# Infinite-Story: A Training-Free Consistent Text-to-Image Generation

**Jihun Park*, Kyoungmin Lee*, Jongmin Gim*, Hyeonseo Jo, Minseok Oh, Wonhyeok Choi, Kyumin Hwang, Jaeyeul Kim, Minwoo Choi, Sunghoon Im†**

DGIST, South Korea

{pjh2857, kyoungmin, jongmin4422, gustj0510, harrymark0, smu06117, kyumin, jykim94, subminu, sunghoonim}@dgist.ac.kr

## Abstract

We present *Infinite-Story*, a training-free framework for consistent text-to-image (T2I) generation tailored for multi-prompt storytelling scenarios. Built upon a scale-wise autoregressive model, our method addresses two key challenges in consistent T2I generation: identity inconsistency and style inconsistency. To overcome these issues, we introduce three complementary techniques: *Identity Prompt Replacement*, which mitigates context bias in text encoders to align identity attributes across prompts; and a unified attention guidance mechanism comprising *Adaptive Style Injection* and *Synchronized Guidance Adaptation*, which jointly enforce global style and identity appearance consistency while preserving prompt fidelity. Unlike prior diffusion-based approaches that require fine-tuning or suffer from slow inference, Infinite-Story operates entirely at test time, delivering high identity and style consistency across diverse prompts. Extensive experiments demonstrate that our method achieves state-of-the-art generation performance, while offering over $6\times$ faster inference (1.72 seconds per image) than the existing fastest consistent T2I models, highlighting its effectiveness and practicality for real-world visual storytelling.

## Introduction

Large-scale diffusion-based Text-to-Image (T2I) generation models (Rombach et al. 2022; Saharia et al. 2022; Podell et al. 2023; Labs 2024) have demonstrated remarkable performance, establishing themselves as essential tools across a wide range of creative tasks, including design prototyping, content generation, visual communication, and advertising. However, the lack of consistency in generated images has posed limitations on user experience, particularly in scenarios that require coherence across multiple images, such as storytelling, character-driven content creation, comic strip generation, and sequential visual narratives.

To enforce consistency across generated images, various approaches have been proposed, including personalized image generation (Ruiz et al. 2023; Ye et al. 2023), style-aligned image generation (Park et al. 2025; Hertz et al. 2024; Sohn et al. 2023), and consistent text-to-image generation (Avrahami et al. 2024; Liu et al. 2025; Tewel et al. 2024; Wang et al. 2024; Zhou et al. 2024b). While consistent text-to-image generation is particularly foundational for visual storytelling tasks, prior works have largely focused on maintaining identity consistency across scenes. However, they often overlook style consistency between generated image sets, which is crucial for producing visually coherent narratives that span multiple scenes, as illustrated in Figure 2-(Top). In addition, most consistent text-to-image generation methods are based on diffusion models, which—even without fine-tuning—typically require over 10 seconds per image during inference. This surpasses the threshold at which users begin to lose focus during interactive sessions, according to Nielsen's usability guidelines (Nielsen 1994).

Recently, scale-wise autoregressive models (Tian et al. 2024; Voronov et al. 2024; Han et al. 2024) have emerged as a promising alternative, offering faster inference by adopting a next-scale prediction paradigm. These models achieve competitive image quality while significantly improving inference speed compared to both traditional autoregressive (Van Den Oord, Vinyals et al. 2017; Esser, Rombach, and Ommer 2021; Chang et al. 2022, 2023) and diffusion-based models (Podell et al. 2023; Labs 2024). While they effectively mitigate the latency issues inherent in diffusion approaches, scale-wise models continue to face challenges in ensuring consistency across generated images, such as identity inconsistency, style inconsistency, and a combination of both.

To address these challenges, we introduce *Infinite-Story*, a training-free framework for consistent text-to-image generation built on a scale-wise autoregressive model (Han et al. 2024), without modifying the architecture or requiring additional training. Our approach generates a set of images that retain consistency in both identity and style across varying prompts by designating one image in each batch as a reference and propagating its identity and style to guide the remaining samples.

To this end, we propose three lightweight yet effective techniques: *Identity Prompt Replacement*, which mitigates the context bias of text encoders to align identity-related attributes across prompts. Also, we propose a unified attention guidance that consists of *Adaptive Style Injection* and *Synchronized Guidance Adaptation*, enhancing both identity appearance and global visual style consistency via reference

*Equal contribution

†Corresponding author


Figure 1: Image sequences produced by Infinite-Story, maintaining identity and style consistency across varied prompts.


Figure 2: Qualitative comparison with 1Prompt1Story (Liu et al. 2025). While 1Prompt1Story preserves identity consistency, it struggles with style coherence across images. In contrast, Infinite-Story maintains both identity and style consistency, producing a unified and coherent visual narrative.

feature injection into early-stage self-attention layers, while ensuring prompt fidelity through synchronized adaptation across conditional and unconditional branches. These techniques are seamlessly integrated into the inference pipeline without any need for additional fine-tuning or training. By combining these components, Infinite-Story achieves state-of-the-art generation quality, as illustrated in Figure 1 and Figure 2-(Bottom). It outperforms existing methods in both quantitative and qualitative evaluations, while also offering up to 6× faster inference time (1.72 seconds per image) than the fastest diffusion-based consistent T2I models, as shown in Figure 3.

