IdentityStory: Taming Your Identity-Preserving Generator for Human-Centric Story Generation

Donghao Zhou[1*], Jingyu Lin[2*], Guibao Shen[3], Quande Liu[4], Jialin Gao[1], Lihao Liu[5], Lan Du[2], Cunjian Chen[2†], Chi-Wing Fu[1], Xiaowei Hu[6†], Pheng-Ann Heng[1]

[1]The Chinese University of Hong Kong
[2]Monash University
[3]The Hong Kong University of Science and Technology (Guangzhou)
[4]Kling Team, Kuaishou Technology
[5]Amazon
[6]South China University of Technology

Abstract

Recent visual generative models enable story generation with consistent characters from text, but human-centric story generation faces additional challenges, such as maintaining detailed and diverse human face consistency and coordinating multiple characters across different images. This paper presents **IdentityStory**, a framework for human-centric story generation that ensures consistent character identity across multiple sequential images. By taming identity-preserving generators, the framework features two key components: *Iterative Identity Discovery*, which extracts cohesive character identities, and *Re-denoising Identity Injection*, which re-denoises images to inject identities while preserving desired context. Experiments on the ConsiStory-Human benchmark demonstrate that IdentityStory outperforms existing methods, particularly in face consistency, and supports multi-character combinations. The framework also shows strong potential for applications such as infinite-length story generation and dynamic character composition.

Page — https://correr-zhou.github.io/IdentityStory/
Code — https://github.com/correr-zhou/IdentityStory
Extended version — https://arxiv.org/pdf/2512.23519

1 Introduction

Recent visual generative models (Rombach et al. 2022; Podell et al. 2023; Black Forest Labs 2024) enable users to create high-quality images from text, but they still struggle to maintain character consistency across multiple generated images due to their stochastic nature. This limitation promotes the task of story generation, which aims to yield a series of images with consistent characters solely using text. While this task has already made an impact in education (Carter 1993) and entertainment (Klimmt et al. 2012), it exhibits greater potential in human-centric scenarios such as film storyboarding (Hart 2013; Halligan 2013), advertisement design (Escalas 2003; Megehee and Woodside 2010), and artistic production (Cetinic and She 2022). Therefore, we aim to further advance **human-centric story generation** in this work, exploring story generation specifically with humans as characters.

Figure 1: Human-centric story generation. Our IdentityStory can *solely rely on text* to generate a series of images that consistently depict human characters and faithfully align with text prompts, outperforming the state-of-the-art method (Zhou et al. 2024b). Text prompts are omitted due to space constraints. Please refer to our extended version for details. *Zoom in for better view.*

As shown in Figure 1, this task adopts a text prompt set where each prompt shares the same character descriptions as input, aiming to generate multiple images with consistent characters and the corresponding visual content. Human-centric story generation is more challenging compared to story generation on other subjects, since (i) human faces inherently contain richer details (Wang and Deng 2021) and exhibit diverse variations (Zhou et al. 2021; Kammoun et al. 2022), imposing greater difficulty on consistency maintenance, and (ii) the coordination of multiple characters across different images demands a more flexible generation framework. Existing state-of-the-art methods of story generation cannot achieve satisfactory results (Figure 1&6), as they rely on attention-sharing mechanisms (*e.g.*, ConsiStory (Tewel et al. 2024), Story Diffusion (Zhou et al. 2024b)) or global semantic modulation (*e.g.*, Story-Adapter (Mao et al. 2024), 1Prompt1Story (Liu et al. 2025)), lacking the ability to precisely and flexibly maintain the cross-image identity consistency of human characters.

Recently, identity-preserving generators (Li et al. 2024; Ye et al. 2023; Wang et al. 2024a) have emerged within

*Equal contribution.
†Corresponding authors.

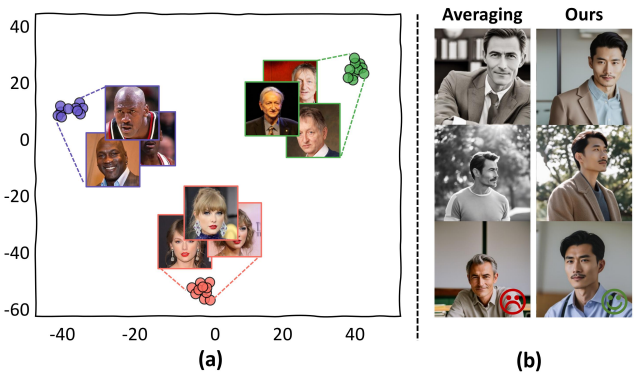

Figure 2: Identity embeddings of identity-preserving generators. Here we use PhotoMaker (Li et al. 2024) as an example: (a) We extract embeddings of three identities using PhotoMaker's image encoder and visualize them in 2D with t-SNE (Van der Maaten and Hinton 2008), observing that *embeddings of different identities are highly distinguishable*. (b) We collect 20 character descriptions and extract identity embeddings using the naive averaging approach and our method. Then, each embedding generates 15 images for computing pairwise face similarity with ArcFace (Deng et al. 2019), showing our method performs better.

