STORYMAKER: TOWARDS HOLISTIC CONSISTENT CHARACTERS IN TEXT-TO-IMAGE GENERATION

TECHNICAL REPORT

Zhengguang Zhou, Jing Li, Huaxia Li,* Nemo Chen, Xu Tang

Xiaohongshu Inc.

Figure 1: Visualization of images generated by our StoryMaker. The first three rows depict a story about a day in the life of an "office worker," while the last two rows tell a story inspired by the movie "Before Sunrise."

---

*Project Leader

ABSTRACT

Tuning-free personalized image generation methods have achieved significant success in maintaining facial consistency, *i.e.*, identities, even with multiple characters. However, the lack of holistic consistency in scenes with multiple characters hampers these methods' ability to create a cohesive narrative. In this paper, we introduce StoryMaker, a personalization solution that preserves not only facial consistency but also clothing, hairstyles, and body consistency, thus facilitating the creation of a story through a series of images. StoryMaker incorporates conditions based on face identities and cropped character images, which include clothing, hairstyles, and bodies. Specifically, we integrate the facial identity information with the cropped character images using the Positional-aware Perceiver Resampler (PPR) to obtain distinct character features. To prevent intermingling of multiple characters and the background, we separately constrain the cross-attention impact regions of different characters and the background using MSE loss with segmentation masks. Additionally, we train the generation network conditioned on poses to promote decoupling from poses. A LoRA is also employed to enhance fidelity and quality. Experiments underscore the effectiveness of our approach. StoryMaker supports numerous applications and is compatible with other societal plug-ins. Our source codes and model weights are available at https://github.com/RedAIGC/StoryMaker.

1 Introduction

Diffusion-based image generation methods, such as DALL-E [Ramesh et al., 2021], Imagen [Saharia et al., 2022], and Stable Diffusion [Rombach et al., 2021], have recently made significant advancements. However, personalizing generated content using texts alone remains challenging. To address this, test-time fine-tuning methods [Avrahami et al., 2023, Gal et al., 2022, Kumari et al., 2023, Ruiz et al., 2023] have been proposed to produce images with specific subjects. Nevertheless, their generalization ability is constrained by the limited number of images and the high cost of fine-tuning. Consequently, tuning-free methods [Li et al., 2024a, Ma et al., 2024, Wei et al., 2023, Xiao et al., 2023, Wei et al., 2024, Kim et al., 2024, Ye et al., 2023a, Wang et al., 2024a, Han et al., 2024] trained on large-scale datasets have been introduced. These methods employ a visual encoder to integrate visual information into the generator without the need for lengthy fine-tuning. While Xiao et al. [2023], Wei et al. [2024], Kim et al. [2024] preserve facial identities, they fail to maintain the holistic consistency including consistent clothing, hairstyles, and bodies, thereby limiting their applications.

In this paper, we introduce StoryMaker, which pursues the holistic consistency, not only preserving facial identities but also clothing, hairstyles, and bodies. StoryMaker allows variation in backgrounds, character poses, and styles through text prompts, enabling the generation of a series of images with consistent characters, thereby creating a narrative. StoryMaker also facilitates applications such as clothing swapping and image variation and is compatible with plug-ins like LoRA for stylization.

To preserve clothing, hairstyles, and bodies in addition to faces, StoryMaker conditions the generation on face identities and cropped character images, which include clothing, hairstyles, and bodies. After extracting information from the reference image, we integrate face identities and cropped character images using the Positional-aware Perceiver Resampler (PPR) to derive character features.

As it is more difficult to retain clothing, hair styles and bodies, other than only face identities, StoryMaker regularizes the cross-attention impact region among different characters as well as the background. Unlike MM-Diff [Wei et al., 2024], which only separates different foregrounds, we include a learnable background embedding to encourage differentiation from the background. An ID loss is introduced to further regularize identities. To decouple generation from the poses of cropped character images, enhancing diversity and utility, we train our model on predicted poses with ControlNet [Zhang et al., 2023]. During inference, ControlNet can be omitted, allowing poses to be guided directly by text prompts. Alternatively, referred poses can be provided to ControlNet. A LoRA is employed to improve fidelity and quality. By combining these elements, StoryMaker generates image series with consistent faces, clothing, hairstyles, and bodies, thereby constructing a coherent story.

In summary, the main contributions of this paper are: i) We address the task of generating a series of images with consistent faces, clothing, hairstyles, and bodies, while allowing variations in backgrounds, poses, and styles via text prompts, enabling narrative creation. ii) To tackle this complex task, we propose StoryMaker, which first extracts information from reference images and refines it using the Positional-aware Perceiver Resampler. To prevent different characters and the background from interleaving each other, we regularize the cross-attention impact region using MSE loss with segmentation masks and train the backbone network conditioned on poses by ControlNet to facilitate decoupling. We also train a LoRA to enhance fidelity and quality. iii) Experiments demonstrate that our proposed StoryMaker achieves excellent performance and has diverse applications in real-world scenarios.