In summary, our primary contributions include:

- We present *Infinite-Story*, the first training-free, scale-wise autoregressive framework for consistent text-to-image generation.
- We introduce an *Identity Prompt Replacement* technique that aligns identity attributes across prompts by unifying identity prompt embeddings.
- We propose a unified attention guidance approach that combines *Adaptive Style Injection* and *Synchronized Guidance Adaptation* to achieve consistent overall style and identity appearance while preserving prompt fidelity.

## Related Work

### Text-to-image generation

Large-scale image-text datasets (Changpinyo et al. 2021; Lin et al. 2014; Schuhmann et al. 2022; Byeon et al. 2022) have enabled conditional image synthesis by bridging language and vision. This has spurred the development of powerful Text-to-Image (T2I) models—diffusion-based (Rombach et al. 2022; Saharia et al. 2022; Podell et al. 2023;

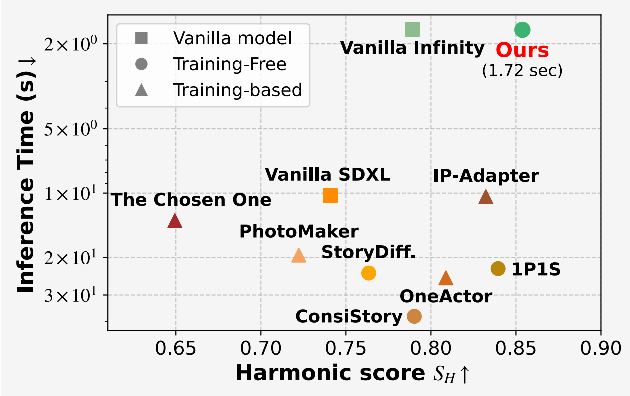


Figure 3: Inference time and harmonic score $S_H$ comparison with state-of-the-art consistent T2I models.

Labs 2024), GAN-based (Kang et al. 2023), and autoregressive (AR)-based (Chang et al. 2023; Han et al. 2024; Tang et al. 2024)—capable of producing high-quality images from text prompts. Diffusion models dominate with strong synthesis quality, supporting tasks like image editing (Brooks, Holynski, and Efros 2023; Hertz, Aberman, and Cohen-Or 2023; Wang et al. 2023) and translation (Tumanyan et al. 2023; Parmar et al. 2023), but suffer from slow inference. AR models have advanced from next-token prediction (Van Den Oord, Vinyals et al. 2017; Esser, Rombach, and Ommer 2021) to faster masked token generation (Chang et al. 2022, 2023; Kondratyuk et al. 2023), with next-scale prediction (Tian et al. 2024) improving efficiency further (Han et al. 2024; Tang et al. 2024; Voronov et al. 2024). Nonetheless, T2I models still struggle with maintaining subject identity consistency across images, limiting their applicability in areas like storytelling, content creation, and branding.

### Personalized image generation

Personalized image generation enables scene exploration with user-specific features. Existing methods are broadly categorized into subject-driven and style-driven approaches. Subject-driven methods (Gal et al. 2022; Ye et al. 2023; Ruiz et al. 2023) typically fine-tune or adapt pre-trained encoders to inject concept embeddings from reference images, but often require external datasets, limiting generality. Recent works address this with parameter-efficient fine-tuning by updating limited model components like attention layers (Nam et al. 2024; Kumari et al. 2023). Style-driven methods instead focus on visual consistency by optimizing style features via LoRA-based tuning (Frenkel et al. 2024; Shah et al. 2024; Sohn et al. 2023; Hu et al. 2022; Ryu 2022) or by adapting attention for stylistic coherence (Hertz et al. 2024; Park et al. 2025). Despite their strengths, most methods rely on diffusion models, which are slow and unsuitable for interactive use.

### Consistent text-to-image generation

Consistent text-to-image generation, which aims to preserve identity across multiple images, has become a key focus within personalized image generation. Recent studies (Kumari et al. 2023; Li et al. 2024; Zhou et al. 2024b; Tewel et al. 2024) show that adjusting attention layer weights effectively modulates identity. Other approaches (Mou et al. 2023) incorporate structured control to aid identity preservation. Foundational works (Radford et al. 2021; Chen et al. 2025) highlight the linguistic strength of transformer-based text encoders, while enhanced textual conditioning (Hertz et al. 2022; Gal et al. 2022) further improves identity consistency. Building on this, (Liu et al. 2025) leverage prompt embedding variations to maintain coherent identities across images. Inspired by these insights, we introduce a training-free consistent text-to-image generation method through the manipulation of prompt embeddings and attention mechanisms.

## Method

### Overall pipeline

In this paper, we aim to generate $N$ multiple images $\mathbf{I} = \{I^n\}_{n=1}^N$ from corresponding text prompts $\mathbf{t} = \{t^n\}_{n=1}^N$, each composed of the same identity prompts $\mathbf{t}_{\text{iden}} = \{t_{\text{iden}}^n\}_{n=1}^N$ and distinct expression prompts $\mathbf{t}_{\text{exp}} = \{t_{\text{exp}}^n\}_{n=1}^N$, with the objective of maintaining consistent identity and overall style. All prompts are concatenated and processed in parallel as a batch.