Figure 3: Text alignment degradation of identity-preserving generators. We present qualitative results of vanilla SDXL (Podell et al. 2023), PhotoMaker (Li et al. 2024), and InstantID (Wang et al. 2024a). The prompt for the case above is "*A surreal digital artwork of a young Black girl with curly hair and bright eyes, standing on a rainbow bridge high above the clouds*", and the prompt for the case below is "*A hyper-realistic illustration of a middle-aged Latina woman with dark, curly hair and strong cheekbones, wearing an apron in a cozy home kitchen*".

the task of subject-driven generation (requires reference images as input) (Gal et al. 2022; Ruiz et al. 2023; Rout et al. 2024; Zhou et al. 2024a) to ensure identity preservation in generated images. Therefore, an intriguing question arises: *Can we harness identity-preserving generators for human-centric story generation?* In this work, we present **IdentityStory**, a framework designed to unleash the full potential of identity-preserving generators for this task, ensuring precise and flexible consistency maintenance for human characters across multiple images (Figure 1).

To achieve this, we first develop *Iterative Identity Discovery* to extract identities (Figure 4(a)). We find that identity-preserving generators process a well-constructed identity space (Figure 2(a)), where identity representation can be obtained by aggregating character image embeddings. After generating diverse character images from descriptions and projecting them into the identity space, we use Singular Value Decomposition (SVD) to iteratively filter out low-relevance embeddings and extract cohesive identities.

Next, we design *Re-denoising Identity Injection* to inject identities (Figure 4(b)). To address text alignment degradation of identity-preserving generators (Figure 3), we first use a general generator to create a more text-aligned prototype image. Meanwhile, we cache noisy images during generation to preserve environmental semantics and segment the prototype image to extract character layouts. Using a progressive masking strategy, we then re-denoise with identity-preserving generators to inject identities. We perform experiments based on *ConsiStory-Human*, a new benchmark improved from (Tewel et al. 2024), demonstrating that IdentityStory achieves overall superior performance and especially excels at precise face consistency maintenance. Notably, our method achieves a pairwise face similarity score of 55.5%, over double that of the second-best method (27.1%). Moreover, we show the practicality of IdentityStory to more applications such as community tool integration, infinite-length story generation, and dynamic character composition.

2 Related Works

Story Generation is first formulated as the task of story visualization by StoryGAN (Li et al. 2019) and has evolved across diverse technical paradigms such as generative adversarial networks (Li et al. 2019; Li, Kong, and Zhou 2020), large language-vision models (Shen and Elhoseiny 2023; Yang et al. 2024), and diffusion models (Tewel et al. 2024; Zhou et al. 2024b; Mao et al. 2024; Liu et al. 2025; Song et al. 2025). Early works (Li, Kong, and Zhou 2020; Maharana, Hannan, and Bansal 2022) tend to train models on close-domain datasets. While these methods manage to reproduce specified characters, the lack of generalization severely hinders their ability to handle unseen subjects or adapt to various contexts. Building upon text-to-image (T2I) diffusion models, subsequent works (Tewel et al. 2024; Zhou et al. 2024b; Mao et al. 2024; Liu et al. 2025) succeed in generating consistent characters of open domains. However, these approaches involve either intricate module assembly, complex strategy designs, or heavy memory usage, thus restricting their applicability in practice. More importantly, they failed to precisely yield consistent human characters, while our IdentityStory excels in this aspect by unlocking the capabilities of identity-preserving generators.

Subject-Driven Generation aims to enable T2I diffusion models to generate specific visual concepts from reference images. Initial approaches (Gal et al. 2022; Ruiz et al. 2023) achieved this by optimizing text embeddings or fine-tuning model weights. Subsequent methods (Kumari et al. 2023; Jiang et al. 2024; Shi et al. 2024; Zhou et al. 2024a) expanded on these approaches to handle multiple visual con-

