STORY-ADAPTER: A TRAINING-FREE ITERATIVE FRAMEWORK FOR LONG STORY VISUALIZATION

Anonymous authors
Paper under double-blind review

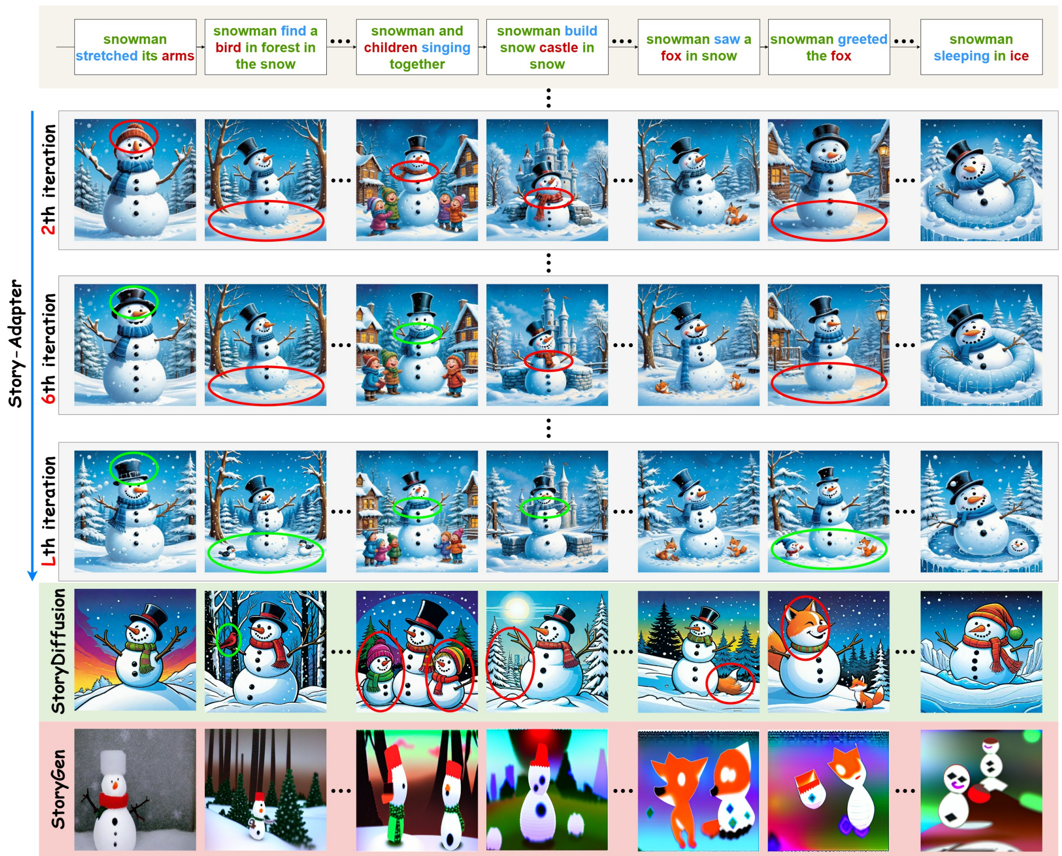

Figure 1: A long story of "*snowman*" visualized by our Story-Adapter from different iterations, compared with those visualized by previous StoryDiffusion (Zhou et al., 2024) and StoryGen (Liu et al., 2024). Notable differences are highlighted in green and red. Zoom in for a better view.

ABSTRACT

Story visualization, the task of generating coherent images based on a narrative, has seen significant advancements with the emergence of text-to-image models, particularly diffusion models. However, maintaining semantic consistency, generating high-quality fine-grained interactions, and ensuring computational feasibility remain challenging, especially in long story visualization (*i.e.*, up to 100 frames). In this work, we propose a training-free and computationally efficient framework, termed **Story-Adapter**, to enhance the generative capability of long stories. Specifically, we propose an *iterative* paradigm to refine each generated image, leveraging both the text prompt and all generated images from the previous iteration. Central to our framework is a training-free global reference cross-attention module, which aggregates all generated images from the previous iteration to preserve semantic consistency across the entire story, while minimizing computational costs with global embeddings. This iterative process progressively optimizes image generation by repeatedly incorporating text constraints, resulting in more precise and fine-grained interactions. Extensive experiments validate the superiority of Story-Adapter in improving both semantic consistency and generative capability for fine-grained interactions, particularly in long story scenarios.