Our method is based on the Infinity architecture (Han et al. 2024), which employs a next-scale prediction scheme (Tian et al. 2024). The model consists of a pre-trained text encoder $E_T$ employing Flan-T5 (Chung et al. 2024), a transformer $G$ that autoregressively predicts quantized residual $s$-th feature maps $\mathbf{R}_s$ over steps $\mathbf{S} = \{1, 2, ..., S\}$, and a decoder $D$ that reconstructs images from the final feature maps:

$$\mathbf{I} = D(\mathbf{F}_S), \quad \mathbf{F}_s = \sum_{i=1}^{s} \text{up}_{H\times W}(\mathbf{R}_i), \; \mathbf{R}_s \in \mathbb{R}^{N\times h_s \times w_s},$$
$$\mathbf{R}_s = G(\mathbf{F}_{s-1}, \mathbf{T}), \quad \mathbf{T} = E_T(\mathbf{t}) = \left\{\left(T_{\text{iden}}^n, T_{\text{exp}}^n\right)\right\}_{n=1}^N, \tag{1}$$

where $h_s, w_s$ are the spatial sizes at step $s \in \mathbf{S}$, $\text{up}_{H\times W}(\cdot)$ denotes a bilinear upsampling function to upsample to $H \times W$ size, and $\mathbf{T}$ denotes the encoded identity and expression features. The initial feature map $\mathbf{F}_0$ is initialized from $\mathbf{T}$.

As shown in Figure 4, *Identity Prompt Replacement* is first applied to ensure consistent identity attributes. During generation, both *Adaptive Style Injection* and *Synchronized Guidance Adaptation* are applied to self-attention layers in early generation steps $\mathbf{S}_{\text{early}}$, promoting consistent identity appearance and global style across all generated images.

### Identity Prompt Replacement

It is well known that generative models reflect biases in their training data distributions (Zhou et al. 2024a; Wei, Kumar, and Zhang 2025). For instance, the prompt "a dog

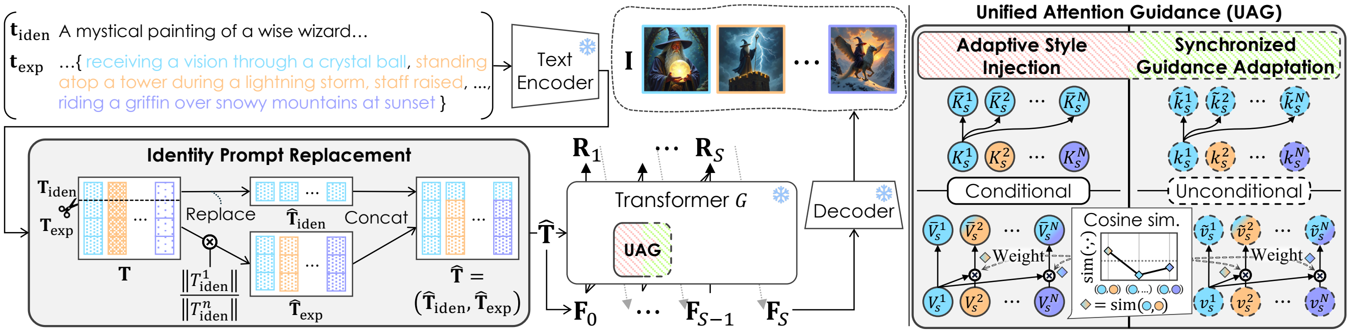


Figure 4: Overall pipeline of our method. The text encoder $E_T$ (Chung et al. 2024) processes a set of text prompts $\mathbf{t}$, producing contextual embeddings $\mathbf{T}$ that condition the transformer. *Identity Prompt Replacement* is applied to $\mathbf{T}$ before generation to ensure consistent identity representation across prompts. During generation, Unified Attention Guidance (UAG), which consists of *Adaptive Style Injection* and *Synchronized Guidance Adaptation*, is applied to early-stage self-attention layers to achieve consistent identity appearance and overall style alignment while preserving prompt fidelity. The transformer autoregressively produces residual feature maps, which are decoded into final images $\mathbf{I}$ via the image decoder.

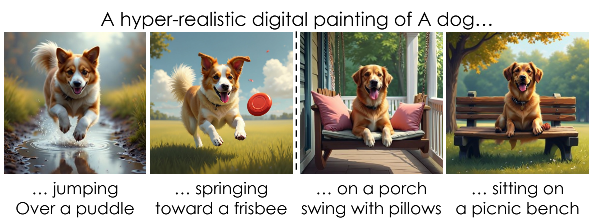


Figure 5: Context-bias in text-to-image generation.

springing toward a frisbee" (dynamic expression) often generates a Welsh corgi, while "a dog on a porch swing with pillows" (static expression) tends to produce calm, domesticated dogs like a Golden retriever, as illustrated in Figure 5—highlighting how prompt phrasing shapes semantic interpretation. This bias stems from the self-attention mechanism, where identity representations (e.g., "a dog") are influenced by their surrounding context, leading to inconsistent identity attributes—including gender, age, and species—across prompts.