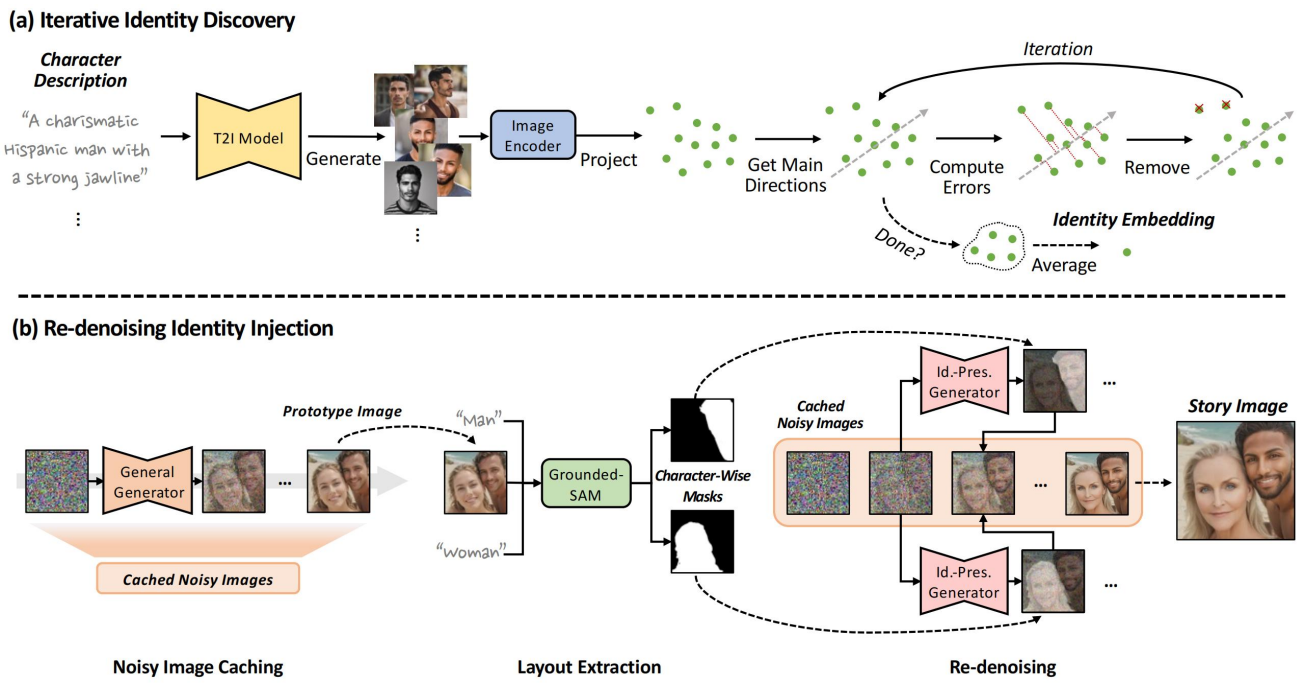

Figure 4: Pipeline of IdentityStory. This framework consists of two key techniques, including (a) *Iterative Identity Discovery* (Section 3.1), which utilizes Singular Value Decomposition (SVD) to iteratively filter out low-relevance embeddings to extract cohesive identities, and (b) *Re-denoising Identity Injection* (Section 3.2), which uses identity-preserving generators to inject extracted identities based on the noisy cached images and character layouts produced from general generators.

cepts. Recently, learning-based approaches (Wei et al. 2023; He et al. 2024; Tan et al. 2024; Cai et al. 2024) have gained widespread adoption due to their strong zero-shot capabilities. Within these, a subset of methods (Li et al. 2024; Wang et al. 2024a; Xiao et al. 2024; Wang et al. 2024b; Guo et al. 2025), which we refer to as identity-preserving generators, specifically focus on generating human-centric images. With additional training on large human datasets, these methods effectively preserve the identity of individuals depicted in reference images (Wei et al. 2025). In this work, our method unlocks their potential and goes beyond their original functionality, enabling them to support text-only input and achieve better text alignment.

3 Methodology

In this work, we explore *human-centric story generation*, aiming to generate visually coherent story images with human characters (Figure 1). Specifically, the input is a story text $\mathcal{T} = \{T_i\}_{i=1}^{n}$ with n text prompts, where each text prompt T_i corresponds to a story image and shares the same character descriptions $\{c_i\}_{i=1}^{s}$ with s characters. The goal is to generate a series of images $\mathcal{I} = \{I_i\}_{i=1}^{n}$ with consistent characters solely using $\mathcal{T} = \{T_i\}_{i=1}^{n}$. We develop **IdentityStory**, aiming to unleash the potential of identity-preserving generators for human-centric story generation (Figure 4). In this framework, *Iterative Identity Discovery* (Sec. 3.1) is first conducted to extract identity embeddings solely from character descriptions via iterative filtering, and then *Re-denoising Identity Injection* (Sec. 3.2) is employed to generate story images by injecting identity embeddings with a re-denoising paradigm. In the following, we delve deeper into the details of these two techniques.

3.1 Iterative Identity Discovery

We observe that identity-preserving generators contain a well-structured identity space (Figure 2(a)), making it possible to obtain an identity embedding by aggregating the embeddings of generated character images. However, simply averaging the embeddings results in low identity uniqueness (Figure 2(b)) due to the diversity of generated images, which reflect various identities. Therefore, we develop *Iterative Identity Discovery* (Figure 4(a)), which leverages Singular Value Decomposition (SVD) to iteratively filter out low-relevance embeddings and ultimately aggregate the most cohesive ones, yielding a more unique identity representation (Figure 2(b)). Below we describe it in detail.

Overall Scheme. We first generate m character images for each character description c_i. Using the image encoder of the identity-preserving generator G_{id}, these images are then projected into the identity space, which results in a combined embedding matrix $E \in \mathbb{R}^{m \times d}$ where d is the embedding dimension. Inspired by (Gu et al. 2014), we obtain the implicit semantic information of E by applying SVD as

$$E = U \Sigma V^T, \tag{1}$$

where $U \in \mathbb{R}^{m \times m}$, $\Sigma \in \mathbb{R}^{m \times d}$, and $V \in \mathbb{R}^{d \times d}$. As a result, V contains right singular vectors of E, where each column denotes a semantic direction of the identity space.

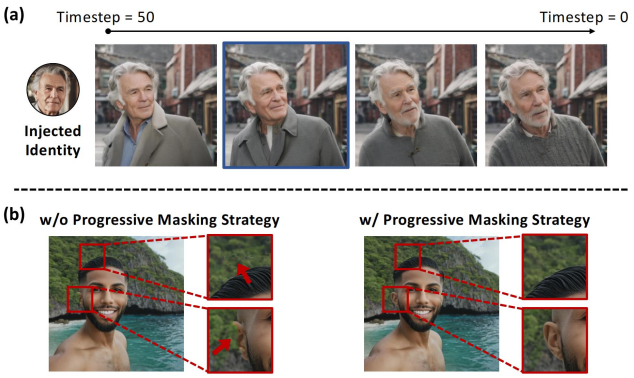

Figure 5: Design choices of Re-denoising Identity Injection. (a) We start from a sweet-spot timestep ($t' = 40$) to re-denoise, balancing image harmony and identity fidelity (*blue frame*). (b) We develop a progressive masking strategy to effectively diminish artifacts (*red frame*).

