*, 2023b. 3, 4

A ADDITIONAL RESULTS

A.1 TRANSITION VIDEOS GENERATION

To further showcase the transition video generation capabilities of our StoryDiffusion, additional videos created by our method are presented in Fig. 7. Our StoryDiffusion is capable of producing high-quality and smooth transition videos as shown in the additional results.

“More Transition Videos generated by StoryDiffusion”

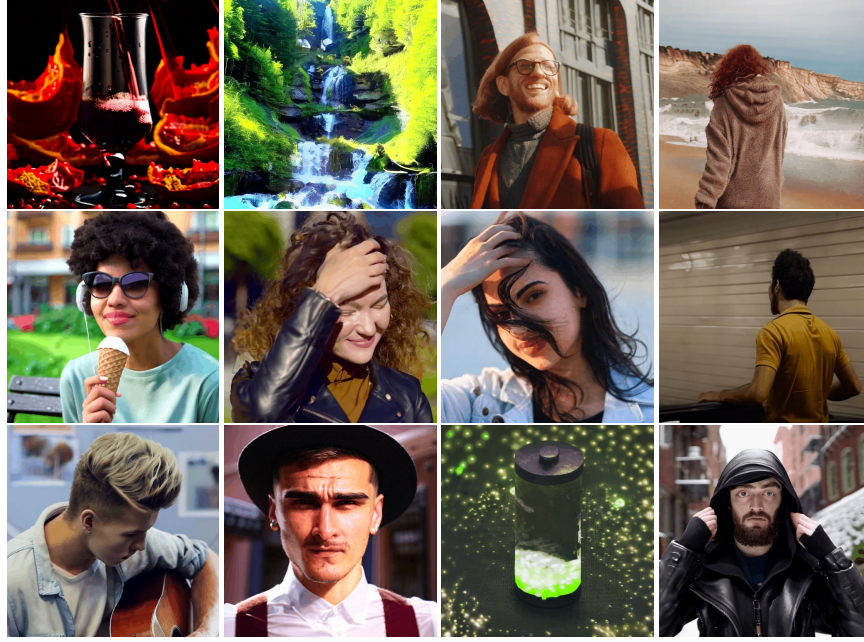


Figure 7: More video results generated by our StoryDiffusion. Best viewed with *Acrobat Reader*. Click the image to play the video.

A.2 CONSISTENT IMAGES GENERATION WITH CONTROLNET

Given that our Consistent Self-Attention is training-free and pluggable, we further explore integration with ControlNet (Zhang et al., 2023b) to introduce pose control in the generation of subject-consistent images. The results of combining our Consistent Self-Attention with ControlNet are displayed in Fig. 8. Our approach is also capable of generating subject-consistent images under the guidance of ControlNet.

B IMPLEMENTATION DETAILS

For training our transition video model, we utilize the AnimateDiff V2 motion module (Guo et al., 2024) as our initial weights of the temporal module. We set our learning rate at $1e-4$ and conduct training 100k iterations on 8 A100 GPUs. To encode the conditional images into the image semantic space, we utilize the OpenCLIP ViT-H-14 (Radford et al., 2021; Cherti et al., 2023) pre-trained model. Our Semantic Motion Predictor incorporates 8 transformer layers, with a hidden dimension of 1024 and 12 attention heads.

C LIMITATIONS

The first limitation arises in our subject-consistent image generation. Similar to current state-of-the-art methods (Ye et al., 2023), 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 using a sliding window approach, our method is not designed specifically for long video generation. 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.



Figure 8: Generation results of our Consistent Self-Attention combined with ControlNet.