To address the context bias inherent in text encoders, we propose an *Identity Prompt Replacement* (IPR) strategy that reduces such bias through the alignment of identity-related attributes across prompts. Specifically, we enforce a consistent representation of identity by replacing all identity embeddings $\mathbf{T}_{\text{iden}} = \{T^n_{\text{iden}}\}^N_{n=1}$ with those extracted from a reference instance (by default, the first sample in the batch). To maintain the proportional relationship between identity and expression features, we further normalize the magnitude of the expression embeddings $\mathbf{T}_{\text{exp}} = \{T^n_{\text{exp}}\}^N_{n=1}$ as follows:

$$\hat{\mathbf{T}} = \left(\hat{\mathbf{T}}_{\text{iden}}, \hat{\mathbf{T}}_{\text{exp}}\right) = \left\{ \left( T^1_{\text{iden}}, \frac{\|T^1_{\text{iden}}\|}{\|T^n_{\text{iden}}\|} \cdot T^n_{\text{exp}} \right) \right\}^N_{n=1}, \tag{2}$$

where $\hat{\mathbf{T}}_{\text{iden}}$ and $\hat{\mathbf{T}}_{\text{exp}}$ denote identity and expression prompt embeddings processed via IPR.

## Unified Attention Guidance

**Adaptive Style Injection** Although the Identity Prompt Replacement (IPR) technique mitigates context-level discrepancies by aligning identity-related attributes across prompts, it remains insufficient in preserving the appearance-level identity and global visual style consistency. To address this, we propose an *Adaptive Style Injection* (ASI) mechanism that aligns both the appearance of the identity and the overall scene style. ASI operates within the self-attention layers during the early generation steps, motivated by prior findings that analyze the functional roles of generation stages for style alignment (Park et al. 2025).

As illustrated in Figure 4-(Right), for each sample in the batch, we replace all key features in the self-attention with those of the reference, i.e., $K^1_s$, which encourages the model to attend to semantically consistent regions anchored by the reference. Also, we compute the cosine similarity between its value features and those of a reference instance, to obtain an adaptive interpolation weight $\alpha^n_s$, which is then used to interpolate the value features, facilitating appearance-level alignment of identity. The Adaptive Style Injection is defined as follows:

$$\begin{aligned} \bar{K}^n_s &= K^1_s, \bar{V}^n_s = \alpha^n_s V^n_s + (1-\alpha^n_s)V^1_s, \\ \alpha^n_s &= \lambda \cdot \text{sim}(V^1_s, V^n_s), \forall n \in \{1, \cdots, N\}, \end{aligned} \tag{3}$$

where $K^1_s$ and $V^1_s$ denote the key and value features of the reference sample, $V^n_s$ is the value feature of the $n$-th sample at $s$-th generation step, and $\lambda$ is a scaling coefficient. This similarity-guided adaptive operation facilitates the smooth and proportional alignment of appearance-level identity and global visual style across the batch, guided by the reference instance.

**Synchronized Guidance Adaptation** While Adaptive Style Injection improves identity appearance and global style consistency, applying it only to the conditional branch disrupts the balance between the conditional and unconditional signals established by Classifier-Free Guidance (CFG) (Ho and Salimans 2021), which is intended to enhance prompt fidelity. Such disruption may undermine the effectiveness of CFG, potentially degrading prompt fidelity in the generated images.

To resolve this, we propose *Synchronized Guidance Adaptation*, which applies the same operation to the unconditional branch using the identical interpolation weights computed from the conditional path, as illustrated in Figure 4-(Right). Specifically, for the unconditional branch, we modify the key and value features as:

$$\tilde{k}_s^n = k_s^1, \quad \tilde{v}_s^n = \alpha_s^n v_s^n + (1-\alpha_s^n) v_s^1, \ \forall n \in \{1, \cdots, N\}, \tag{4}$$

where $k_s^1$ and $v_s^1$ denote the unconditional branch's key and value features of the reference sample, $v_s^n$ is the value feature of the $n$-th sample at $s$-th generation step, and $\alpha_s^n$ is the adaptive weight shared from the conditional branch Equation (3). By synchronizing the feature adaptation across both branches, our approach preserves the intended effect of classifier-free guidance, enabling generated images to faithfully reflect their text prompts while maintaining a consistent subject identity and overall style.

# Experiment

## Implementation details

In our experiment, we leverage the pre-trained Infinity 2B model (Han et al. 2024) as our baseline. The model performs scale-wise prediction over 12 steps and employs a codebook with a dimensionality of $2^{32}$, producing quantized feature maps of resolution $64 \times 64$ with 32 channels. The early-stage step for Adaptive Style Injection and Synchronized Guidance Adaptation is fixed at $\mathbf{S}_{\text{early}} = \{2, 3\}$, and the scaling coefficient $\lambda$ is set to 0.85. All other components of the model architecture remain unchanged, and all parameters are frozen throughout inference.

The number of output images is determined by the number of input text prompts. When generating four $1024 \times 1024$ images in parallel on a single A6000 GPU, the process takes approximately 6.88 seconds in total, or 1.72 seconds per image. For scenarios involving more than four prompts, we adopt a batched generation strategy: in each batch, the identity prompt paired with the first expression prompt is always placed first, while the remaining positions are filled with the other prompts. This approach ensures that identity information remains consistent and is effectively propagated across all generated batches.

## Evaluation Setup

**Benchmark.** We follow the evaluation protocol proposed in 1Prompt1Story (Liu et al. 2025), an extension of the original ConsiStory benchmark (Tewel et al. 2024). ConsiStory+ expands the evaluation space by introducing a more diverse range of subjects, prompt descriptions, and styles. In accordance with this setup, we evaluate both prompt alignment as well as the consistency of subject identity and style over 200 distinct prompt sets, resulting in the generation of up to 1,500 images in total.