2 Related Work

2.1 Subject-Driven Image Generation

Subject-driven text-to-image generation has achieved remarkable progress. Current methods in this domain can be categorized based on whether they necessitate test-time fine-tuning for input images. Early approaches [Ruiz et al., 2023, Gal et al., 2022, 2023] require test-time optimization of specific text tokens to represent target concepts using a limited set of subject images. These fine-tuning methods are time-consuming due to the slow optimization process before inference. Recent methods aim to eliminate the need for fine-tuning by integrating additional modules while keeping the primary pre-trained text-to-image models frozen. Subject-Diffusion [Ma et al., 2024] substitutes text tokens describing subjects with the corresponding image embeddings and trains an adapter module to incorporate fine-grained image features. ELITE [Wei et al., 2023] and FastComposer [Xiao et al., 2023] also map images to text embeddings by training an additional network. Blip-Diffusion [Li et al., 2024b] employs the pre-trained multi-modal encoder BLIP-2 [Li et al., 2023] to infuse image information. IP-Adapter [Ye et al., 2023b] separates image and text features in cross-attention, allowing for independent image feature integration. MoA (Mixture-of-Attention) [Ostashev et al., 2024] enhances image quality by segregating subject and context. The SSR-Encoder [Zhang et al., 2024a] is a recent development that integrates segment information into text features through cross-attention, facilitating selective feature extraction.

Identity-preserving human image generation is a prominent area in subject-driven image generation, given its broad real-world applications. Solutions such as FaceStudio [Yan et al., 2023], IP-Adapter-FaceID [Ye et al., 2023b], FlashFace [Zhang et al., 2024b], and PhotoMaker [Li et al., 2024a] utilize ID embeddings derived from Arcface [Deng et al., 2019] as a condition, which is crucial for maintaining facial fidelity. The leading approach, InstantID [Wang et al., 2024a], introduces IdentityNet, which uses five facial keypoints to control face structure, achieving optimal face similarity. Beyond single-ID customization, some methods [Wei et al., 2024, He et al., 2024a, Jang et al., 2024, Kumari et al., 2023, Avrahami et al., 2023] focus on multi-ID image generation. Some studies [Kim et al., 2024, Kong et al., 2024] use predefined layouts to guide multi-ID image generation, while limits the scalability in real-world scenarios. In contrast, MM-diff [Wei et al., 2024] imposes constrains on the cross-attention maps with different subjects during the training phase, which guarantees the generation of multi-ID images without any predefined input. Recently, UniPortrait [He et al., 2024a] employs an ID routing module to unify multi-ID customization, avoiding identity blending. Our proposed StoryMaker not only preserves faces in image generation, but also ensures consistency in clothing, hairstyle, and bodies. For multi-character generation, we introduce the Positional-aware Perceiver Resampler and attention loss to address multi-character blending.

2.2 Image Story Generation

Maintaining consistent content across a series of generated images has numerous real-world applications, such as story visualization and comic creation. StoryDiffusion [Zhou et al., 2024] employs a consistent self-attention mechanism that adapts information from other images in the batch to ensure character consistency in storytelling sequences. Unlike StoryDiffusion, which adapts from full images, ConsiStory [Tewel et al., 2024] uses a subject-driven shared attention block that only adapts information from masked subjects, with correspondence-based feature injection enhancing subject consistency between images. DreamStory [He et al., 2024b] leverages a Large Language Model (LLM) for better understanding and guidance in generation. Subjects are generated first, followed by a Multi-Subject Consistent Diffusion model, ensuring subject consistency across images by adapting information from other images in self-attention and from texts in cross-attention, similar to ConsiStory. OneActor [Wang et al., 2024b] introduces a cluster-conditioned generation paradigm, achieving controlled, consistent subject generation by tuning an adapter to inject modified prompt embeddings into the fixed U-Net. Our method focuses on generating images with consistent subjects given references, while these methods work without references. Approaches like StoryDiffusion, ConsiStory, and DreamStory are training-free, yet backgrounds can easily be involved with inaccurate masks.