Next, we select the directions with the top k singular values from V, formulating $V_k \in \mathbb{R}^{d\times k}$. The directions in V_k capture the most significant semantics of the embeddings, while the others involve low-relevance semantics or noises. Then, we further compute the reconstruction matrix $W = V_k V_k^T$, where $W \in \mathbb{R}^{d\times d}$ can be regarded as a transformation designed to preserve only the most essential semantics. Using W, we can calculate the reconstruction errors $\varepsilon \in \mathbb{R}^m$ for m identity embeddings as

$$\varepsilon = \frac{1}{d}\sum_{i=1}^{d} \|E_{:,i} - (EW)_{:,i}\|_2, \tag{2}$$

where $\{:, i\}$ represents the matrix indexing. Low-relevance embeddings can deviate from the core semantics, thereby leading to larger errors due to insufficient reconstruction. With a filtering ratio r and an iteration number p, we remove the embeddings with the $(1-r)\cdot m$ largest errors from E, and return to Equation 1 for the next iteration. Through this iterative process, about 21.6% of embeddings are retained and averaged to form the final identity embedding, effectively capturing the core semantics of the character while minimizing the influence of non-identity factors such as poses and facial expressions. Such a technique can extract a more reliable identity representation ensuring character consistency.

3.2 Re-denoising Identity Injection

While identity-preserving generators excel at maintaining identities in generated images (Li et al. 2024; Ye et al. 2023; Wang et al. 2024a), they exhibit suboptimal performance on text alignment. This limitation affects their ability to generate desired visual elements and reasonable character layouts, as illustrated in Figure 3(a). To address this, we design *Re-denoising Identity Injection* (Figure 4(b)), utilizing the complementary strengths of general generators and identity-preserving generators. Specifically, the general generator provides high-quality environmental details and character layouts, while the identity-preserving generator focuses on injecting identities. This approach involves three processes: noisy image caching, layout extraction, and re-denoising, which are detailed as follows.

Noisy Image Caching. First, we utilize a general generator G (*e.g.*, the base model), which is better in text alignment (Figure 3), to generate template images containing faithful visual semantics for subsequent processes. For a story image I_i, we use its text prompt T_i to generate

$$I'_i = G(T_i), \tag{3}$$

where I'_i is the corresponding template image. During the sampling of G, we cache the noisy images of all t timesteps, which can be formulated as

$$\mathbf{z} = \{z_t, z_{t-1}, ..., z_1, z_0\}, \tag{4}$$

where $z_i \in \mathbb{R}^{l\times l}$ is the noisy image at i-th timestep, l is the latent dimension, and t is set to 50 for a DDIM (Song, Meng, and Ermon 2020) scheduler. These predicted noisy images implicitly encapsulate the contextual information for generating the template image, serving as the basis for the re-denoising phase.

Layout Extraction. To obtain proper foreground layouts, previous works often rely on user inputs (Bar-Tal et al. 2023; Avrahami et al. 2023; Xie et al. 2023) or large language models (Phung, Ge, and Huang 2024; Feng et al. 2024). In contrast, we directly extract the character layout from the template image I'_i, as it is generated using the prompt T_i, which aligns well with our intended design. This guarantees that the resultant layout is precise and natural. Based on character words in T_i such as "man" and "woman", we use Grounded-SAM (Ren et al. 2024) to segment out character-wise masks as

$$\mathbf{M} = \{M_1, M_2, ..., M_s\}, \tag{5}$$

where the background mask can be set as $M_{\text{bg}} = \complement(M_1 \cup M_2 ... M_s)$. Using these masks, we can accurately separate each character from the background, allowing for precise identity injection while preserving the original and unbiased environmental details during re-denoising.

Re-denoising. Using the cached noisy images and the character-wise masks, we then inject the identity of characters by performing re-denoising with the identity-preserving generator G_{id}. We start from a timestep of $t' = 40$, a sweet spot that provides a better balance between image harmony and identity fidelity (Figure 5(a)). For the noisy image z_i, we use the denoising network ϵ_{id} of G_{id} to perform

$$z_{i-1,j} = \epsilon_{\text{id}}(c_j, e_j, z_i), \quad j = 1, ..., s, \tag{6}$$

where $z_{i,j}$ is the noisy image for j-th character at the next timestep, which is guided by the identity embedding e_j and the character description c_j. Then, z_i is over-written as

$$z_{i-1} \leftarrow \text{DS}(M_{\text{bg}}) \odot z_{i-1} + \sum_{j=1}^{s} \text{DS}(M_j) \odot z_{i-1,j}, \tag{7}$$

where $\text{DS}(\cdot)$ is a down-sample operation to match the shape of noisy images. Fixed masks could constrain the scope of re-denoising and make it difficult to adequately reconcile local details, leading to undesired artifacts (Figure 5(b)). Thus,

Methods	Text Alignment		Character Consistency		Image Quality	
	CLIP-T↑ (%)	CLIP-T-C↑ (%)	CLIP-I↑ (%)	Face-Sim↑ (%)	Q-Align-Aes↑	Q-Align-Gen↑
ConsiStory (Tewel et al. 2024)	**35.5**	30.1	78.2	17.1	3.75	4.71
StoryDiffusion (Zhou et al. 2024b)	34.0	<u>30.7</u>	<u>85.2</u>	<u>27.1</u>	3.58	4.20
Story-Adapter (Mao et al. 2024)	34.3	29.1	76.6	23.9	3.65	4.42
1Prompt1Story (Liu et al. 2025)	34.9	29.7	79.8	23.5	<u>4.16</u>	<u>4.81</u>
IdentityStory (Ours)	<u>35.4</u>	**31.1**	**85.8**	**55.5**	**4.25**	**4.92**

Table 1: Quantitative comparison. The results of automatic metrics demonstrate IdentityStory's overall superior performance, especially in *face similarity (Face-Sim)*. The best and second-best results are marked in **bold** and <u>underlined</u>.