**Evaluation metrics.** Following 1Prompt1Story, to assess *prompt fidelity*, we compute the average CLIP text score (Radford et al. 2021) between each generated image and its corresponding prompt, denoted as CLIP-T. For *identity consistency*, we utilize DreamSim (Fu et al. 2023), a perceptual similarity metric shown to correlate well with human judgment, as well as CLIP-I (Radford et al. 2021), which measures the cosine similarity between image embeddings. Following the protocol of DreamSim, we remove image backgrounds using CarveKit (Selin 2023) and replace them with random noise, so that the similarity measurements reflect only the subject's identity. The same background removal process is applied to images evaluated with CLIP-I to ensure consistency across identity-based metrics. To assess *style consistency* among images conditioned on the same identity prompt, we follow prior work on style-aligned image generation (Hertz et al. 2024; Park et al. 2025; Frenkel et al. 2024) and compute the average pairwise DINO similarity, which captures alignment in overall visual appearance. For a more comprehensive evaluation, we further report a harmonic score $S_H$, which aggregates four core metrics (CLIP-T, CLIP-I, DreamSim, and DINO) using the harmonic mean. Since DreamSim is a distance-based metric, we convert it to a similarity measure via $[1 - \text{DreamSim}]$ before computing the mean. This composite score provides a balanced view of both prompt fidelity and visual consistency across identity and style.

## Comparison with state-of-the-art consistent text-to-image generation models

**Quantitative comparison.** Table 1 provides a comparative analysis of our method against a variety of state-of-the-art consistent text-to-image generation models, encompassing both training-based and training-free approaches. Despite requiring neither fine-tuning nor training, our method achieves the best $S_H$ score, reflecting a strong balance across all core metrics. Notably, our method achieves the highest DINO similarity, as well as leading scores in CLIP-I and DreamSim, demonstrating superior consistency in both style and identity. The identity metrics are computed after background removal to isolate subject appearance, further validating our approach's robustness in preserving subject identity across generated images. While 1Prompt1Story also delivers competitive results as a training-free baseline, our method surpasses it in both style and identity consistency, and overall $S_H$, while operating over $13\times$ faster. Importantly, these results are achieved with an inference time of just 1.72 seconds per image—significantly faster than most diffusion-based models, which typically exceed 10 seconds. These results underscore that Infinite-Story not only provides high-quality and consistent generations but also does so with remarkable efficiency, making it suitable for practical deployment in real-time generation scenarios.

**Qualitative comparison.** Figure 6 presents qualitative results across two themes—an elf character and a watercolor-style hedgehog—used to evaluate various consistent text-to-image generation models. Some methods, such as IP-Adapter, preserve subject identity effectively, especially in facial structure and posture. However, they often fail to re-

| Method | Train-Free | $S_H \uparrow$ | DINO ↑ | CLIP-T ↑ | CLIP-I ↑ | DreamSim ↓ | Inference Time (s) ↓ |
|---|---|---|---|---|---|---|---|
| Vanilla SDXL (Podell et al. 2023) | - | 0.7408 | 0.6067 | 0.9074 | 0.8793 | 0.3385 | 10.27 |
| Vanilla Infinity (Han et al. 2024) | - | 0.7891 | 0.6965 | 0.8836 | 0.8955 | 0.2780 | 1.71 |
| IP-Adapter (Ye et al. 2023) | ✗ | 0.8323 | 0.7834 | 0.8661 | 0.9243 | 0.2266 | 10.40 |
| PhotoMaker (Li et al. 2024) | ✗ | 0.7223 | 0.6516 | 0.8651 | 0.8465 | 0.3996 | 19.52 |
| The Chosen One (Avrahami et al. 2024) | ✗ | 0.6494 | 0.5824 | 0.8162 | 0.7943 | 0.4893 | 13.47 |
| OneActor (Wang et al. 2024) | ✗ | 0.8088 | 0.7172 | 0.8859 | 0.9070 | 0.2423 | 24.94 |
| ConsiStory (Tewel et al. 2024) | ✓ | 0.7902 | 0.6895 | **0.9019** | 0.8954 | 0.2787 | 37.76 |
| StoryDiffusion (Zhou et al. 2024b) | ✓ | 0.7634 | 0.6783 | 0.8403 | 0.8917 | 0.3212 | 23.68 |
| 1Prompt1Story (Liu et al. 2025) | ✓ | 0.8395 | 0.7687 | 0.8942 | 0.9117 | 0.1993 | 22.57 |
| Ours | ✓ | **0.8538** | **0.8089** | 0.8732 | **0.9267** | **0.1834** | **1.72** |

Table 1: Quantitative comparison with state-of-the-art consistent T2I generation models. Inference time is reported per image. Symbols ↑ and ↓ indicate whether higher or lower values are better. **Bold** and underline denote the best and second-best results, respectively.