3 Preliminaries

We build our model on the state-of-the-art text-to-image model, namely Stable Diffusion XL [Podell et al., 2023]. In this section, we first introduce preliminaries about diffusion models and IP-adapter [Ye et al., 2023b], which form the foundation of our method.
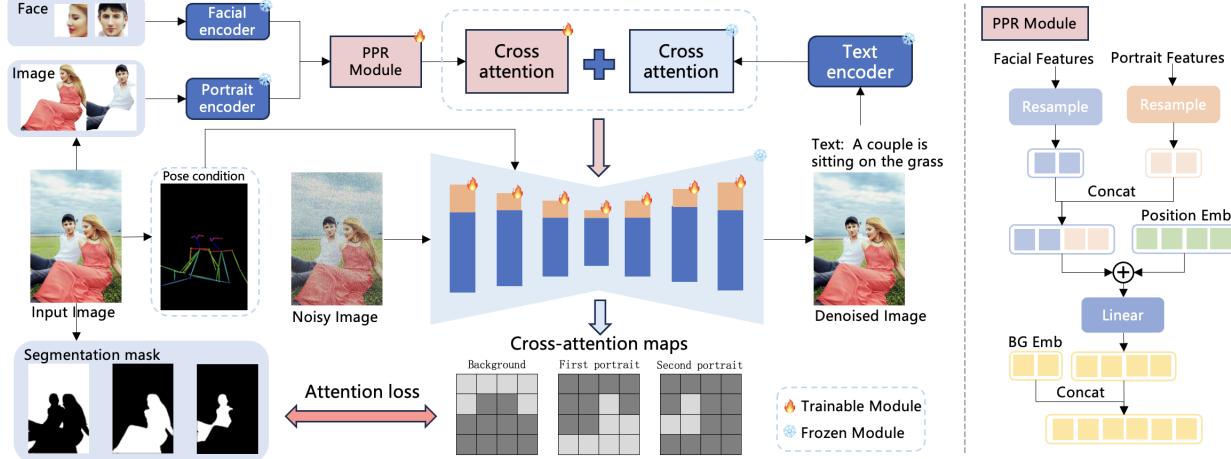
Figure 2: he model architecture of our proposed StoryMaker. The facial image and character image are embedded using the face encoder and image encoder, respectively, and refined through our proposed Positional-aware Perceiver Resampler module. Decoupled cross-attention with LoRAs is employed to inject these embeddings into the diffusion model. At the bottom, we illustrate the attention loss on cross-attention maps with the segmentation mask. The core of the PPR module is also depicted on the right.

3.1 Stable diffusion

The innovation of Stable Diffusion resides in executing the diffusion process within a low-dimensional latent space to enhance computational efficiency. This approach incorporates three primary components: a variational autoencoder (VAE) [Kingma, 2013] for compressing input images into the latent space, a text encoder to transform textual prompts into embeddings, and a U-Net [Ronneberger et al., 2015] for the denoising procedure. For a given input image x of dimensions $H \times W \times 3$, the VAE encoder ε transforms it to a latent representation $z_0 = \varepsilon(x)$ of dimensions $H/8 \times W/8 \times C$, where 8 is the downsampling factor and C is the latent dimension. The denoising process employs a U-Net ϵ_θ to denoise the normally-distributed noise ϵ added to the noisy latent z_t at timestep t, conditioned on c. Here, c denotes the text embeddings generated by the pre-trained CLIP text encoder. The overall training objective is defined as:

$$\mathcal{L}_{SD} = \mathbb{E}_{\varepsilon(x),c,\epsilon \sim \mathcal{N}(0,1),t} \left[\|\epsilon - \epsilon_\theta(z_t,t,c)\|_2^2 \right] \tag{1}$$

During inference, a random noise z_t is drawn from Gaussian noise and iteratively denoised by the U-Net to yeild the initial latent representation $\hat{z}_0$. Subsequently, the VAE decoder D converts the initial latent into the pixel space as $\hat{x} = D(\hat{z}_0)$.

3.2 IP-Adapter

IP-Adapter [Ye et al., 2023b] introduces an image prompt adapter that allows the diffusion model to generate images conditioned on an image prompt. The method comprises two components: an image encoder to extract features from the reference image, and an adapter module with decoupled cross-attention layers to integrate these image features into the pre-trained text-to-image model. Specifically, in the original cross-attention layer of the diffusion model, given the query features Z and text features c_t, the output of cross-attention Z_t is defined by the following equation:

$$Z_t = Attention(Q, K_t, V_t) = Softmax(\frac{QK_t^T}{\sqrt{d}})V_t, \tag{2}$$

where $Q = ZW_q$, $K = c_tW_k^t$, $V = c_tW_v^t$ are the query, key, and value matrices for the attention operation, respectively, and d denotes the channel dimension of the feature. The newly introduced decoupled cross-attention is calculated as follows:

$$Z_{new} = Attention(Q, K_t, V_t) + \gamma \cdot Attention(Q, K_i, V_i), \tag{3}$$

4

where c_i is the image prompt embed and $K_i = c_i W_k^i, V_i = c_i W_v^i$ constitute the added attention operation for the image cross-attention. Here only W_k^i and W_v^i are the trainable weights.

