Storynizor: Consistent Story Generation via Inter-Frame Synchronized and Shuffled ID Injection

Yuhang Ma[1*], Wenting Xu[1*], Chaoyi Zhao[1*], Keqiang Sun[2], Qinfeng Jin[1], Xiaoda Yang[3], Zeng Zhao[1†], Changjie Fan[1], Zhipeng Hu[1]

[1]Fuxi AI Lab, Netease Inc.
[2]Multimedia Laboratory, The Chinese University of Hong Kong
[3]School of Software Technology, Zhejiang University
{mayuhang, xuwenting01, zhaochaoyi, jinqinfeng, hzzhaozeng, hzliubai, fanchangjie, zphu}@corp.netease.com

Abstract

Recent advances in text-to-image diffusion models have spurred significant interest in continuous story image generation. In this paper, we introduce **Storynizor**, a model capable of generating coherent stories with strong inter-frame character consistency, effective foreground-background separation, and diverse pose variation. The core innovation of Storynizor lies in its key modules: **ID-Synchronizer** and **ID-Injector**. The ID-Synchronizer employs an *auto-mask self-attention* module and a *mask perceptual loss* across inter-frame images to improve the consistency of character generation, vividly representing their postures and backgrounds. The ID-Injector utilize a *Shuffling Reference Strategy (SRS)* to integrate ID features into specific locations, enhancing ID-based consistent character generation. Additionally, to facilitate the training of Storynizor, we have curated a novel dataset called **StoryDB** comprising 100,000 images. This dataset contains single and multiple-character sets in diverse environments, layouts, and gestures with detailed descriptions. Experimental results indicate that Storynizor demonstrates superior coherent story generation with high-fidelity character consistency, flexible postures, and vivid backgrounds compared to other character-specific methods.

Introduction

Recent advancements in text-to-image diffusion models has sparked considerable interest in generating continuous story images. Maintaining consistency between frames, ensuring natural and flexible character poses, and achieving a clear separation of foreground and background are critical challenges in this domain.

Many prior works have paid attention to ensuring character consistency. For instance, IP-Adapter (Ye et al. 2023), Arc2Face (Papantoniou et al. 2024), and InstantID (Wang et al. 2024) extract identity features from a reference image and inject them into the diffusion model. While effective in single-character scenarios, these methods often struggle with stiff postures and are limited in handling more complex multicharacter interactions.

Figure 1: Comparison of Storynizor with existing methods. Storynizor shows superior performance when implemented in the original SD-base checkpoint in text-image alignment and inter-frame consistency.

Other approaches, such as Mix-of-Show (Gu et al. 2024) and OMG (Kong et al. 2024), focus on multi-character generation by utilizing attention maps to position characters within a frame. These methods successfully achieve varied poses and maintain character consistency but lack inter-frame coherence, as they operate on a frame-by-frame basis without ensuring consistency across the sequence.

To achieve narrative coherence, methods like ConsiStory (Tewel et al. 2024) and StoryDiffusion (Zhou et al. 2024) have attempted to fuse character features across frames to enhance inter-frame consistency. However, the absence of an identity injection mechanism in these approaches results in inaccurate alignment with reference images. Moreover, when considering a pre-trained diffusion model like the original checkpoint of SD1.5, their training-free nature often leads to semantic degradation, and collapsible cross-frame results, as illustrated in Fig. 1.

As shown in Tab. 1, prior works have focused on specific aspects of generating continuous story images, but none of them has comprehensively addressed all key challenges.

In this paper, we introduce Storynizor, the first model capable of generating multicharacter stories with high

*These authors contributed equally.
†Corresponding Authors.

	IDC	FHP	MS	IFC	FBD
IP-Adapter (Ye et al. 2023)	✓	✗	✗	✗	✗
InctantID (Wang et al. 2024)	✓	✗	✗	✗	✗
OMG (Kong et al. 2024)	✓	✓	✓	✗	✗
ConsiStory (Tewel et al. 2024)	✗	✓	✓	✓	✗
StoryDiffusion (Zhou et al. 2024)	✗	✓	✓	✓	✗
FastComposer (Xiao et al. 2023)	✗	✓	✓	✗	✓
Storynizor (ours)	✓	✓	✓	✓	✓

Table 1: Comparison between our proposed Storynizor and state-of-the-art character-specific methods. in ID consistency (IDC), flexible human pose (FHP), multi-subject (MS), iner-frame consistency (IFC) and F/B disentanglement (FBD).

inter-frame character consistency, effective foreground-background separation, and rich pose variation.

As shown in Fig. 2, given arbitrary numbers of reference images and several text prompts from a story, our Storynizor generate corresponding story images, with consistent character identity, vivid character postures and maintaining high consistency across frames.

The core innovation of Storynizor lies in key modules: the ID-Synchronizer to ensure that identity features are consistently maintained across frames and the ID-Injector to introduce ID-specific features from reference images.