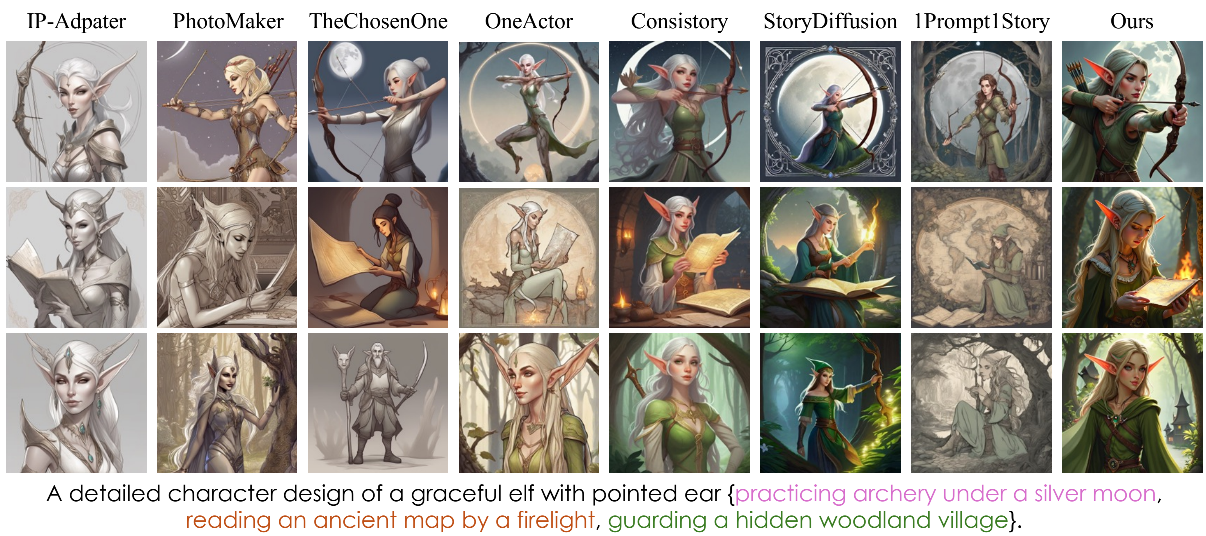


Figure 6: Qualitative comparison with state-of-the-art consistent T2I generation models.

| # | Component | | | Quantitative Metrics | | | | |
|---|---|---|---|---|---|---|---|---|
| | IPR | ASI | SGA | $S_H \uparrow$ | DINO ↑ | CLIP-T ↑ | CLIP-I ↑ | DreamSim ↓ |
| (a) | | | | 0.7891 | 0.6965 | **0.8836** | 0.8955 | 0.2780 |
| (b) | ✓ | | | 0.8013 | 0.7119 | 0.8814 | 0.9046 | 0.2569 |
| (c) | ✓ | ✓ | | 0.8481 | 0.8082 | 0.8625 | 0.9242 | 0.1931 |
| (d) | ✓ | ✓ | ✓ | **0.8538** | **0.8089** | 0.8732 | **0.9267** | **0.1834** |

Table 2: Ablation study on the Identity Prompt Replacement (IPR), Adaptive Style Injection (ASI), and Synchronized Guidance Adaptation (SGA). The symbol ↑ indicates that higher values are better, and ↓ indicates that lower values are better. The best and second-best results are highlighted in **bold** and underline, respectively.

flect prompt-specific nuances, as seen in the elf example, where different expression settings (e.g., "guarding a hidden woodland village") result in minimal contextual variation. On the other hand, OneActor and 1Prompt1Story capture expression prompts well but show style shifts in background and rendering details that disrupt visual cohesion. StoryDiffusion and ConsiStory demonstrate style consistency, yet they exhibit inconsistencies in subject identity across the prompt set. PhotoMaker and The Chosen One, while producing aesthetically pleasing results, tend to underperform across all three aspects—prompt fidelity, identity consistency, and style consistency.

Our model successfully addresses all three aspects. In both the elf and hedgehog scenarios, the generated images reflect clear variation across prompts while consistently preserving subject identity and maintaining a unified visual style. These results confirm that our method generates image sequences that are identity-consistent, style-consistent, and faithful to the prompt.

## Ablation study

**Quantitative Analysis.** Table 2 presents an ablation study evaluating the contributions of each proposed component. Starting with *Identity Prompt Replacement* (IPR), as shown in row (b), we observe notable gains in CLIP-I and DreamSim, confirming its effectiveness in aligning identity-related

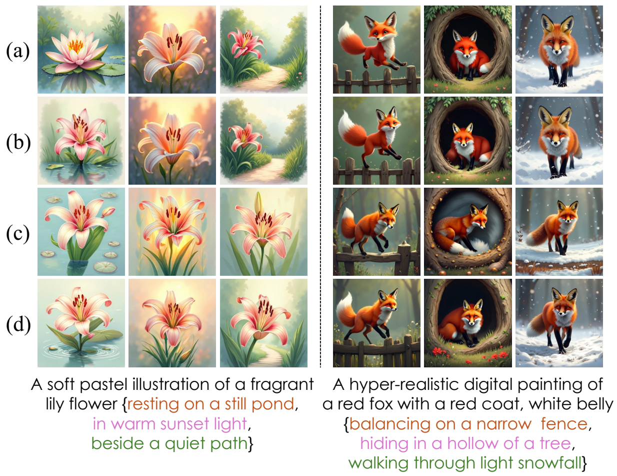


Figure 7: Qualitative analysis of ablation study. The results from (a)-(d) correspond to the configurations in Table 2.

attributes across prompts by mitigating the context bias of text encoders. When *Adaptive Style Injection* (ASI) is added in (c), DINO similarity increases significantly, indicating improved global style consistency. Additionally, CLIP-I and DreamSim scores also improve, reflecting enhanced alignment in identity appearance. Finally, in (d), applying *Synchronized Guidance Adaptation* (SGA) helps restore the balance between the conditional and unconditional branches of CFG, leads to a meaningful gain in CLIP-T, and further consolidates overall consistency. Although there is a slight trade-off in prompt fidelity compared to the baseline, the full configuration achieves the highest harmonic score $S_H$, indicating that our method successfully balances identity coherence, style consistency, and prompt fidelity—all without any additional fine-tuning.