Figure 6: Qualitative comparison. Compared to other methods, our IdentityStory exhibits remarkable performance in handling human-centric scenarios, enabling consistent generation of human characters with only text as input. *Zoom in for better view.*

we further develop a progressive masking strategy, updating M_j at each timestep i as

$$M_j \leftarrow \text{Dilate}(M_j, K_i), \quad j = 1, ..., s, \tag{8}$$

where $\text{Dilate}(M_j, K_i)$ denotes dilating the mask M_j with the kernel size $K_i = \frac{i-t'}{t-t'} \cdot K_{\max}$ and $K_{\max}$ is the predefined maximum kernel size. This strategy enables re-denoising to progressively refine more local details, effectively eliminating unwanted artifacts (Figure 5(b)).

4 Experiments

4.1 Setups

Implementation Details. We select PhotoMaker (Li et al. 2024) as our identity-preserving generator and SDXL (Podell et al. 2023) as the base model. For Iterative Identity Discovery, we set the character image number $m = 64$, the filtering ratio $r = 60\%$, and the iteration number $p = 3$. For Re-denoising Identity Injection, we set the initial timestep $t' = 40$ and the maximum kernel size $M_{\max} = 50$.

Benchmark and Compared Methods. We present *ConsiStory-Human*, an enhanced version of the ConsiStory benchmark (Tewel et al. 2024), which is designed specifically for human-centric scenarios. It features diverse characters and story descriptions, comprising 100 prompt sets, each with 10 text prompts. Using ConsiStory-Human, each method generates 1,000 images for a thorough evaluation. We compare our method with state-of-the-art story generation methods, including ConsiStory (Tewel et al. 2024), StoryDiffusion (Zhou et al. 2024b), Story-Adapter (Mao et al. 2024), and 1Prompt1Story (Liu et al. 2025).

Evaluation Metrics. We evaluate methods based on the following aspects: (i) *Text alignment*: We compute the average CLIP (Radford et al. 2021) score between each generated story image and its corresponding text prompt, denoted as CLIP-T. Additionally, we segment characters from each story image and compute the average CLIP score between each character image and its character description, denoted as CLIP-T-C. (ii) *Character consistency*: We measure the average similarity between segmented character images within each prompt set using CLIP (Radford et al. 2021), referred to as CLIP-I. Moreover, we compute the average pairwise face similarity with ArcFace (Deng et al. 2019), denoted as Face-Sim. (iii) *Image quality*: We assess both the aesthetic quality and general quality of the generated story images using Q-Align (Wu et al. 2023), with scores represented as Q-Align-Aes and Q-Align-Gen.

4.2 Experimental Results

Quantitative Comparison. We present the quantitative results in Table 1. In terms of text alignment, our method achieves the best CLIP-I-C and performs just slightly below ConsiStory in CLIP-I. However, the better CLIP-T of ConsiStory comes at the cost of its poor character consistency. In contrast, our method shows remarkable performance in character consistency, particularly in face similarity. Specif-

Methods	Text Align.↑	Char. Consis.↑	Img. Qual.↑
ConsiStory (Tewel et al. 2024)	81.88	<u>78.58</u>	82.91
StoryDiffusion (Zhou et al. 2024b)	80.74	77.52	85.47
Story-Adapter (Mao et al. 2024)	81.08	77.00	83.07
1Prompt1Story (Liu et al. 2025)	<u>82.20</u>	78.31	<u>86.48</u>
IdentityStory (Ours)	**84.41**	**82.83**	**88.74**

Table 2: MLLM-as-a-judge evaluation. The proposed IdentityStory achieves the highest average across all three metrics, further showcasing its superior performance.

Methods	Text Align.↑	Char. Consis.↑	Img.Qual.↑
ConsiStory (Tewel et al. 2024)	6.5%	6.5%	8.5%
StoryDiffusion (Zhou et al. 2024b)	<u>10.3%</u>	<u>11.0%</u>	<u>14.3%</u>
Story-Adapter (Mao et al. 2024)	5.0%	5.2%	3.8%
1Prompt1Story (Liu et al. 2025)	9.0%	10.8%	8.7%
IdentityStory (Ours)	**69.2%**	**66.5%**	**64.7%**

Table 3: User study. The selection rates on the three metrics clearly indicate that IdentityStory outperforms other methods in terms of human preference.

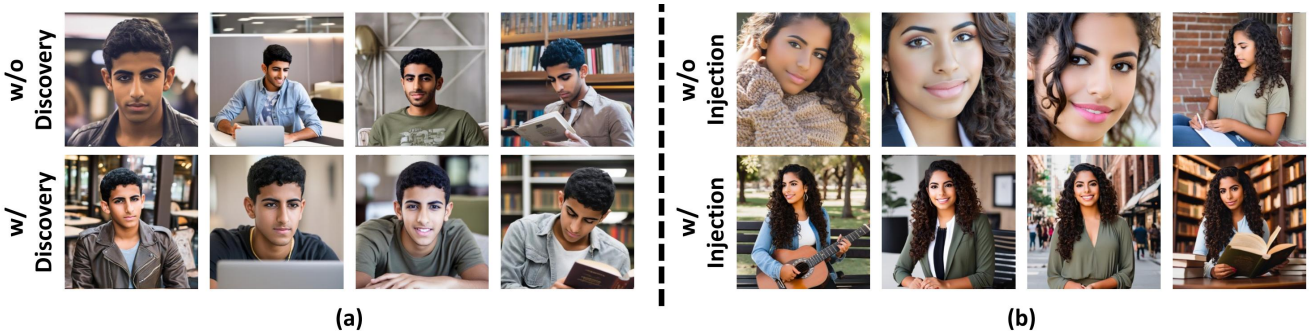

Figure 7: Ablation of key techniques. We demonstrate the effectiveness of (a) Iterative Identity Discovery, which improves character consistency, and (b) Re-denoising Identity Injection, which enhances text alignment.

ically, we reach a Face-Sim of 55.5%, more than double the second-best of 21.1%. Moreover, Q-Align indicates that our method can generate top-quality images, excelling in both general and aesthetic aspects.