4 Method

4.1 Overview

Given a reference image containing one or two characters, StoryMaker seeks to generate a series of new images featuring the same characters, maintaining not only identical faces, $i.e.$, identities, but also their clothing, hairstyles, and bodies. A narrative can then be created by altering the background, the characters' poses, and the style according to the text prompts.

Specifically, we first extract the facial information, $i.e.$, identities, of the characters using the face encoder, and the details of their clothing, hairstyles, and bodies via the character image encoder. We then refine this information using the proposed Positional-aware Perceiver Resampler. To control the backbone generation network, we inject the refined information into the decoupled cross-attention module proposed by IP-Adapter [Ye et al., 2023b]. To prevent multiple characters and the background from interleaving, we constrain the impact region of the cross-attention for different characters and background separately. ID loss is additionally utilized to maintain the characters' identities. Furthermore, to decouple pose information from the reference image, we train the network conditioned on detected poses by ControlNet [Zhang et al., 2023]. For enhanced fidelity and quality, we also train the U-Net with LoRA [Hu et al., 2021]. Once trained, we can either discard the entire ControlNet and control the characters' poses through text prompts or guide image generation with new poses during inference. The complete pipeline of our proposed method is illustrated in Figure 2.

4.2 Reference Information Extraction

Since the facial features extracted by the face recognition model effectively capture semantic details and enhance fidelity, similar to InstantID [Wang et al., 2024a] and IP-Adapter-FaceID [Ye et al., 2023b], we utilize Arcface [Deng et al., 2019] to detect faces and obtain aligned facial embeddings from the reference image. To maintain consistency in hairstyles, clothing, and bodies, we first segment the reference image to crop the characters. Following recent works such as IP-Adapter [Ye et al., 2023b] and MM-Diff [Wei et al., 2024], we use the pretrained CLIP vision encoder, known for its rich content and style, to extract features of the hairstyles, clothing, and bodies of the characters. During training, the face encoder, $i.e.$, Arcface model, and the image encoder, $i.e.$, CLIP vision encoder, are kept frozen.

4.3 Reference Information Refinement by Positional-aware Perceiver Resampler

Following InstantID [Wang et al., 2024a] and IP-adapter [Ye et al., 2023b], we utilize two independent resampler modules to transform the facial features, $i.e.$, F_{face}, and the character features, $i.e.$, $F_{character}$, into facial embeddings and character embeddings, respectively. These embeddings are concatenated and augmented with positional embeddings, i.e., E_{pos}, which serve to distinguish different characters. To differentiate the foreground from the background, we introduce a learnable background embedding, $i.e.$, E_{bg} and concatenate it into the final embedding. Denoting the two independent resampler modules as R_1 and R_2, the Positional-aware Perceiver Resampler is formulated as follows:

$$E_1 = R_1(F_{face}) \tag{4}$$
$$E_2 = R_2(F_{character}) \tag{5}$$
$$E_i = MLP(Cat(E_1, E_2) + E_{pos}) \tag{6}$$
$$c_i = Cat(E_{bg}, Reshape(E_i, (N * L, D))) \tag{7}$$

where L represent the number of tokens and the dimension of the character embeddings, respectively, and N denotes the number of characters in the reference image. The image prompt embed for image cross-attention is c_i. We denote the L tokens of the background embedding as E_{bg}, resulting in the dimension of c_i is $((N + 1) \cdot L) \times D$.

4.4 Decoupled Cross-attention

After extracting the reference information, we utilize the decoupled cross-attention to embed it into the text-to-image model, following IP-Adapter [Ye et al., 2023b].

4.5 Pose Decoupling from Character Images

Pose diversity is essential for storytelling. Training conditioned solely on character images can lead to the network overfitting to the poses of the reference images, resulting in generated characters with identical poses. To facilitate decoupling poses from character images, we condition the training on poses using Pose-ControlNet [Zhang et al., 2023]. During inference, we can either discard ControlNet and employ text prompts to control the poses of generated characters or guide generation with a newly provided pose.

4.6 Training with LoRA

Furthermore, to enhance ID consistency, fidelity, and quality akin to IP-Adapter-FaceID, LoRA layers [Hu et al., 2021] are integrated into each attention layer of the diffusion model. Specifically, in each cross-attention layer, Q, K_t, V_t, K_i and V_i are modified as follows:

$$
\begin{cases}
Q & = Z(W_q + \Delta W_q), \\
K_t & = c_t(W_k^t + \Delta W_k^t), \\
V_t & = c_t(W_v^t + \Delta W_v^t), \\
K_i & = c_i(W_k^i + \Delta W_k^i), \\
V_i & = c_i(W_v^i + \Delta W_v^i)
\end{cases}
\tag{8}
$$

We freeze the U-Net model, and only the ΔW is trainable.