1 INTRODUCTION

Story visualization aims to generate a sequence of coherent images from text prompts, reflecting the narrative's progression and enabling users, even without an artistic background, to visually present their stories (Li et al., 2019; Maharana & Bansal, 2021; Chen et al., 2022). Recent advancements in text-to-image models, particularly diffusion models, have significantly improved the quality of generated visuals, producing high-quality, creative, and aesthetically pleasing images (Saharia et al., 2022; Rombach et al., 2022; Kang et al., 2023). These models greatly outperform earlier approaches such as generative adversarial networks (Brock, 2018) in terms of image quality.

However, story visualization remains challenging, particularly in maintaining semantic consistency and capturing complex interactions as the story length increases. Two main paradigms have emerged in this domain. The **Auto-Regressive paradigm**(Fig. 2A), which generates frames sequentially (Pan et al., 2024; Liu et al., 2024), often struggles with semantic consistency due to error accumulation and the inability to reference future frames, leading to inconsistencies in the overall narrative. Although techniques like Consistent Self-Attention (CSA) (Zhou et al., 2024) can help mitigate these inconsistencies, their reliance on intermediate denoising features results in high memory consumption, limiting scalability for longer stories. To address these challenges, Zhou et al. (2024) further propose the **Reference-Image paradigm**, which employs fixed reference images to guide the visualization process. However, as shown in Fig. 2B, while using only the initial frames as reference images alleviate scalability issues, it fails to provide the global semantic coherence necessary for long-story visualization, ultimately resulting in the propagation of errors from the reference images to subsequent frames. As such, both paradigms experience quality degradation when visualizing long stories. Additionally, they inherit the limitations from Stable Diffusion (SD) (Rombach et al., 2022), particularly in generating fine-grained interactions (as shown in Fig. 1).

To address these limitations, we present **Story-Adapter**, an *iterative framework* that adapts pre-trained SD models for long story visualization. Unlike existing methods that generate images auto-regressively or rely on static reference images (Fig. 2 A&B), our approach prioritizes semantic consistency by incorporating *all* generated images from previous iterations into the current one. This process offers two key advantages. 1) It offers a comprehensive view of the entire narrative, thereby reducing error accumulation and mitigating the propagation of flaws from reference images. 2) By continuously engaging with text prompts, Story-Adapter optimizes generative quality for details based on insights from earlier iterations. As illustrated in Fig.1, our framework enhances both semantic consistency and the quality of fine-grained interactions across iterations, resulting in more coherent and higher-quality visualizations. For example, the image depicting complex character interactions, such as "the snowman greeting the fox" demonstrates substantial improvement over iterations compared to previous methods(Liu et al., 2024; Zhou et al., 2024).

During initialization, only text prompts of the story are utilized to generate reference images. In subsequent iterations, the global embeddings of all images generated in the previous round, along with the text embeddings, collaboratively guide the image generation process. To implement the iterative paradigm efficiently, we propose a plug-and-play *Global Reference Cross-Attention (GRCA)*, where all global image embeddings act as keys and values. This significantly reduces computational costs, as global embeddings operate at a lower dimensionality than the intermediate denoising features used in CSA. Additionally, to strike a balance between visual consistency and text controllability, we introduce a linear weighting strategy in the iterative paradigm to fuse both modalities.

Extensive experiments demonstrate that Story-Adapter consistently outperforms existing methods for visualizing both regular-length and long stories (up to 100 frames). Specifically, in the context of regular-length story visualization using the StorySalon benchmark dataset (Liu et al., 2024), Story-Adapter exceeds the baseline model, StoryGen(Liu et al., 2024), achieving a 9.4% improvement in average Character-Character Similarity (aCCS)(Cheng et al., 2024) and a 21.71 reduction in average Fréchet Inception Distance (aFID) (Cheng et al., 2024). For long story visualization, Story-Adapter also demonstrates solid advancements, achieving gains of 3.4% in aCCS and 8.14 in