Specifically, our approach builds upon the UNet architecture, where the ID-Synchronizer, composed of Auto-mask Space-Attention (AMSA) trained with the Mask Perceptual Loss, plays a crucial role in preventing the attention mask leakage and enhancing the consistency of characters throughout the sequence of frames.

In parallel, the ID-Injector extracts essential features from reference characters and integrates them into specific locations within the network. To make sure the ID-Injector learns the identity information from the reference character images without simply replicating the image feature from the reference image, we introduce a Shuffling Reference Strategy (SRS). Concretely, we randomly sample pairs of reference and ground-truth images from the same character set, with variations in layout, scenarios, and gestures. This strategy significantly boosts the generalization of the model and maintain consistency across diverse poses and environments, leading to notable improvements in performance.

To train Storynizor effectively and support the Shuffling Reference Strategy (SRS), we further curated a novel dataset, called StoryDB, by selecting multiple sets of characters and collecting images of each character set in various environments, layouts, and gestures. This diverse and carefully structured dataset allows the model to maintain identity consistency while performing different actions in diverse scenarios. Tab. 2 shows the comparison between StoryDB and other existing datasets, indicating that none of existing datasets contains the same level of diversity and multi-contextual character interactions as StoryDB.

The contributions of this paper are four folded:

- We **introduce Storynizor**, the first model capable of generating multi-character stories with high inter-frame character consistency, effective foreground-background separation, and rich pose variation.
- We develop two key modules—**ID-Injector and ID-Synchronizer**—integrated into a UNet-based architecture, ensuring consistent character identity and posture across sequential frames.
- We curate **a novel dataset named StoryDB** featuring multiple character sets in various environments, layouts, and gestures, enabling the model to maintain identity consistency across different scenarios and actions.

	MC	IDC	CC	TD	VB	VP	GI
Deepfashion(Liu et al. 2016)	✗	✗	✓	✗	✗	✗	✓
VITON(Han et al. 2018)	✗	✓	✗	✗	✗	✗	✗
StreetTryOn(Cui et al. 2024)	✗	✗	✓	✗	✓	✗	✓
StoryDB (ours)	✓	✓	✓	✓	✓	✓	✓

Table 2: Comparison between our proposed StoryDB and other related datasets in Multi-characters (MC), ID Consistency (IDC), Clothes Consistency (CC), Text Description (TD), Various Background (VB), Various Pose (VP), Group Images (GI).

Related Work

Diffusion Models. Models like DALL-E2 (Ramesh et al. 2022) and Imagen (Saharia et al. 2022) required high resources, while LDMs (Rombach et al. 2022) reduced this by using latent space. However, diffusion models still lack consistent character generation.

Consistent Character Generation. Consistent character generation in diffusion models ensures coherence across images. Early methods like LoRA (Hu et al. 2021) required fine-tuning, while training-free methods like IP-Adapter (Ye et al. 2023) and InstantID (Wang et al. 2024) avoid it. However, image-conditioned methods struggle with balancing consistency and text-image alignment in story generation.

Continious Story Generation. Continuous story generation ensures character consistency (Liu et al. 2024; Zhou et al. 2024; Tewel et al. 2024; Ma et al. 2024). Inspired by diffusion models in image, video (Guo et al. 2023), and 3D generation (Shi et al. 2023), methods like ConsiStory (Tewel et al. 2024) and StoryDiffusion (Zhou et al. 2024) use cross-frame attention but struggle with semantic accuracy, causing inconsistencies and misalignment.

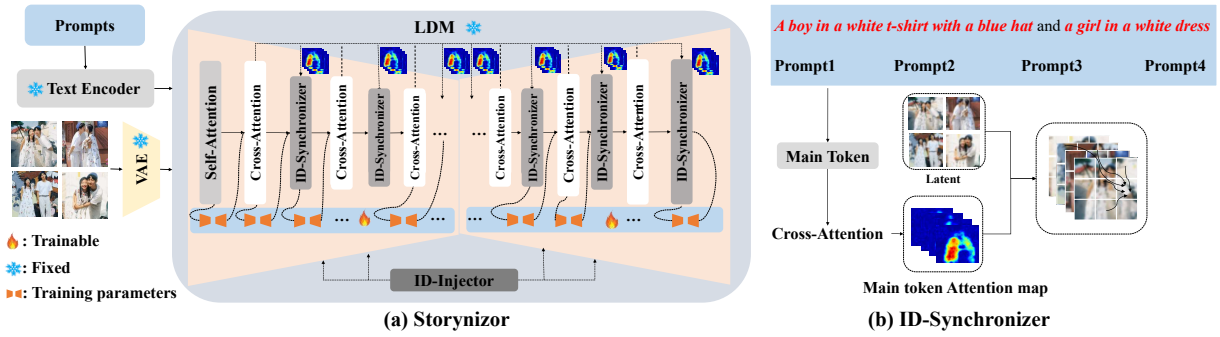
Method

We propose a pretrained story generation model called Storynizor, which generates multi-character stories with high inter-frame character consistency, effective foreground background separation, and rich pose variation under a series of prompt conditioning and ID images (optional). To modeling our task, the prompt $\mathcal{T}$ is set as the following:

$$\mathcal{T} = \{\mathcal{T}_n\}, n = 1, ..., N \tag{1}$$

where N denotes the total numbers of prompts. $\mathcal{T}_n$ contains the description of characters P and the actions of each character$A$:

$$\mathcal{T}_n = \{P_n, A_n\} = \{P_n^m, A_n^m\}, m = 1, ..., M, \tag{2}$$



(a) Storynizor (b) ID-Synchronizer

Figure 2: Overview of our proposed (a) Storynizor. Storynizor mainly contains two modules, ID-Injector and ID-Synchronizer. ID-Injector extracts ID features of reference characters with a Shuffling Reference Strategy (SRS), while ID-Synchronizer introduces a mask perceptual loss to modify cross-attention masks and utilizes an auto-mask self-attention module to ensure consistent generation of main characters across inter-frames, as well as vivid background.

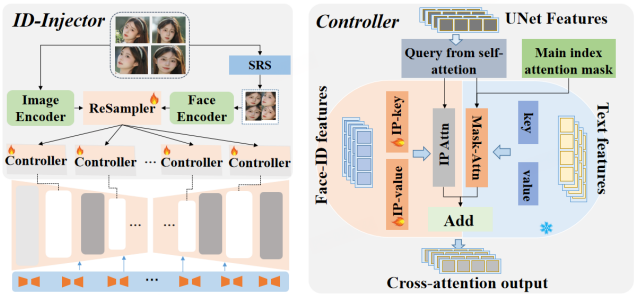


Figure 3: The structure of ID-Injector. The reference ID images are shuffled through **Shuffling Reference Strategy(SPS)**. A Resampler and several inter-frame controllers are introduced to integrate reference ID images.

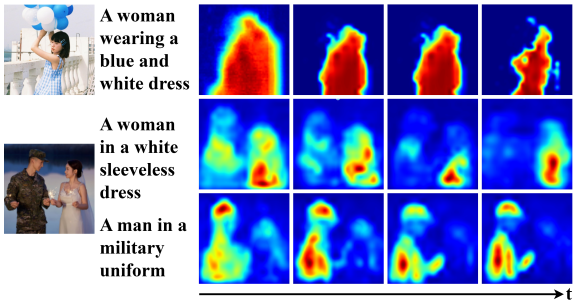


Figure 4: Cross attention map of each character during training. As the number of training steps increases, character attention maps gradually converge to accuracy within the constraints of mask perceptual loss.

where M represents the total number of characters, A_n^m represents the action of the m-th character in the n-th prompt. Notably, P_n^m refers to the description of the m-th character in the n-th prompt. Then, the series of multi-character stories generation can be formulated as follows:

$$\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_N = \mathcal{F}(z_1, ..., z_N | \mathcal{T}, \mathcal{I}_R, \theta), \tag{3}$$

where z denotes the latent noise, $\mathcal{I}_R$ represents reference images of characters. θ_i defines the parameters of Storynizor.

The pipeline of Storynizor is shown in Fig. 2(a). In contrast to existing methods, our work makes improvements in two aspects: (1) It consists of an ID-Synchronizer $\mathcal{S}$ which uses an auto-mask spacial attention module to obtain masks during diffusion process, and pay more attention to the character regions across frames, resulting in more precise consistent character and diverse background generation. (2) An ID-Injector Φ is introduced as a component, which extracts ID features of reference characters and injects them into ID-Synchronizer to generate images with instant Face-ID.

ID-Synchronizer

Previous works (Tewel et al. 2024) typically consider a spacial self-attention module to ensure consistency among inter-frames. Given a series of latent noise features $x_t \in \mathbb{R}^{B\times F\times H\times W\times C}$ and a single text prompts y, they formulate hidden states as $z_t \in \mathbb{R}^{B\times FHW\times C}$ for spacial self-attention to inherit all the module weights from the original 2D self-attention in diffusion model. ID-Synchronizer also begins with this well-explored design. However, the shared visual features across images produce nearly identical backgrounds. While maintaining minimal variation in backgrounds or layout among frames is typical for tasks like video and 3D-object generation, generating narrative images for stories demands vibrant backgrounds tailored to specific text prompts. Therefore, we introduce an Auto-mask Self-attention (AMSA) to our ID-Synchronizer to ensure consistent character generation in vivid backgrounds and postures. AMSA leverages attention masks of primary subjects, acquired from the cross-attention modules of the UNet, to concentrate on regions containing characters. It then employs spatial self-attention to these specific areas within the noise features across frames, as illustrated in Fig. 2(b). AMSA requires precise cross attention maps to achieve an excellent generation of different background and consistent characters across images. Acknowledging the constrained semantic representation of the original text encoder in Stable Diffusion, we introduce a Mask Perceptual Loss to improve the semantic representation of each character.