4.7 Loss Constraints on Cross-attention Maps with Masks

To prevent multiple characters and the background from interleaving, we regularize the influence region of cross-attention using the embeddings of different characters and the background. Unlike MM-Diff [Wei et al., 2024], which does not consider the background, we introduce a learnable background embedding to address it. We constrain the influence region by calculating the MSE loss between the softmax values of cross-attention and the segmentation masks predicted by a pre-trained network. This design, *i.e.*, introducing a learnable background embedding, encourages a better separation not only within the foreground characters but also between foreground and background. As seen in Equation 7, the first L tokens of image prompt c_i represent the background, with each subsequent set of L tokens representing each character. In each layer of image cross-attention, we obtain the cross-attention map A of size $h \times w$ for each character by summing all its L tokens as:

$$
P = Softmax(QK^T / \sqrt{d}),
\tag{9}
$$

$$
A = \sum_{k=1}^{L} P_k
\tag{10}
$$

Our proposed attention loss $\mathcal{L}_{attn}$ can be formulated as follows:

$$
\mathcal{L}_{attn} = \frac{1}{N+1} \sum_{k=1}^{N+1} \|A_k - M_k\|_2^2,
\tag{11}
$$

where N is number of characters in the reference image, and "+1" represents the background.

4.8 Overall Loss

In training, we average $\mathcal{L}_{attn}$ across all M layers and combine it with the diffusion loss as follows:

$$
\mathcal{L} = \mathcal{L}_{SD} + \frac{\lambda}{M} \sum_{l=1}^{M} \mathcal{L}_{attn}
\tag{12}
$$

where $\mathcal{L}$ is our final training objective and λ is a weighting scalar.

5 Experiments

5.1 Setup

5.1.1 Datasets

We collect an internal character dataset consisting of a total of 500K images, including 300K single-character images and 200K two-character images. Image captions are automatically generated using CogVLM [Wang et al., 2023].

Table 1: Quantitative comparisons on character-conditioned generation.The best results are in **bold**.

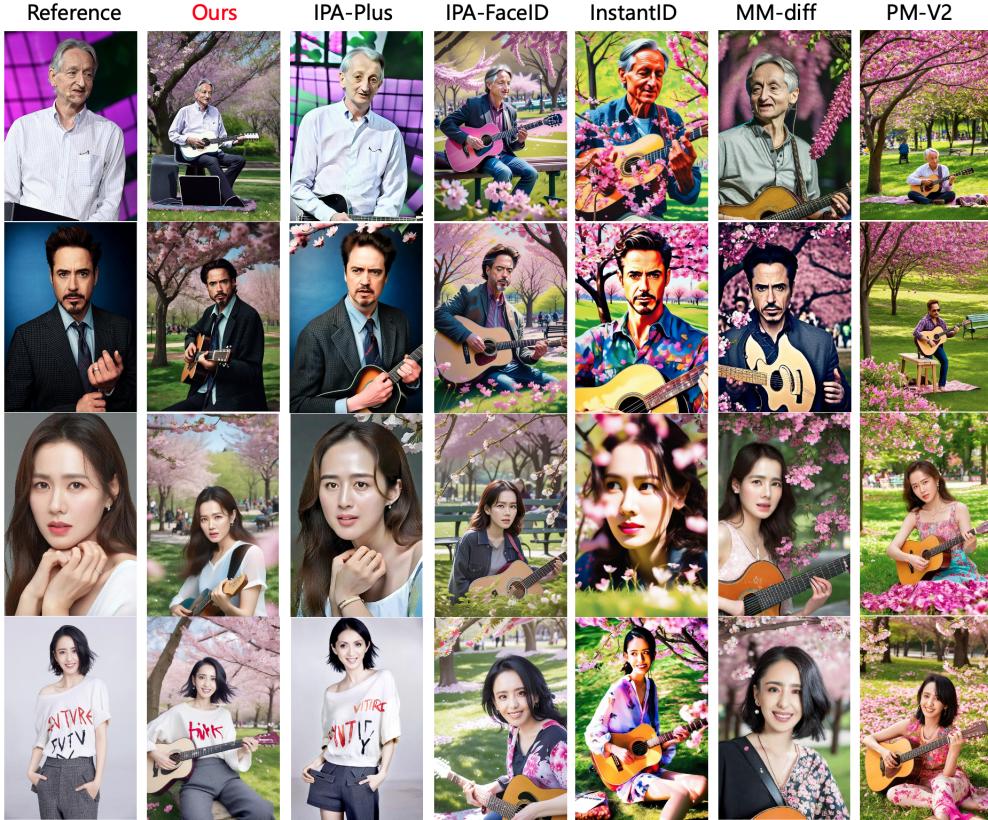| Method	Multi-person	Clothing, etc.	Face Sim. ↑ (%)	CLIP-T ↑ (%)	CLIP-I ↑ (%)
MM-Diff [Wei et al., 2024]	✔	✗	60.70	19.81	<u>78.46</u>
PhotoMaker-V2 [Li et al., 2024a]	✗	✗	61.83	**23.71**	75.31
IP-adapter-FaceID [Ye et al., 2023b]	✗	✗	66.66	22.97	68.72
InstantID [Wang et al., 2024a]	✗	✗	**73.90**	<u>23.39</u>	72.50
StoryMaker(Ours)	✔	✔	<u>67.36</u>	21.75	**79.51**

Figure 3: Visual comparison on single character condition generation.