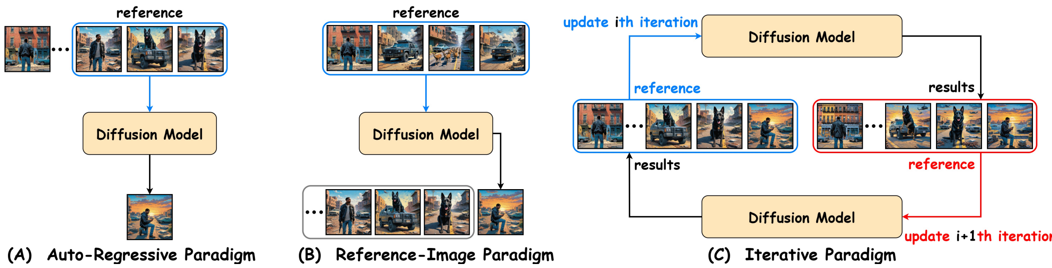

Figure 2: Comparison of paradigms for long story visualization: (A) Auto-Regressive (AR): generates frames sequentially referencing on previous finite frames (*e.g.* the previous three frames); (B) Reference-Image (RI): employs fixed reference images (*e.g.* the beginning four frames) as reference images; (C) Iterative Paradigm: leverages all frames from the previous iteration as reference images.

aFID compared to StoryDiffusion (Zhou et al., 2024), demonstrating the superior generative quality of Story-Adapter, particularly in terms of semantic consistency and fine-grained interactions.

2 Related Work

2.1 Diffusion Models

Diffusion models (Ho et al., 2020; Song et al., 2020b; Sohl-Dickstein et al., 2015; Nichol & Dhariwal, 2021) have emerged as powerful tools for data distribution modeling through iterative denoising. Recent advancements in sampling techniques (Xiao et al., 2021; Song et al., 2020a; Luo et al., 2023), backbone architectures (Peebles & Xie, 2023; Lu et al., 2023), and latent space denoising (Rombach et al., 2022; Podell et al., 2023) have led to their widespread adoption in various generative tasks, including video (Esser et al., 2023; Yang et al., 2024), 3D (Luo & Hu, 2021; Xu et al., 2024), audio (Ruan et al., 2023; Huang et al., 2023), and human motion generation (Zhang et al., 2022; Karunratanakul et al., 2023). While Text-to-Image diffusion models (Saharia et al., 2022; Zhang et al., 2023; Rombach et al., 2022; Podell et al., 2023) have gained significant attention, challenges persist in generating coherent image sequences for tasks like story visualization due to the inherent randomness and fine-grained interaction generation.

2.2 Story Visualization

Story visualization (Chen et al., 2022; Li, 2022) has evolved from GAN-based approaches like StoryGAN (Li et al., 2019) to more advanced techniques. Recent developments leverage diffusion models (Shen et al., 2024; Tao et al., 2024) and combine them with auto-regressive paradigm, as seen in AR-LDM (Pan et al., 2024) and StoryGen (Liu et al., 2024). These methods have improved coherence in image sequences and extended to open-ended story visualization. However, challenges remain in maintaining semantic consistency for the whole story and avoiding error accumulation, especially for longer narratives (Wang et al., 2023; Zhou et al., 2024; Liu et al., 2024).

2.3 Subject-Consistent Image Generation

The consistency of the generated subjects is critical for tasks such as story visualization and video generation. Recent advancements in subject-consistent image generation have focused on reducing computational resources while maintaining consistency. Early approaches like Gal et al. (2022); Ruiz et al. (2023) require extensive fine-tuning, prompting more efficient methods (Ryu, 2023; Han et al., 2023; Kumari et al., 2023; Yuan et al., 2023). Notable progress includes IP-Adapter (Ye et al., 2023) with its decoupled cross-attention design and technique like PhotoMaker (Li et al., 2024) that accelerates generation using identity images. Recently, StoryDiffusion (Zhou et al., 2024) introduced Consistent Self-Attention (CSA) to boost the frame-wise subject consistency but still faces limitations in long image sequences. In contrast, Story-Adapter maintains image semantic consistency in long image sequences by using cross-attention on *global embeddings* from all generated images of the previous iteration and the corresponding text features. Along with our iteration paradigm, the whole generations are gradually improved *w.r.t* semantic consistency and generative quality for fine-grained interactions.