Auto-mask Space-Attention. Our aim is to ensure consistent character portrayal across inter-frame generation while integrating lively backgrounds. To achieve this, ID-Synchronizer extends the original self-attention module into a spatial self-attention module. Specifically, we rearrange the hidden states $z_t^{i,n}$ of each frame in the i-th layer of the diffusion model by formulating it as following:

$$z_t^i = [z_t^{i,1} \oplus z_t^{i,2} \oplus ... \oplus z_t^{i,N}] \tag{4}$$

where $x_t^i \in \mathbb{R}^{(B\times N)\times H\times W\times C}$, $\oplus$ denotes concatenation. Given that the self-attention mechanism in the diffusion model primarily handles visual information, we implement an auto-mask mechanism to incorporate attention masks of the main character region into spatial attention. This ensures that during the AMSA process, attention is masked, enabling each image to concentrate exclusively on the main character region of other frames within the batch.

In our task, the cross-attention maps are obtained to capture the areas of multiple characters in the latent image. Considering maintaining the text alignment in the story generation task, we do not make any changes to the cross-attention modules in the diffusion model. During the training process, each self-attention layer receives cross-attention maps from all preceding layers. We capture the cross-attention map of each frame in a series sample by calculating between the text embedding of P_n obtained in Eq. 2 and each noise image latent z_t of i-th UNet layer following Eq. 5:

$$\begin{aligned} q_t^i = W_q^i \cdot z_t^i, \quad k_n^i = W_k^i \cdot \mathcal{E}(P_n) \\ m_{P_n,t}^i = \sum_{i=1}^{i} \text{Softmax}(\frac{q_t^i \cdot k_n^i}{\sqrt{d_k}}), \quad n = 1, \ldots, N \end{aligned} \tag{5}$$

where n denotes n-th frame mentioned in Eq. 1, W_q^i,W_k^i are projection metrics in the cross attention module of the i-th layer, $\mathcal{E}$ represents the text encoder that encodes P into text embeddings. Thus, the masks across the inter-frame collection are defined as follows:

$$M_{P,t}^i = [m_{P_1,t}^i \oplus m_{P_2,t}^i \oplus ... \oplus m_{P_N,t}^i], \tag{6}$$

where n denotes each frame in a series of training samples, i refers to the i-th layer of UNet.

With the formulated latent noise z_t^i in Eq. 4 and the attention masks $M_{P,t}^i$ obtained by Eq. 6, the hidden states of i-th layer of the diffusion model are finally calculated as follows:

$$\begin{aligned} Q^i = W_q^i z_t^i, K^i = W_k^i z_t^i, V^i = W_v^i z_t^i \\ {z'}_t^i = Softmax(Q^i \cdot K^i / \sqrt{d_k} + \log M_{P,t}^i) \cdot V^i \end{aligned} \tag{7}$$

where W_q^i,W_k^i,W_v^i are projection matrices, ${z'}_t^i$ is the new hidden states of i-th layer of UNet after AMSA.

Mask Perceptual Loss AMSA's effectiveness relies on accurate cross-attention maps for high-quality, diverse background generation while maintaining character consistency. To enhance character semantic representation, we introduce a mask perceptual loss. We use a pre-trained segmentation model to obtain ground truth mask images for each character from training samples. Cross-attention maps are generated for each character and compared to the ground truth masks. We incorporate Dice loss(Sudre et al. 2017) as an additional constraint to optimize cross-attention masks. Thus, the loss function is reconstructed as follows:

$$\mathcal{L} = \mathcal{L}_{LDM} + \alpha \sum_{i=1}^{N} (1 - (2 * \sum_{i=1}^{M} p_i * g_i) / (\sum_{i=1}^{M} p_i^2 + \sum_{i=1}^{M} g_i^2), \tag{8}$$

where p_i refers the i-th pixel value of predict mask converted from $M_{\mathcal{T}_{tokens},t}^k$ and g_i represents the ith pixel value of ground truth mask images. M is the total number of pixels. N is the total characters in a training sample. $\mathcal{L}_{LDM}$ represents the original loss of latent diffusion models, α is the hyperparameter of the weight of mask loss. Fig. 4 illustrates the evolution of attention maps throughout the training process. Over the course of training, the cross-attention maps progressively become more accurate and increasingly resemble the ground truth masks.

ID-Injector

Since the ID-Injector is trained alongside the ID-Synchronizer, it necessitates inter-frame feature injection. Given arbitrary numbers of ID images, Storynizor develops an optional inter-frame ID-Injector, which can receive additional face ID features for continuous story generation across frames. We adopt an ID encoder $\mathcal{E}_f$ to extract ID features from given face images $\mathcal{I}_R$ and a CLIP encoder $\mathcal{E}_I$ to extract image embeddings of this face. Then we develop a Resampler $\mathcal{P}_r$ to project the face images to the condition space of the latent diffusion model. Given a set of reference images $\mathcal{I}_R = \{\mathcal{I}_n, n = 1, ..., N$, the inter-frame face embedding is defined as the following:

$$c_f = \mathcal{P}_r(\mathcal{E}_f(\mathcal{I}_R), \mathcal{E}_I(\mathcal{I}_R)), \tag{9}$$

where $c_f \in \mathbb{R}^{(B\times N)\times T\times h}$, $T \times h$ refers to the dimension of face condition embedding of each frame, $B \times N$ refers to the batch size and numbers of frames. Subsequently, another inter-frame cross-attention adaptive module is introduced into the latent diffusion model to support face images as prompts, illustrated in Fig. 3(right).