We employ the buffalo_l [Deng et al., 2019] model to detect and obtain the ID-embedding of each face. Character segmentation masks are acquired using our internal instance segmentation model.

5.1.2 Training Details

We train our model based on Stable Diffusion XL [Rombach et al., 2022]. Similar to IP-Adapter-FaceID [Ye et al., 2023b], we utilize buffalo_l [Deng et al., 2019] as the face recognition model and OpenCLIP ViT-H/14 [Ilharco et al., 2021] as the image encoder. The rank of trainable LoRA weights is set to 128. During training, we freeze the original weights of the base model and train only the PPR module and LoRA weights. Additionally, we initialize the weights of the resample module for the face and character from IP-Adapter-FaceID and IP-Adapter, respectively. Our model is trained for 8k steps on 8 NVIDIA A100 GPUs with a batch size of 8 per GPU. We use AdamW with a learning rate of 1e-4 for the first 4k steps and 5e-5 for the last 4k steps. We set λ to 0.1. Training images are resized to a 1024×1024 resolution. The text caption is randomly dropped by 10% during training, and the cropped character image is randomly dropped by 5%. During inference, we use the UniPC [Zhao et al., 2024] sampler with 25 steps and set the classifier-free guidance to 7.5.
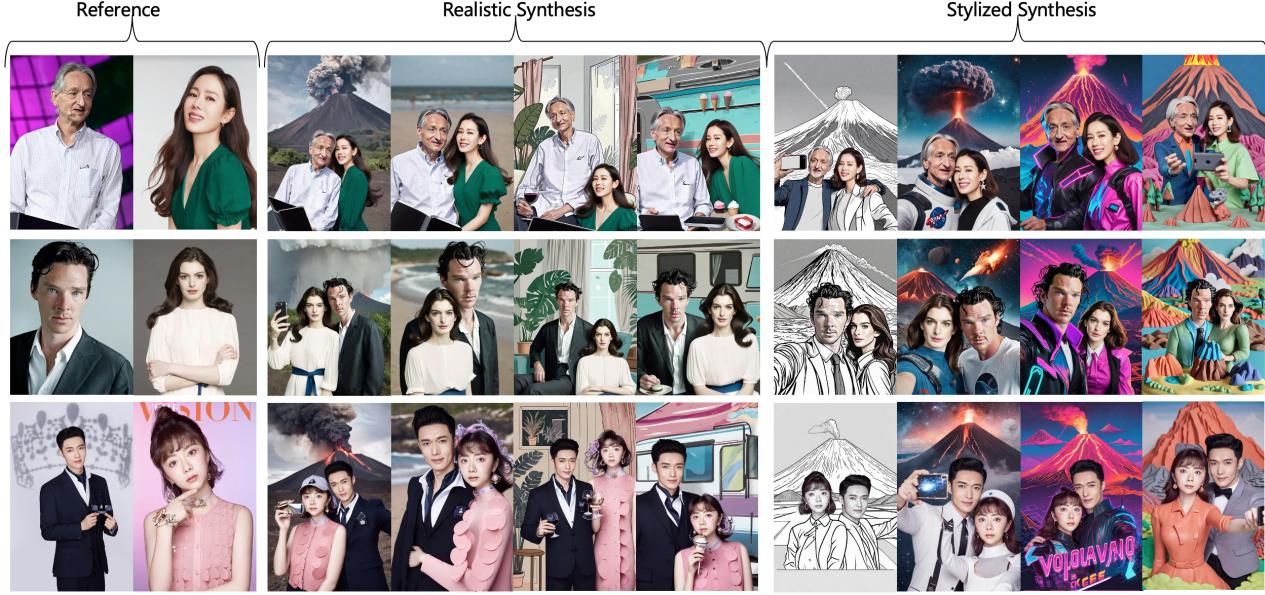
Figure 4: Visualization of two-character image generation. The first two columns display two different reference character images. The middle four columns illustrate StoryMaker's ability for realistic synthesis. The last four columns demonstrate results of stylized synthesis, where the character embedding is set to zero.