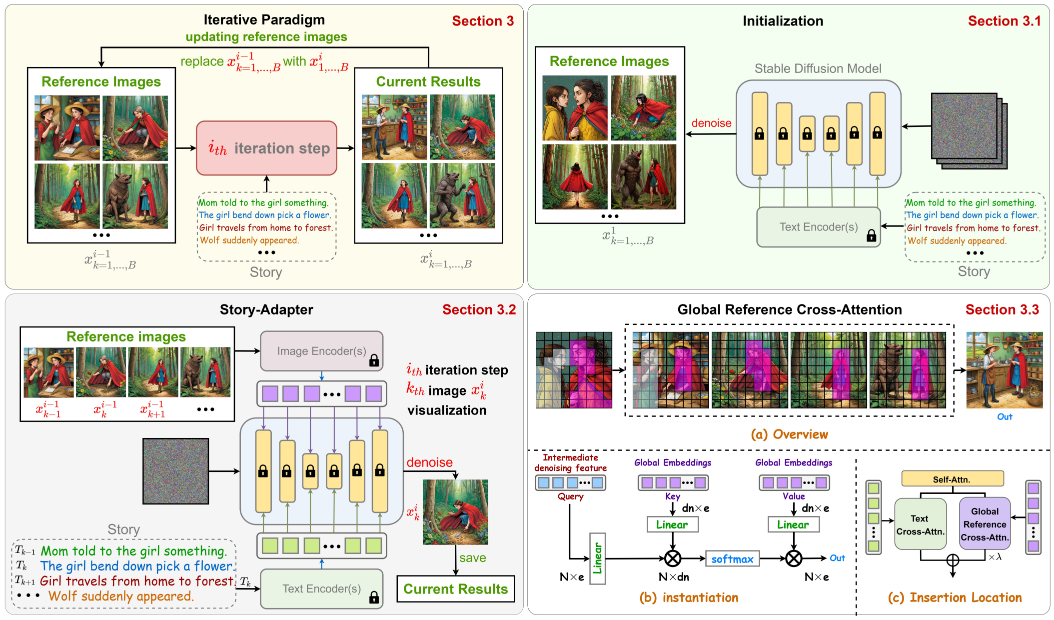

Figure 3: Illustration of the proposed iterative paradigm, which consists of initialization, iterations in Story-Adapter, and implementation of Global Reference Cross-Attention.

3 Method

Compared to regular-length stories, long stories contain more characters and more complex interactions, leading to higher requirements for semantic consistency and fine-grained interaction generation. To address the above challenges, we resort to an *iterative paradigm* that progressively refines all the generated images, *w.r.t.* semantic consistency and visual details in multiple rounds. We instantiate the iterative paradigm by equipping a fixed Stable-Diffusion (SD) model with a cross-attention mechanism, termed **Story-Adapter**. The pipeline is demonstrated in Fig. 3.

3.1 Initialization

To build the initialization for iteration, we only employ text prompt T_k for the k_{th} image in the story to guide the fixed $\mathrm{SD}(z, T_k)$ in generating the initial images, where z is the random noise. All generated images from the initial step will be stored as reference images for the first iteration. We denote $i = 0$ as the initialization of Story-Adapter. Thus, the whole initialization process can be represented as:

$$
\begin{aligned}
&x_k^{i=0} = \mathrm{SD}(z, T_k), k \in [1, B], \\
&x_{1,\cdots,B}^{i=0} = [x_1^{i=0}, x_2^{i=0}, \cdots, x_k^{i=0}, \cdots, x_{B-1}^{i=0}, x_B^{i=0}],
\end{aligned} \tag{1}
$$

where B denotes the length of the story. Compared to subject-consistent image generation methods (Ye et al., 2023) that introduce reference image guidance, initialization which relies only on text prompts more faithfully visualizes the corresponding content in the story. The following iterations benefit from the rich visual content provided by the initialization of the reference images.

3.2 Story-Adapter

This subsection demonstrates how each image is updated within an iteration in Story-Adapter. Formally, for the i_{th} iteration, we use all visualizations from the previous iteration $x^{i-1}_{1,\cdots,B}$ as the reference images R to refine the generated images in the current round. For the generation of the k_{th} image of a long story, we define a function $\mathrm{SD}_{\mathrm{GRCA}}(z, T_k, x^{i-1}_{1,\cdots,B})$ to represent the *whole* denoising process with our Global Reference Cross-Attention (Sec. 3.3).