Shuffling Reference Strategy (SRS). Recent works (Li et al. 2023; Xiao et al. 2023) demonstrate various approaches to inject personalized features into diffusion models, such as original ID embedding, average ID embedding, stacked ID embedding and ID embedding with face keypoints. However, when used with ID-Synchronizer, with the integration of spatial attention modules in AMSA, the generated images are notably influenced by the initial image conditions, leading to consistent facial poses throughout story generation process. Consequently, generated facial poses in images tend to align more closely with input images.

We develop a new Shuffling Reference Strategy to our Storynizor. As illustrated in Fig. 3(a), after packaging a set of reference images with the same ID, the SPS module is utilized to shuffle the set, resulting in a shuffled $\mathcal{I}_R^{'}$. Subsequently, injecting this shuffled set into the Resampler $\mathcal{P}_r$ yields a shuffled ID embedding.

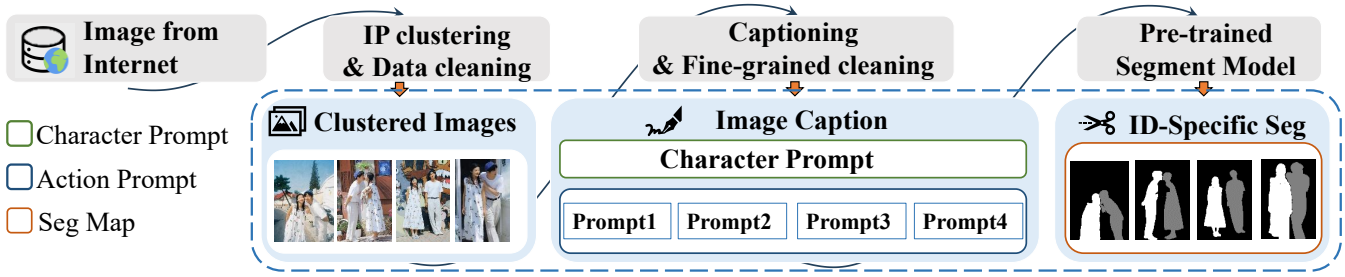


Figure 5: StoryDB Visualization and Data processing pipeline. Character Description: A woman in a floral dress, a man in white T-shirt and grey pants. Prompt1: They are on a sunny street with trees and a white structure in background. Prompt2: They are standing in front of a café, they are kissing. Prompt3: They stand side by side against a vibrant graffiti wall. Prompt4: The man is looking to the right side, they're standing beside a white wall.

Specifically, each training sample comprises N images and N associated prompts. We only consider single-character generation in training our ID-Injector. The training dataset contains:

$$\mathcal{I}_R = \{\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_N\} \tag{10}$$

This bucket $\mathcal{I}_R$ can serve as a unified face condition space. During the training process, we shuffle the bucket $\mathcal{I}_R$ with the following:

$$\mathcal{I}_R^{'} = \{\mathcal{I}_{s_1}, \mathcal{I}_{s_2}, ..., \mathcal{I}_{s_N}\} \tag{11}$$

where s_n indicates a shuffled index of the reference images. Thus, Eq. 9 can be written as the following to apply SPS into inter-frame ID-Injector:

$$c_f = \mathcal{P}_r(\mathcal{E}_f(\mathcal{I}_R^{'}), \mathcal{E}_I(\mathcal{I}_R^{'})), \tag{12}$$

The feature set c_f comprises a collection of individual ID features for each frame. Through the use of SPS, we can guarantee that every ID feature within c_f is paired with another latent noise within the diffusion model.

To inject c_f into ID-Synchronizer, we leverage the intrinsic cross-attention mechanism within the diffusion model, expanding it into an inter-frame generation as follows:

$$\begin{aligned} Q^i = W_q^i z_t^i, K^i = W_k^i c_f, V^i = W_v^i c_f \\ {z'}_t^i = Softmax(Q^i \cdot K^i / \sqrt{d_k}) \cdot V^i \end{aligned} \tag{13}$$

where W_q^i,W_k^i,W_v^i are projection matrices, ${z'}_t^i$ is the new hidden state of i-th layer of UNet after the inter-frame cross attention mechanism.

In contrast to other methods, SPS allows each image to condition on a reference image with the same ID but different from itself. This unified representation significantly enhances the robustness of the facial pose in the generated images, particularly in inter-frame generation.