5.1.3 Evaluation Metrics

To compare with other methods, we evaluate our methods in a single-character setting. We collect a dataset of 40 characters and adopt 20 unique text prompts from FastComposer [Xiao et al., 2023] and generate 4 images for each prompt. Following FastComposer [Xiao et al., 2023] and MM-Diff [Wei et al., 2024], we use CLIP image similarity (CLIP-I) to compare the generated images with reference images. For identity preservation, we employ buffalo_l [Deng et al., 2019] model to detect and calculate the cosine similarity (Face Sim.) between two face images. Additionally, we assess the image-text similarity using the CLIP-score (CLIP-T).

5.2 Results

5.2.1 Quantitative Evaluation

As shown in Table 1, we compare our StoryMaker with four tuning-free character generation models, including MM-Diff [Wei et al., 2024], PhotoMaker-V2 [Li et al., 2024a], InstantID [Wang et al., 2024a], and IP-Adapter-FaceID [Ye et al., 2023b]. Our proposed StoryMaker achieves the highest CLIP-I score among previous methods due to the consistency of the entire portrait, including face, hairstyle, and clothing, though it has a relatively lower CLIP-T, slightly compromising text prompt adherence. For face similarity, our method outperforms others except for InstantID. We attribute InstantID's superior performance to the extensive training data and the IdentityNet controlling module. It should be noted that among all evaluated methods, only MM-Diff and our method can preserve the ID of multiple persons. Moreover, StoryMaker is the only approach that maintains consistency not only in faces but also in clothing, hairstyles, and bodies.

5.2.2 Visualization

Single-Character Image Generation. As shown in Figure 3, compared to IP-Adapter-FaceID, InstantID, MM-Diff, and PhotoMaker-V2, which are designed for identity preservation, the proposed StoryMaker not only maintains face fidelity but also clothing consistency. While IP-Adapter-Plus performs well on clothing consistency, it falls short in text prompts following and face fidelity.

Multiple-characters Image Generation. We further demonstrate the performance of multiple-character image generation. As shown in Figure 4, with a text prompt, our method can generate different poses of two characters while maintaining consistency in faces, clothing, and hairstyles. Additionally, due to the use of two independent resampler

(a) Gender editing with text prompts.   (b) Clothing swap   (c) + LoRA

(d) Image variation   (e) Pose control

Portrait1 $\alpha = 1$ 0.8   0.7   0.6   0.5   0.4   0.3   0.2   0.0   Portrait2

(f) Portrait interpolation with a interpolation coefficient α.

Figure 5: Diverse applications of StoryMaker.

modules, we can set the character embedding (E_2 in Equation 7) to all zero, while maintaining only ID-preserving and generating stylized synthesis in the last four columns in Figure 4.

Personalized Story Diffusion. Given reference character images, our proposed StoryMaker can generate consistent character images based on arbitrary prompts, enabling the creation of a story using a series of prompts. As illustrated in the top three rows of Figure 1, our method generates a series of images of a single person according to a short story composed of five text prompts describing "A day in the life of an office worker." The poses of the generated characters vary without being controlled by given pose maps. In the bottom two images of Figure 1, we present a story featuring the movie "Before Sunrise," generated with two characters. To achieve optimal results, we control the generation using specified poses.

Applications. The excellent performance of our method in aligning IDs, clothing, maintaining prompt consistency, and enhancing the diversity and quality of generated images provides a strong foundation for diverse downstream applications. As shown in Figure 5(a), a man or woman could become a boy or girl while maintaining clothing consistency. Additionally, StoryMaker demonstrates a surprising ability for clothing swapping (Figure 5(b)), achieved by replacing the character image with a clothing image, indicating that the character embedding contains clothing information. Moreover, similar to IP-Adapter [Ye et al., 2023b] and InstantID [Wang et al., 2024a], StoryMaker functions as a plug-and-play module, capable of integrating with LoRA or ControlNet to generate diverse images while maintaining character consistency, as shown in Figure 5(c,e). Due to the character-preserving capability, human image variations can be realized, as illustrated in Figure 5(d). Furthermore, we explore character interpolation between two characters, showcasing StoryMaker's ability to blend features from multiple characters, as demonstrated in Figure 5(f).

6 Conclusion

In this paper, we introduce StoryMaker, a novel approach for personalized image generation that excels maintaining consistency not only in facial identities but also in clothing, hairstyles, and bodies across multiple characters scenes. Our method enhances narrative creation by allowing background, pose, and style variations via text prompts, enabling diverse and coherent storytelling. StoryMaker leverages the Positional-aware Perceiver Resampler to obtain distinct character embeddings by fusing the features extracted from the face image and the cropped character image. To prevent intermingling of multiple characters and the background, we separately constrain the cross-attention impact regions of different characters and the background using MSE loss with segmentation masks. By incorporating pose decoupling through ControlNet and fidelity enhancements with LoRA, StoryMaker consistently generates high-quality images with matched identities and visual consistency. Our extensive experiments demonstrate StoryMaker's superior performance in maintaining character identity and consistency, especially in multi-character scenarios, outperforming existing tuning-free models. The model's versatility is further highlighted through various applications such as clothing swapping, character interpolation, and integration with other generative plug-ins. We believe StoryMaker significantly contributes to personalized image generation and opens possibilities for wide applications in digital storytelling, comics, and beyond, where individuality and narrative coherence are essential.