Algorithm 1: Pseudo-Code of Story-Adapter.

```
# diffusion model:θ, iteration epochs:L, starting weight factor:λ_s,
    ending weight factor:λ_e, i_th iteration j_th diffusion step k_th
    intermediate denoising features:I^i_{k,j}, story length:B, diffusion
    steps:J, decoder:D
# Initialize I^0_{k,j}, I^i_{k,j}~N(0, I), k~(1, B), i~(1, L), j~(0, J)
# Initialize Story-Adapter iteration
for j in reversed(range(0, J)):
   # Init z~N(0, I) if j>1 else z=0
   I^0_{k,j-1}=(1/sqrt(α_j))*I^0_{k,j}-(1-α_j)*θ(I^0_{k,j},j,T_k)/sqrt(1-α_j))+σ_t*z
R=concat([x^0_1,...,x^0_k,...,x^0_B]), x^0_k=D(I^0_{k,0})

# Insert GRCA to θ and initialize weighting factor list λ_list
λ_list=linspace(λ_s,λ_e,L)
# Story-Adapter Iteration
for i,λ in enumerate(λ_list):
   for j in reversed(range(0, J)):
     I^i_{k,j-1}=(1/sqrt(α_j))*(I^i_{k,j}-(1-α_j)*θ(I^i_{k,j},j,T_k,R,λ)/sqrt(1-α_j))+σ_t*z
   R=concat([x^i_1,...,x^i_k,...,x^i_B]), x^i_k=D(I^i_{k,0})
```

Thus the i_{th} iteration can be expressed as:

$$
\begin{aligned}
x_k^i &= \mathrm{SD}_{\mathrm{GRCA}}(z, T_k, x_{1,\cdots,B}^{i-1}), k \in [1, B], \\
x_{1,\cdots,B}^i &= [x_1^i, x_2^i, \cdots, x_k^i, \cdots, x_{B-1}^i, x_B^i],
\end{aligned} \tag{2}
$$

As iterations proceed, the reference images evolve to be more coherent, as Story-Adapter consistently improves the semantic consistency in a global view. Additionally, generative quality for fine-grained interactions is also constantly optimized as Story-Adapter repeatedly engages text prompt constraints during iterations.

3.3 Global Reference Cross-Attention

Although incorporating image context or reference images extends text-to-image generation to character-consistent image sequences, existing AR paradigms (Pan et al., 2024; Liu et al., 2024) suffer from error accumulation over long stories, while RI paradigms (Zhou et al., 2024) may propagate flaws from the reference images.

In contrast, we propose an efficient plug-and-play augmentation module to equip SD models, called Global Reference Cross-Attention (GRCA). We utilize a pre-trained CLIP (Radford et al., 2021) image encoder to extract a global embedding c for each reference image from the previous round, effectively preserving the semantics of reference images using only a few tokens. The token simplification allows GRCA to incorporate all reference images as guidance in the cross-attention process without incurring significant computational overhead.

In the i_{th} iteration of Story-Adapter, given all the reference images in the previous round $x_{1,\cdots,B}^{i-1} \in \mathbb{R}^{B\times h\times w\times 3}$, h, w denote reference image resolution. We define a function $\mathrm{Attention}(Q, K, V)$ to indicate the attention calculation, where Q, K, and V represent the query, key, and value in the attention, respectively. GRCA in the visualization for the k_{th} image can be specified as:

$$
\begin{aligned}
&c_{1,\cdots,B}^i = \mathrm{CLIP}(x_{1,\cdots,B}^{i-1}), c_{1,\cdots,B}^i \in \mathbb{R}^{B\times d}, \\
&c_{1,\cdots,B}^i = c_{1,\cdots,B}^i W_c, W_c \in \mathbb{R}^{d\times ne}, \\
&c_{1,\cdots,B}^i = \mathrm{flatten}(c_{1,\cdots,B}^i), c_{1,\cdots,B}^i \in \mathbb{R}^{1\times Bn\times e}, \\
&Q_k^i = I_k W_q, K_k^i = c_{1,\cdots,B}^i W_k, V_k^i = c_{1,\cdots,B}^i W_v, \\
&\mathrm{GRCA}(I_k, x_{1,\cdots,B}^{i-1}) = \mathrm{Attention}(Q_k^i, K_k^i, V_k^i),
\end{aligned} \tag{3}
$$

Where W_c is the projection matrix of global embeddings transformed into reference tokens. d and e denote the embedded dimension of global embeddings and the projection dimension of the projection matrix, respectively. n indicates the number of reference tokens for a single reference image, $n = 4$ if not specified. $\mathrm{flatten}(.)$ represents a flatten operation for vectors. W_q is the mapping weight matrix for the intermediate denoising feature I in SD. W_k, W_v are the mapping weight matrices of the reference tokens.