StoryDB Dataset Construction

Storynizor aims to generate consistent character images across diverse backgrounds. However, as shown in Tab. 2, existing open-source datasets suffer from either limited background diversity or inconsistent character attributes. We introduce **StoryDB**, a character-centric image-text pair dataset comprising 10,000 groups, each featuring the same character in consistent attire across different scenes, totaling 100,000 images. Each group contains 5-12 images with corresponding prompts, indexed shared prompt elements, and character mask images. StoryDB not only supports Storynizor's training but also serves as a resource for future research in story generation and IP-consistent generation.

Image downloading. Initially, we collect images from the internet and open-source datasets to create a comprehensive character dataset comprising real humans, cartoon characters, and animals. We then calculate the aesthetic score of each image to filter the dataset during the download process.

IP clustering. We cluster identical IPs to generate several smaller datasets. Subsequently, we segment the images using category-specific keywords, and then calculate text-image and image-image score using CLIP. Non-compliant samples are then filtered out based on these scores.

Fine-grained filter and captioning. We use GPT-4v to align and caption images within each category. Images are collectively input to GPT-4v for character alignment. Aligned images are captioned; non-compliant ones are rejected. GPT-4v labels image sets with the same character description. Finally, we manually correct non-compliant images in grouped sets to meet training dataset requirements.

Tokenized and segmentation. Once we obtain the image-text pairs, we extract the identical descriptions in the group prompts, which are essential for generating the cross-attention map in Eq. 5. Then, we use these descriptions to generate character mask images with the pre-trained segmentation model, Segment Anything. These mask images act as ground truth to refine cross-attention maps during training.

Experiments

Implementation Details

We utilize the original checkpoint of Stable Diffusion Model-1.5 as the backbone for both ID-Synchronizer and ID-Injector. Training is conducted on 8 NVIDIA A100 GPUs, with 5% probability of dropping out text and face conditions. Inference uses DDIM (Song, Meng, and Ermon 2020) with 30 steps and a guidance scale of 7.0 on an NVIDIA A30 GPU, with the resolution of 768 $\times$ 768.

ID-Synchronizer. We train the ID-Synchronizer with its

Figure 6: Qualitative comparison of Storynizor and other consistent story generation methods. We observe Storynizor outperforms other methods when generating consistent characters with vivid backgrounds and flexible poses in prompt-only story generation. Additionally, it achieves high-fidelity ID preservation in prompt-ID story generation.

UNet parameters frozen, using the StoryDB Dataset. We train 50,000 iterations with a batch size of 4 and learning rate of 5×10^{-5} at a resolution of 512 $\times$ 512. The ID-Synchronizer is further fine-tuned at a resolution of 768 $\times$ 768 for high-fidelity generation, with a batch size of 1 for 50,000 iterations.

ID-Injector. We use a total of 80 million text-image pairs, comprising 50M from LAION-Face(Zheng et al. 2022) and 30M from the internet. We train 2 epochs with a learning rate of 1×10^{-4} and a batch size of 128 with the resolution of 512 $\times$ 512. In the second stage, we incorporate the pre-trained ID-Injector into Storynizor. We train the ID-Injector with ID-Synchronizer frozen, using the StoryDB for 5 epochs with a learning rate of 1×10^{-4} and a batch size of 4, at the resolution of 768 $\times$ 768.

Evaluation Dataset and Metrics

We use GPT-4v to generate 100 character prompts and 100 story prompts, combining them randomly into 10k test groups. Each group contains 4-story prompts and 1 character prompt. We adopt CLIP-T for text-image alignment. CLIP-I and DINO-v2 (Oquab et al. 2023) are utilized to evaluate the similarity across inter-frame generated images. For ID-based generation, we randomly select 100 faces from FFHQ (Karras, Laine, and Aila 2019) and use Arcface (Deng et al. 2019) distance to evaluate the face similarity of the given image and the generated images (Face Sim(R)) and the face similarity among inter-frame generated images (Face Sim).

Quantitative Evaluation

Table 3 presents quantitative results. In prompt-only generation, Storynizor excels in text-image consistency and inter-image coherence. For prompt-ID guided generation, InstantID scores high in facial similarity but suffers from low diversity, indicated by low CLIP-T scores, due to generating images too similar to the reference. PhotoMaker matches Storynizor in text similarity but falls short in story continuity and facial consistency. Overall, Storynizor achieves the highest comprehensive score, showcasing its superior story generation capabilities. As shown in Table 7, Storynizor consistently performs well when generating images with one to four characters, highlighting its strength in multi-character generation.

Methods	Models	Clip-T↑	Clip-I↑	Dino-I↑	Face Sim↑	Face Sim (R)↑
p-only	Storygen	25.21	67.45	67.42	10.82	-
	Consistory	29.01	<u>76.24</u>	<u>79.22</u>	<u>30.84</u>	-
	Storydiffusion	<u>30.01</u>	72.56	70.34	23.44	-
	Storynizor	**33.28**	**83.33**	**86.62**	**41.55**	-
p-ID	IP-Adapter	28.26	66.43	65.83	26.57	20.57
	PhotoMaker	**32.46**	66.23	67.38	27.34	24.34
	InstantID	25.44	<u>79.46</u>	<u>81.66</u>	**68.36**	**69.00**
	Storynizor	<u>32.42</u>	**80.86**	**82.26**	<u>39.64</u>	<u>36.46</u>

Table 3: Quantitative comparison (%) of Storynizor with other methods for prompt-only (p-only) and prompt-ID (p-ID) consistent story generation, with best and second-best results bolded and underlined.