**Qualitative analysis.** Figure 7 presents qualitative results from our ablation study on the proposed methods. Without any proposed methods (a), the generated images exhibit severe inconsistency in both subject identity and visual style. For instance, the flower species and rendering styles vary across scenes, and the red fox's appearance—such as fur texture and facial shape—fluctuates noticeably. Introducing only *Identity Prompt Replacement* (IPR) in (b) improves identity-related attributes by mitigating the context bias of text encoders. The lily maintains a more unified floral structure across prompts, and the red fox preserves more consistent facial features and body proportions. However, global style elements—such as lighting and rendering—remain inconsistent. When *Adaptive Style Injection* (ASI) is added (c), both global style and appearance-level identity consistency are further enhanced. The flower exhibits stable coloration and stroke patterns, while the red fox retains consistent shading and background textures across diverse scenes. Nevertheless, some prompt-specific semantics remain underemphasized, and visual artifacts such as unnatural outlines or distorted textures occasionally appear—likely due to strong style injection overriding localized details. Finally, applying the full method with *Synchronized Guidance Adaptation* (SGA) in (d) restores balance between the conditional and unconditional branches, enabling better preservation of prompt fidelity. This results in visually coherent outputs that maintain consistent subject appearance and unified stylistic rendering, while accurately reflecting prompt-specific variations—evidenced by appropriate posture, context, and lighting across prompts. These qualitative trends are consistent with the quantitative improvements observed in Table 2.

| Method | Identity ↑ | Style ↑ | Prompt ↑ |
|---|---|---|---|
| 1Prompt1Story (Liu et al. 2025) | 18.0% | 13.2% | 28.2% |
| OneActor (Wang et al. 2024) | 7.2% | 7.2% | 10.6% |
| IP-Adapter (Ye et al. 2023) | 16.4% | 29.6% | 4.7% |
| **Ours** | **58.4%** | **50.0%** | **56.5%** |

Table 3: User study preference percentages.

### User study

To complement our quantitative evaluation, we conduct a user study, with results shown in Table 3. A total of 55 participants were asked to assess three core criteria: identity consistency, prompt fidelity, and style consistency. We compare images generated by our model with those from 1Prompt1Story (Liu et al. 2025), OneActor (Wang et al. 2024), and IP-Adapter (Ye et al. 2023), which ranked highest in our quantitative benchmarks. The results of the user study indicate that our model consistently outperforms competing methods in all three aspects—identity, prompt, and style consistency—demonstrating strong human-perceived performance across a variety of prompts. Details of the user study protocol are provided in the supplementary material.

## Conclusion

In this paper, we present *Infinite-Story*, a training-free framework for consistent text-to-image generation tailored to multi-prompt scenarios. Built upon a scale-wise autoregressive backbone, our method tackles two key challenges in consistent generation—identity inconsistency and style inconsistency—without requiring model fine-tuning or training. To this end, we introduce three lightweight yet effective techniques: *Identity Prompt Replacement*, which mitigates the context bias of text encoders to align identity-related attributes, and a unified attention guidance strategy that combines *Adaptive Style Injection* and *Synchronized Guidance Adaptation* to align appearance-level identity and global style while preserving prompt fidelity. Extensive experiments demonstrate that Infinite-Story achieves state-of-the-art consistency in both identity and style while maintaining generation diversity. Notably, our approach operates over 6× faster than leading diffusion-based methods, underscoring its practicality for real-time and interactive applications such as visual storytelling and character-driven content generation. Future work includes extending our method to support temporal consistency in video generation and exploring more adaptive reference selection strategies.

## Acknowledgments

This research was supported by the 2024 innovation base artificial intelligence data convergence project project with the funding of the 2024 government (Ministry of Science and ICT) (S2201-24-1002), Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. RS-2025-02219277, AI Star Fellowship Support Project(DGIST)), Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2025-25420118) and LG AI STAR Talent Development Program for Leading Large-Scale Generative AI Models in the Physical AI Domain (RS-2025-25442149).

## References

Avrahami, O.; Hertz, A.; Vinker, Y.; Arar, M.; Fruchter, S.; Fried, O.; Cohen-Or, D.; and Lischinski, D. 2024. The chosen one: Consistent characters in text-to-image diffusion models. In *ACM SIGGRAPH 2024 conference papers*, 1–12.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18392–18402.

Byeon, M.; Park, B.; Kim, H.; Lee, S.; Baek, W.; and Kim, S. 2022. COYO-700M: Image-Text Pair Dataset. https://github.com/kakaobrain/coyo-dataset.

Chang, H.; Zhang, H.; Barber, J.; Maschinot, A.; Lezama, J.; Jiang, L.; Yang, M.-H.; Murphy, K.; Freeman, W. T.; Rubinstein, M.; et al. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*.

Chang, H.; Zhang, H.; Jiang, L.; Liu, C.; and Freeman, W. T. 2022. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11315–11325.

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3558–3568.

Chen, H.-Y.; Lai, Z.; Zhang, H.; Wang, X.; Eichner, M.; You, K.; Cao, M.; Zhang, B.; Yang, Y.; and Gan, Z. 2025. Contrastive Localized Language-Image Pre-Training. arXiv:2410.02746.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70): 1–53.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883.