Eventually, we merge the outputs from GRCA with the outputs from text cross-attention, to guide the visualization of k_{th} image in the story. In particular, with corresponding text prompt T_k and all reference images $x^{i-1}_{1,\cdots,B}$, the intermediate denoising feature I^i_k is obtained as follows:

$$I^i_k = \mathrm{Attention}(I^i_k, T_k, T_k) + \lambda \mathrm{GRCA}(I^i_k, x^{i-1}_{1,\cdots,B}). \tag{4}$$

where λ is a balance factor for controlling the influence of GRCA on the visualization results. We propose a linear weighting strategy to adjust the weight factor for each iteration, where the weight factor increases linearly with a low value to trade off visual consistency and text alignment in the iterative paradigm. Since the existing diffusion models contain a cross-attention design associated with the reference image, our GRCA could be directly plugged in and reuse the cross-attention weights without training. We demonstrate the procedure of Story-Adapter, along with the linear weighting strategy in Algo. 1.

4 Experiments

In this section, we first introduce the datasets, the evaluation metrics, and implementation details. Then we compare Story-Adapter with previous AR-based and RI-based methods for visualization of both regular-length and long stories. Finally, we validate the effectiveness of the proposed iterative paradigm and Global Reference Cross-Attention (GRCA) through extensive ablations. Additional experimental results, comparison on subject-consistent generation, and human evaluation can be found in the *Appendix*.

4.1 Dataset and Evaluation

We use the StorySalon dataset (Liu et al., 2024) to benchmark performance for *regular-length* story visualization. For *long* story visualization, we curate multiple long stories using GPT-4o (OpenAI, 2024). To evaluate the efficacy of Story-Adapter, we report CLIP text-image similarity (CLIP-T) (Radford et al., 2021), average Fréchet Inception Distance (aFID) (Cheng et al., 2024), and Character-Character Similarity (aCCS) (Cheng et al., 2024). CLIP-T is to measure image-text alignment, both aFID and aCCS are used to evaluate semantic consistency among generated images.

4.2 Implementation Details

To ensure a fair comparison, we used the weights of IP-Adapter (Ye et al., 2023) and IP-AdapterXL (Ye et al., 2023), respectively, resulting in two models: **Story-Adapter** and **Story-AdapterXL**. We utilized DDIM (Song et al., 2020a) for 50-step sampling with an unclassified classifier guidance score set to 7.5. For the hyperparameters in our iterative paradigm, we set the number of story iterations to 10 by default. The weight factor λ is set to 0.3 for the initial iteration and 0.5 for the final iteration, with linearly interpolated values for the intermediate iterations by our linear weighting strategy.

4.3 Regular-length Story Visualization

Based on the standard setup on StorySalon dataset (Liu et al., 2024), we compare with existing story visualization methods and Stable Diffusion Model (SDM) baselines, including StoryDiffusion (Zhou et al., 2024), StoryGen (Liu et al., 2024), AR-LDM (Pan et al., 2024), SDM (Rombach et al., 2022), Finetuned-SDM (fine-tuned on StorySalon), and Prompt-SDM. For Prompt-SDM, we use prompts of “cartoon-style images”. To adhere to copyright restrictions and ensure fair comparisons, we exclusively utilize text prompts from the open-source subset of the StorySalon test set for evaluation. This subset comprises 6,026 prompts, with an average of 14 frames per story and the longest story containing up to 44 frames.