Models	Text Alignments		Consistent Generation	
(Storynizor vs. *)	Win(%)	Lose(%)	Win(%)	Lose(%)
IP-Adapter	91.2	8.8	69.8	30.2
InstantID	100	0.0	100	0.0
Photomaker	74.8	25.2	79.3	20.7
Storygen	97.8	2.2	99.2	0.8
Consistory	69.2	30.8	61.7	38.3
Storydiffusion	65.8	34.2	63.2	36.8

Table 4: Human evaluation on Storynizor and other existing consistent story generation methods.

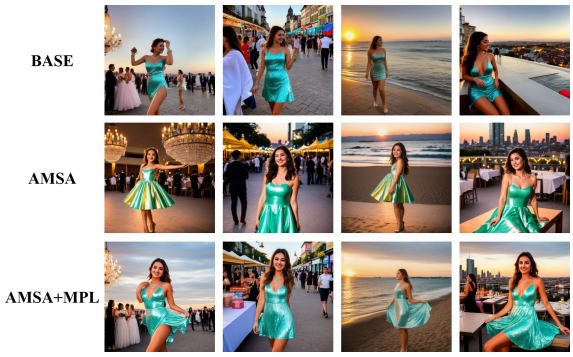


Figure 7: Qualitative ablation results of Storynizor with ASMA and MPL. Text prompt: A beautiful girl in a shiny dress danced in a ballroom event / walked through a street festival / stood on a beach / sat at a roof restaurant.

AMSA	MPL	CLIP-T↑	CLIP-I↑	DINO-I↑	Face Sim↑
✗	✗	30.46	78.83	81.29	29.90
✓	✗	32.51	81.66	83.74	31.90
✓	✓	**32.59**	**83.28**	**85.58**	**36.55**

Table 5: Quantitative ablation result (%) of auto-mask self-attention (AMSA), and mask perceptual loss (MPL). The experiments are conducted with the resolution of 512 × 512.

Method	CLIP-T↑	CLIP-I↑	DINO-I↑	Face Sim↑	Face Sim(R)↑
Stacked-ID	30.39	70.74	72.71	19.26	17.72
SRS	**32.59**	**71.65**	**75.63**	**36.48**	**32.57**

Table 6: Quantitative ablation result (%) of different types of ID injections. Stacked-ID denotes that the reference ID image is identical to the latent image. SPS refers to our shuffle reference strategy.

Qualitative Evaluation

Figure 6 compares the results of different methods. Storynizor excels in maintaining detail consistency and diversity. In multi-character generation, Storygen struggles with gender attire confusion and lacks text-image alignment. Consistory produces repetitive character layouts and fails to capture specific semantic features. Storydiffusion has issues with semantic clarity and consistency in clothing details. In contrast, Storynizor ensures character consistency and background diversity while maintaining semantic alignment. For prompt-ID guided generation, InstantID closely matches reference faces but lacks pose diversity and semantic fidelity. IP-Adapter faces significant semantic loss. PhotoMaker offers varied angles but lacks the narrative coherence of Storynizor. Overall, Storynizor surpasses other methods by generating coherent, diverse narratives while maintaining reference ID consistency.

Human Evaluation

We conduct a user study with 25 experts to evaluate Storynizor against previous methods. Each expert evaluate samples used for quantitative comparison. As shown in Tab. 4, the results indicate a preference for Storynizor over other methods in both text alignments and consistent story generation.

Method	CLIP-T↑	CLIP-I↑	DINO-I↑	Face Sim↑
1-character	33.40	84.39	86.63	42.56
2-character	32.01	83.78	86.20	41.53
3-character	32.31	83.67	86.01	41.23
4-character	31.57	82.89	85.32	40.34

Table 7: Quantitative ablation result (%) of Storynizor with multi-characters.

Method	CLIP-T↑	CLIP-I↑	DINO-I↑	Face Sim↑
StoryDB	**33.28**	**83.33**	**86.62**	**41.55**
DeepFashion	30.15	80.25	83.2	32.64

Table 8: Quantitative ablation result (%) of Storynizor with StoryDB and DeepFashion.

Ablation Studies

Influence of AMSA and MPL of ID-Synchronizer. We performe an ablation study on two components: (1) the Auto-Mask Self-Attention module (AMSA) and (2) Mask Perceptual Loss (MPL). Table 5 shows that integrating AMSA and MPL significantly improves all metrics for our model. As illustrated in Fig., AMSA enhances character consistency, while MPL improves the consistency of fine details in the generated results.