7 Limitations

In the absence of an explicit pose guide, the posture of generated characters often exhibits anomalies and lacks harmony. Moreover, generating three or more characters simultaneously presents significant challenges. The fidelity and detail of the generated clothing remain unsatisfactory.

References

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. arXiv:2112.10752, 2021.

Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–12, 2023.

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 arXiv:2208.01618*, 2022.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8640–8650, 2024a.

Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–12, 2024.

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15943–15953, 2023.

Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023.

Zhichao Wei, Qingkun Su, Long Qin, and Weizhi Wang. Mm-diff: High-fidelity image personalization via multi-modal condition integration. *arXiv preprint arXiv:2403.15059*, 2024.

Chanran Kim, Jeongin Lee, Shichang Joung, Bongmo Kim, and Yeul-Min Baek. Instantfamily: Masked attention for zero-shot multi-id image generation. *arXiv preprint arXiv:2404.19427*, 2024.

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 arXiv:2308.06721*, 2023a.

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024a.

Yucheng Han, Rui Wang, Chi Zhang, Juntao Hu, Pei Cheng, Bin Fu, and Hanwang Zhang. Emma: Your text-to-image diffusion model can secretly accept multi-modal prompts. *arXiv preprint arXiv:2406.09162*, 2024.

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*, pages 3836–3847, 2023.

Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models. *ACM Transactions on Graphics (TOG)*, 42(4):1–13, 2023.

Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36, 2024b.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

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 arxiv:2308.06721*, 2023b.

Daniil Ostashev, Yuwei Fang, Sergey Tulyakov, Kfir Aberman, et al. Moa: Mixture-of-attention for subject-context disentanglement in personalized image generation. *arXiv preprint arXiv:2404.11565*, 2024.

Yuxuan Zhang, Yiren Song, Jiaming Liu, Rui Wang, Jinpeng Yu, Hao Tang, Huaxia Li, Xu Tang, Yao Hu, Han Pan, et al. Ssr-encoder: Encoding selective subject representation for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8069–8078, 2024a.

Yuxuan Yan, Chi Zhang, Rui Wang, Yichao Zhou, Gege Zhang, Pei Cheng, Gang Yu, and Bin Fu. Facestudio: Put your face everywhere in seconds. *arXiv preprint arXiv:2312.02663*, 2023.

Shilong Zhang, Lianghua Huang, Xi Chen, Yifei Zhang, Zhi-Fan Wu, Yutong Feng, Wei Wang, Yujun Shen, Yu Liu, and Ping Luo. Flashface: Human image personalization with high-fidelity identity preservation. *arXiv preprint arXiv:2403.17008*, 2024b.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

Junjie He, Yifeng Geng, and Liefeng Bo. Uniportrait: A unified framework for identity-preserving single-and multi-human image personalization. *arXiv preprint arXiv:2408.05939*, 2024a.

Sangwon Jang, Jaehyeong Jo, Kimin Lee, and Sung Ju Hwang. Identity decoupling for multi-subject personalization of text-to-image models. *arXiv preprint arXiv:2404.04243*, 2024.

Zhe Kong, Yong Zhang, Tianyu Yang, Tao Wang, Kaihao Zhang, Bizhu Wu, Guanying Chen, Wei Liu, and Wenhan Luo. Omg: Occlusion-friendly personalized multi-concept generation in diffusion models. *arXiv preprint arXiv:2403.10983*, 2024.

Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *arXiv preprint arXiv:2405.01434*, 2024.

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)*, 43(4):1–18, 2024.

Huiguo He, Huan Yang, Zixi Tuo, Yuan Zhou, Qiuyue Wang, Yuhang Zhang, Zeyu Liu, Wenhao Huang, Hongyang Chao, and Jian Yin. Dreamstory: Open-domain story visualization by llm-guided multi-subject consistent diffusion. *arXiv preprint arXiv:2407.12899*, 2024b.

Jiahao Wang, Caixia Yan, Haonan Lin, and Weizhan Zhang. Oneactor: Consistent character generation via cluster-conditioned guidance. *arXiv preprint arXiv:2404.10267*, 2024b.

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 arXiv:2307.01952*, 2023.

DP Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL `https://doi.org/10.5281/zenodo.5143773`. If you use this software, please cite it as below.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.