Qualitative Comparison. The qualitative results are illustrated in Figure 6. Our method exhibits superior performance in preserving character consistency, especially for fine-grained facial features, across different images while faithfully following text prompts. However, other methods struggle with human-centric scenarios, as they rely on attention or prompts to maintain consistency. Due to its flexibility, our method can also effectively handle multiple characters, whereas other methods cause identity disorder or even blending. Notably, they also display inconsistencies in style, indicating their instability.

MLLM-as-a-Judge Evaluation. To conduct additional evaluation, we utilize GPT-4o (Hurst et al. 2024) to score the generated images. For each case, GPT-4o is provided with the story text alongside the five outputs from other methods and our IdentityStory. We then instruct GPT-4o to independently assign a score between 0 and 100 to each generated image based on three criteria, including text alignment, character consistency, and image quality. Finally, we average its ratings for each method and criterion, presenting the results in Table 2. As we can observe, IdentityStory achieves the highest averages across all aspects, further demonstrating its superior performance in text alignment, character consistency, and image quality compared to other methods.

User Study. To evaluate human preference, we design a questionnaire showing 20 randomly selected groups of generated images, each paired with the corresponding story text. Participants were asked to select the best result in each group in terms of text alignment, character consistency, and image quality. In total, we collected 30 responses and the averaged selection rates are reported in Table 3. The results indicate that IdentityStory outperforms others across all aspects, showing that its generated images are better aligned with human preference.

4.3 Ablation Studies

Key Techniques. We ablate two key techniques of IdentityStory to demonstrate their significance. When removing Iterative Identity Discovery, we adopt the average identity embedding of the generated character images for the subsequent process, showing that it contributes to superior character consistency (Figure 7(a)). When removing Re-denoising Identity Injection, we use the identity-preserving generator to directly produce final results, verifying that it helps achieve better text alignment (Figure 7(b)).

Iterative Strategy of Iterative Identity Discovery. In Figure 2(b), we have demonstrated that our Iterative Identity Discovery can achieve better identity embeddings compared to the naive averaging approach. In Figure 8(a), we follow the same setting and further adjust the iterative numbers of Iterative Identity Discovery, showing the iterative strategy contributes to a better face similarity result.

Injection Mechanism of Re-denoising Identity Injection. We adopt a layout-guided mechanism to precisely inject identities into cached noisy images while preserving environmental details. To evaluate its effectiveness, we abandon the layout guidance and directly average cached noisy images and the identity generator's output to perform iden-

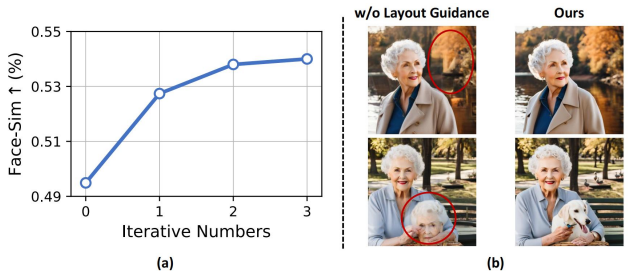

Figure 8: (a) Ablation of the iterative strategy of Iterative Identity Discovery. Face similarity (Face-Sim) increases steadily with iterative numbers, showing the effectiveness of the iterative strategy. (b) Ablation of the injection mechanism of Re-denoising Identity Injection. Our layout-guided mechanism can effectively inject identities while maintaining high-quality environmental details.

Figure 9: Infinite-length story generation with IdentityStory. Our method can conveniently generate a series of images with consistent characters at any length.

tity injection. This paradigm results in disordered visual elements, as it fails to prevent undesired semantics from being mixed into the cached noisy images (Figure 8(b)).

4.4 More Applications

Integration with Other Community Tools. IdentityStory can seamlessly integrate with other community tools to accommodate a wider range of use scenarios (Figure 10). For instance, IdentityStory can enable face pose controls with ControlNet (Zhang, Rao, and Agrawala 2023), and collaborate with stylized base models (SG161222 2024; Cagliostro Research Lab 2024) for enhancing the generation of special styles, highlighting its flexibility for diverse user needs.

Infinite-Length Story Generation. Decoupling ID extraction from generation, infinitely extending a story, which is one of the key advantages of IdentityStory, becomes feasible by reusing the extracted ID embeddings. Traditional methods are constrained by memory consumption (Tewel et al. 2024; Zhou et al. 2024b; Mao et al. 2024) or text length limitations (Liu et al. 2025), while our method only requires extracting the identity once, after which it can continuously generate images with the same character (Figure 9).