Benefits of using SRS to shuffle the input IDs. Our ID-Injector integrates personality traits from face images into cross-frame story generation. An ablation study identifies the optimal injection mode, as shown in Tab. 6. The results demonstrate that our shuffling reference strategy (SRS) surpasses stacked ID embedding in facial similarity and textual alignment, confirming SRS's superiority.

Benefits of training with StoryDB. To validate the effectiveness of our proposed StoryDB, we conduct comparative experiments by training our model on both StoryDB and DeepFashion datasets. As shown in Tab. 8, the results demonstrate that training on StoryDB yields significantly improved performance across all metrics compared to DeepFashion, confirming the efficacy of StoryDB.

Multi-character Generation. Tab. 7 demonstrates that Storynizor is capable of generating images with multiple characters. Our experiments involve generating 1 to 4 characters using Storynizor, showing that performance remains largely unaffected. These results indicate that Storynizor adeptly handles multi-character generation.

Conclusion

In conclusion, we present Storynizor, a model for generating cohesive story images with consistent characters, distinct foreground-background elements, and diverse poses. It combines ID-Synchronizer with AMSA for character consistency and vivid features, and the ID-Injector uses Shuffling Reference Strategy (SRS) for flexible face poses and consistent portrayal. Additionally, we introduce StoryDB, a 100,000-image dataset featuring diverse character sets in various settings, supporting Storynizor's training and future research.

Acknowledgments

This work is supported by the Key Research and Development Program of Zhejiang Province (No. 2022C01011)

References

Cui, A.; Mahajan, J.; Shah, V.; Gomathinayagam, P.; Liu, C.; and Lazebnik, S. 2024. Street tryon: Learning in-the-wild virtual try-on from unpaired person images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8235–8239.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4690–4699.

Gu, Y.; Wang, X.; Wu, J. Z.; Shi, Y.; Chen, Y.; Fan, Z.; Xiao, W.; Zhao, R.; Chang, S.; Wu, W.; et al. 2024. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Advances in Neural Information Processing Systems*, 36.

Guo, Y.; Yang, C.; Rao, A.; Liang, Z.; Wang, Y.; Qiao, Y.; Agrawala, M.; Lin, D.; and Dai, B. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*.

Han, X.; Wu, Z.; Wu, Z.; Yu, R.; and Davis, L. S. 2018. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7543–7552.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.

Kong, Z.; Zhang, Y.; Yang, T.; Wang, T.; Zhang, K.; Wu, B.; Chen, G.; Liu, W.; and Luo, W. 2024. Omg: Occlusion-friendly personalized multi-concept generation in diffusion models. *arXiv preprint arXiv:2403.10983*.

Li, Z.; Cao, M.; Wang, X.; Qi, Z.; Cheng, M.-M.; and Shan, Y. 2023. Photomaker: Customizing realistic human photos via stacked id embedding. *arXiv preprint arXiv:2312.04461*.

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

Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; and Tang, X. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1096–1104.

Ma, Y.; Xu, W.; Tang, J.; Jin, Q.; Zhang, R.; Zhao, Z.; Fan, C.; and Hu, Z. 2024. Character-Adapter: Prompt-Guided Region Control for High-Fidelity Character Customization. arXiv:2406.16537.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H. V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Howes, R.; Huang, P.-Y.; Xu, H.; Sharma, V.; Li, S.-W.; Galuba, W.; Rabbat, M.; Assran, M.; Ballas, N.; Synnaeve, G.; Misra, I.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2023. DINOv2: Learning Robust Visual Features without Supervision.

Papantoniou, F. P.; Lattas, A.; Moschoglou, S.; Deng, J.; Kainz, B.; and Zafeiriou, S. 2024. Arc2face: A foundation model of human faces. *arXiv preprint arXiv:2403.11641*.

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

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

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

Shi, Y.; Wang, P.; Ye, J.; Long, M.; Li, K.; and Yang, X. 2023. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*.

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

Sudre, C. H.; Li, W.; Vercauteren, T.; Ourselin, S.; and Jorge Cardoso, M. 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, 240–248. Springer.

Tewel, Y.; Kaduri, O.; Gal, R.; Kasten, Y.; Wolf, L.; Chechik, G.; and Atzmon, Y. 2024. Training-Free Consistent Text-to-Image Generation. *arXiv preprint arXiv:2402.03286*.

Wang, Q.; Bai, X.; Wang, H.; Qin, Z.; and Chen, A. 2024. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*.

Xiao, G.; Yin, T.; Freeman, W. T.; Durand, F.; and Han, S. 2023. FastComposer: Tuning-Free Multi-Subject Image Generation with Localized Attention. *arXiv preprint arXiv:2305.10431*.

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

Zheng, Y.; Yang, H.; Zhang, T.; Bao, J.; Chen, D.; Huang, Y.; Yuan, L.; Chen, D.; Zeng, M.; and Wen, F. 2022. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18697–18709.

Zhou, Y.; Zhou, D.; Cheng, M.-M.; Feng, J.; and Hou, Q. 2024. StoryDiffusion: Consistent Self-Attention for Long-Range Image and Video Generation. *arXiv preprint arXiv:2405.01434*.