Frenkel, Y.; Vinker, Y.; Shamir, A.; and Cohen-Or, D. 2024. Implicit style-content separation using b-lora. In *European Conference on Computer Vision*, 181–198. Springer.

Fu, S.; Tamir, N.; Sundaram, S.; Chai, L.; Zhang, R.; Dekel, T.; and Isola, P. 2023. DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data. *Advances in Neural Information Processing Systems*, 36: 50742–50768.

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. *arXiv preprint arXiv:2208.01618*.

Han, J.; Liu, J.; Jiang, Y.; Yan, B.; Zhang, Y.; Yuan, Z.; Peng, B.; and Liu, X. 2024. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. *arXiv preprint arXiv:2412.04431*.

Hertz, A.; Aberman, K.; and Cohen-Or, D. 2023. Delta denoising score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2328–2337.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control.

Hertz, A.; Voynov, A.; Fruchter, S.; and Cohen-Or, D. 2024. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4775–4785.

Ho, J.; and Salimans, T. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Kang, M.; Zhu, J.-Y.; Zhang, R.; Park, J.; Shechtman, E.; Paris, S.; and Park, T. 2023. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10124–10134.

Kondratyuk, D.; Yu, L.; Gu, X.; Lezama, J.; Huang, J.; Schindler, G.; Hornung, R.; Birodkar, V.; Yan, J.; Chiu, M.-C.; et al. 2023. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*.

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.

Labs, B. F. 2024. FLUX. https://github.com/black-forest-labs/flux.

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.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, 740–755. Springer.

Liu, T.; Wang, K.; Li, S.; van de Weijer, J.; Khan, F.; Yang, S.; Wang, Y.; Yang, J.; and Cheng, 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*.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; Shan, Y.; and Qie, X. 2023. T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models. arXiv:2302.08453.

Nam, J.; Kim, H.; Lee, D.; Jin, S.; Kim, S.; and Chang, S. 2024. Dreammatcher: appearance matching self-attention for semantically-consistent text-to-image personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8100–8110.

Nielsen, J. 1994. *Usability engineering*. Morgan Kaufmann.

Park, J.; Gim, J.; Lee, K.; Oh, M.; Choi, M.; Kim, J.; Park, W. C.; and Im, S. 2025. A Training-Free Style-aligned Image Generation with Scale-wise Autoregressive Model. *arXiv preprint arXiv:2504.06144*.

Parmar, G.; Kumar Singh, K.; Zhang, R.; Li, Y.; Lu, J.; and Zhu, J.-Y. 2023. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 conference proceedings*, 1–11.

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. *arXiv preprint arXiv:2307.01952*.

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.

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.

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*, 22500–22510.

Ryu, S. 2022. Low-rank adaptation for fast text-to-image diffusion fine-tuning, 2022. *URL https://github. com/cloneofsimo/lora*, 10: 19.

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.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35: 25278–25294.

Selin, N. 2023. CarveKit: Automated high-quality background removal framework. https://github.com/OPHoperHPO/image-background-remove-tool.

Shah, V.; Ruiz, N.; Cole, F.; Lu, E.; Lazebnik, S.; Li, Y.; and Jampani, V. 2024. Ziplora: Any subject in any style by effectively merging loras. In *European Conference on Computer Vision*, 422–438. Springer.

Sohn, K.; Ruiz, N.; Lee, K.; Chin, D. C.; Blok, I.; Chang, H.; Barber, J.; Jiang, L.; Entis, G.; Li, Y.; et al. 2023. Styledrop: Text-to-image generation in any style. *arXiv preprint arXiv:2306.00983*.

Tang, H.; Wu, Y.; Yang, S.; Xie, E.; Chen, J.; Chen, J.; Zhang, Z.; Cai, H.; Lu, Y.; and Han, S. 2024. Hart: Efficient visual generation with hybrid autoregressive transformer. *arXiv preprint arXiv:2410.10812*.

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)*, 43(4): 1–18.

Tian, K.; Jiang, Y.; Yuan, Z.; Peng, B.; and Wang, L. 2024. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*.

Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1921–1930.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Voronov, A.; Kuznedelev, D.; Khoroshikh, M.; Khrulkov, V.; and Baranchuk, D. 2024. Switti: Designing Scale-Wise Transformers for Text-to-Image Synthesis. *arXiv preprint arXiv:2412.01819*.

Wang, J.; Yan, C.; Lin, H.; Zhang, W.; Wang, M.; Gong, T.; Dai, G.; and Sun, H. 2024. OneActor: Consistent Subject Generation via Cluster-Conditioned Guidance. *Advances in Neural Information Processing Systems*, 37: 21502–21536.

Wang, S.; Saharia, C.; Montgomery, C.; Pont-Tuset, J.; Noy, S.; Pellegrini, S.; Onoe, Y.; Laszlo, S.; Fleet, D. J.; Soricut, R.; et al. 2023. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18359–18369.

Wei, X.; Kumar, N.; and Zhang, H. 2025. Addressing bias in generative AI: Challenges and research opportunities in information management. *arXiv preprint arXiv:2502.10407*.

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*.

Zhou, M.; Abhishek, V.; Derdenger, T.; Kim, J.; and Srinivasan, K. 2024a. Bias in generative AI. *arXiv preprint arXiv:2403.02726*.

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.