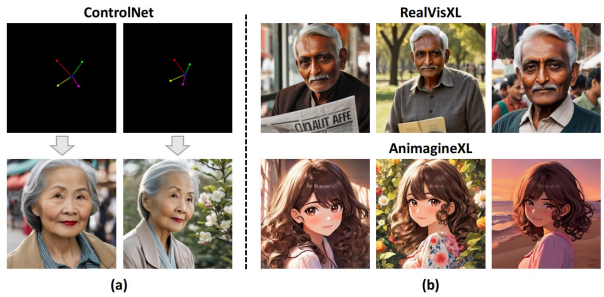

Figure 10: Combining IdentityStory with community tools, IdentityStory can collaborate with community tools, such as (a) ControlNet (Zhang, Rao, and Agrawala 2023) and (b) stylized base models (RealVisXL (SG161222 2024) and AnimagineXL (Cagliostro Research Lab 2024)).

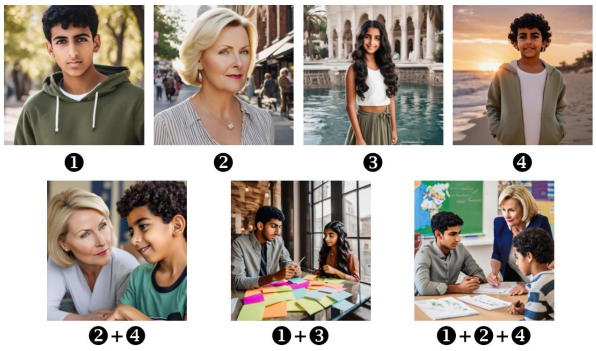

Figure 11: Dynamic character composition with IdentityStory. Our method can dynamically combine characters in different story images.

Dynamic Character Composition. IdentityStory can also dynamically combine characters across different images within a story, as shown in Figure 11. This capability allows characters to be flexibly introduced or combined across various scenes, which is highly practical for handling complex narratives without losing coherence or visual harmony.

5 Conclusion

In this work, we introduce *human-centric story generation* and develop *IdentityStory*, which combines *Iterative Identity Discovery* to extract identities via iterative filtering and *Re-denoising Identity Injection* to inject identities via re-denoising. Experiments show that IdentityStory sets a new standard for this task. Looking ahead, we plan to extend our framework to include broader image and video generation, further enhancing the creativity of visual content. Additionally, we aim to explore the integration of multi-modal inputs, such as audio and textual cues, to enrich the storytelling process and create more immersive human-centric narratives.

Acknowledgments

This study was supported in part by the InnoHK initiative of the Innovation and Technology Commission of the Hong

Kong Special Administrative Region Government via the Hong Kong Centre for Logistics Robotics, by the Faculty Initiatives Research of Monash University (Contract No. 2901912), by the NVIDIA Academic Hardware Grant Program, and by the Research Start-up Fund for Prof. Xiaowei Hu at the Guangzhou International Campus, South China University of Technology (Grant No. K3250310).

References

Avrahami, O.; Hayes, T.; Gafni, O.; Gupta, S.; Taigman, Y.; Parikh, D.; Lischinski, D.; Fried, O.; and Yin, X. 2023. Spatext: Spatio-textual representation for controllable image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18370–18380.

Bar-Tal, O.; Yariv, L.; Lipman, Y.; and Dekel, T. 2023. MultiDiffusion: Fusing Diffusion Paths for Controlled Image Generation. In *International Conference on Machine Learning*, 1737–1752. PMLR.

Black Forest Labs. 2024. FLUX. https://github.com/black-forest-labs/flux.

Cagliostro Research Lab. 2024. Animagine XL 3.0.

Cai, S.; Chan, E.; Zhang, Y.; Guibas, L.; Wu, J.; and Wetzstein, G. 2024. Diffusion self-distillation for zero-shot customized image generation. *arXiv preprint arXiv:2411.18616*.

Carter, K. 1993. The place of story in the study of teaching and teacher education. *Educational researcher*, 22(1): 5–18.

Cetinic, E.; and She, J. 2022. Understanding and creating art with AI: Review and outlook. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(2): 1–22.

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.

Escalas, J. E. 2003. Advertising narratives: what are they and how do they work? In *Representing consumers*, 283–305. Routledge.

Feng, Y.; Gong, B.; Chen, D.; Shen, Y.; Liu, Y.; and Zhou, J. 2024. Ranni: Taming text-to-image diffusion for accurate instruction following. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4744–4753.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-or, D. 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *The Eleventh International Conference on Learning Representations*.

Gu, S.; Zhang, L.; Zuo, W.; and Feng, X. 2014. Weighted nuclear norm minimization with application to image denoising. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2862–2869.

Guo, Z.; Wu, Y.; Zhuowei, C.; Zhang, P.; He, Q.; et al. 2025. Pulid: Pure and lightning id customization via contrastive alignment. *Advances in Neural Information Processing Systems*, 37: 36777–36804.

Halligan, F. 2013. *Movie Storyboards: The art of visualizing screenplays*. Chronicle Books.

Hart, J. 2013. *The Art of the Storyboard: A filmmaker's introduction*. Routledge.

He, J.; Li, H.; Hu, Y.; Shen, G.; Cai, Y.; Qiu, W.; and Chen, Y.-C. 2024. DisEnvisioner: Disentangled and Enriched Visual Prompt for Customized Image Generation. *arXiv preprint arXiv:2410.02067*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

Jiang, J.; Zhang, Y.; Feng, K.; Wu, X.; Li, W.; Pei, R.; Li, F.; and Zuo, W. 2024. MC ^{2}: Multi-concept Guidance for Customized Multi-concept Generation. *arXiv preprint arXiv:2404.05268*.

Kammoun, A.; Slama, R.; Tabia, H.; Ouni, T.; and Abid, M. 2022. Generative adversarial networks for face generation: A survey. *ACM Computing Surveys*, 55(5): 1–37.

Klimmt, C.; Roth, C.; Vermeulen, I.; Vorderer, P.; and Roth, F. S. 2012. Forecasting the Experience of Future Entertainment Technology: "Interactive Storytelling" and Media Enjoyment. *Games and Culture*, 7(3): 187–208.

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

Li, C.; Kong, L.; and Zhou, Z. 2020. Improved-storygan for sequential images visualization. *Journal of Visual Communication and Image Representation*, 73: 102956.

Li, Y.; Gan, Z.; Shen, Y.; Liu, J.; Cheng, Y.; Wu, Y.; Carin, L.; Carlson, D.; and Gao, J. 2019. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Li, Z.; Cao, M.; Wang, X.; Qi, Z.; Cheng, M.-M.; and Shan, Y. 2024. PhotoMaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8640–8650.

Liu, T.; Wang, K.; Li, S.; van de Weijer, J.; Khan, F. S.; Yang, S.; Wang, Y.; Yang, J.; and Cheng, M.-M. 2025. One-Prompt-One-Story: Free-Lunch Consistent Text-to-Image Generation Using a Single Prompt. In *The Thirteenth International Conference on Learning Representations*.

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

Mao, J.; Huang, X.; Xie, Y.; Chang, Y.; Hui, M.; Xu, B.; and Zhou, Y. 2024. Story-Adapter: A Training-free Iterative Framework for Long Story Visualization. *arXiv preprint arXiv:2410.06244*.

Megehee, C. M.; and Woodside, A. G. 2010. Creating visual narrative art for decoding stories that consumers and brands tell. *Psychology & Marketing*, 27(6): 603–622.

Phung, Q.; Ge, S.; and Huang, J.-B. 2024. Grounded text-to-image synthesis with attention refocusing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7932–7942.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *The Twelfth International Conference on Learning Representations*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; Huang, X.; Chen, Y.; Yan, F.; et al. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*.

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

Rout, L.; Chen, Y.; Ruiz, N.; Kumar, A.; Caramanis, C.; Shakkottai, S.; and Chu, W.-S. 2024. RB-Modulation: Training-Free Personalization of Diffusion Models using Stochastic Optimal Control. *arXiv preprint arXiv:2405.17401*.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

SG161222. 2024. RealVisXL V4.0.

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

Shi, Q.; Qi, L.; Wu, J.; Bai, J.; Wang, J.; Tong, Y.; Li, X.; and Yang, M.-H. 2024. RelationBooth: Towards Relation-Aware Customized Object Generation. *arXiv preprint arXiv:2410.23280*.

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

Song, Q.; Zhou, D.; Lin, J.; Shen, F.; Wang, J.; Hu, X.; Chen, C.; and Heng, P.-A. 2025. SceneDecorator: Towards Scene-Oriented Story Generation with Scene Planning and Scene Consistency. *arXiv preprint arXiv:2510.22994*.

Tan, Z.; Liu, S.; Yang, X.; Xue, Q.; and Wang, X. 2024. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 3.

Tewel, Y.; Kaduri, O.; Gal, R.; Kasten, Y.; Wolf, L.; Chechik, G.; and Atzmon, Y. 2024. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, M.; and Deng, W. 2021. Deep face recognition: A survey. *Neurocomputing*, 429: 215–244.

Wang, Q.; Bai, X.; Wang, H.; Qin, Z.; Chen, A.; Li, H.; Tang, X.; and Hu, Y. 2024a. InstantID: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*.

Wang, Q.; Jia, X.; Li, X.; Li, T.; Ma, L.; Zhuge, Y.; and Lu, H. 2024b. Stableidentity: Inserting anybody into anywhere at first sight. *arXiv preprint arXiv:2401.15975*.

Wei, Y.; Zhang, Y.; Ji, Z.; Bai, J.; Zhang, L.; and Zuo, W. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15943–15953.

Wei, Y.; Zheng, Y.; Zhang, Y.; Liu, M.; Ji, Z.; Zhang, L.; and Zuo, W. 2025. Personalized Image Generation with Deep Generative Models: A Decade Survey. *arXiv preprint arXiv:2502.13081*.

Wu, H.; Zhang, Z.; Zhang, W.; Chen, C.; Liao, L.; Li, C.; Gao, Y.; Wang, A.; Zhang, E.; Sun, W.; et al. 2023. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*.

Xiao, G.; Yin, T.; Freeman, W. T.; Durand, F.; and Han, S. 2024. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, 1–20.

Xie, J.; Li, Y.; Huang, Y.; Liu, H.; Zhang, W.; Zheng, Y.; and Shou, M. Z. 2023. BoxDiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7452–7461.

Yang, S.; Ge, Y.; Li, Y.; Chen, Y.; Ge, Y.; Shan, Y.; and Chen, Y. 2024. Seed-story: Multimodal long story generation with large language model. *arXiv preprint arXiv:2407.08683*.

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

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

Zhou, D.; Huang, J.; Bai, J.; Wang, J.; Chen, H.; Chen, G.; Hu, X.; and Heng, P.-A. 2024a. MagicTailor: Component-Controllable Personalization in Text-to-Image Diffusion Models. *arXiv preprint arXiv:2410.13370*.

Zhou, T.; Wang, W.; Liang, Z.; and Shen, J. 2021. Face forensics in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5778–5788.

Zhou, Y.; Zhou, D.; Cheng, M.-M.; Feng, J.; and Hou, Q. 2024b. StoryDiffusion: Consistent self-attention for long-range image and video generation. *Advances in Neural Information Processing Systems*, 37: 110315–110340.