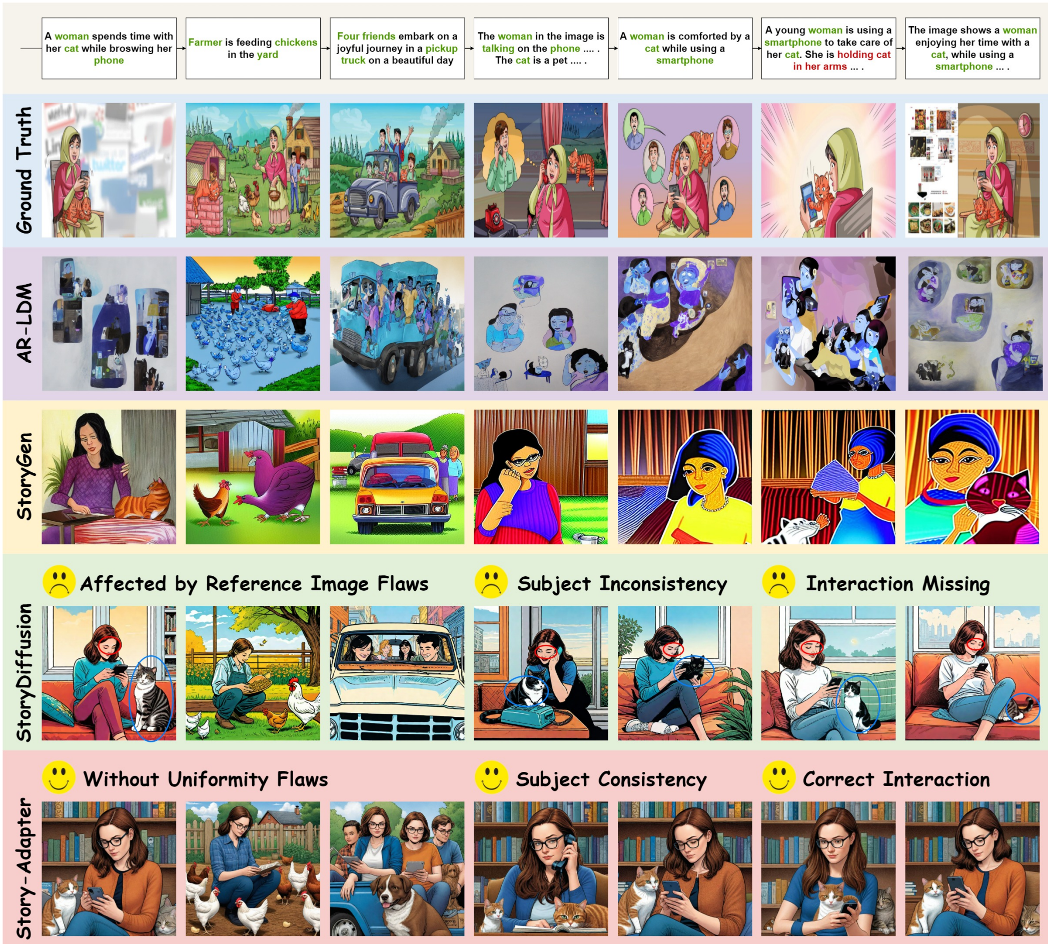

Figure 4: Qualitative comparisons for *regular-length* story visualization. Zoom in for a better view.

Table 1: Quantitative comparison for *regular-length* story visualization.

Method	CLIP-T ↑	aCCS ↑	aFID ↓
SDM (Rombach et al., 2022)	**0.323**	0.662	23.10
Prompt-SDM (Rombach et al., 2022)	0.289	0.699	18.18
Finetuned-SDM (Rombach et al., 2022)	0.309	0.639	23.05
AR-LDM (Pan et al., 2024)	0.237	0.683	40.25
StoryGen (Liu et al., 2024)	0.255	0.724	36.34
Story-Adapter (Ours)	0.305	0.760	16.52
StoryDiffusion (Zhou et al., 2024)	0.311	0.765	14.84
Story-AdapterXL (Ours)	0.310	**0.818**	**14.63**

Table 2: Quantitative comparison for *long* story visualization.

Method	CLIP-T ↑	aCCS ↑	aFID ↓
AR-LDM (Pan et al., 2024)	0.216	0.673	133.62
StoryGen (Liu et al., 2024)	0.223	0.740	126.13
IP-Adapter (Ye et al., 2023)	0.274	0.751	**93.70**
Story-Adapter (Ours)	**0.307**	**0.754**	98.51
IP-AdapterXL (Ye et al., 2023)	0.297	0.787	**88.69**
StoryDiffusion (Zhou et al., 2024)	0.315	0.768	102.44
Story-AdapterXL (Ours)	**0.318**	**0.802**	94.30

Quantitative Evaluation. CLIP-T results in Tab. 1 show that Story-Adapter and StoryDiffusion (Zhou et al., 2024) visualize content more aligned to the text prompt than previous story visualization models (AR-LDM and StoryGen). Meanwhile, since neither Story-Adapter nor most baselines are trained on the StorySalon dataset, we introduce aFID and aCCS metrics for a fair evaluation of the character consistency among generated story images. Results of aFID and aCCS in Tab. 1 illustrate that Story-Adapter achieves higher semantic consistency of the generated images compared to StoryDiffusion. Such results validate the effectiveness of our design for coherent image sequence visualization.

Qualitative Evaluation. In Fig. 4, we provide the qualitative comparison results of the open-ended story visualization. Although AR-LDM and StoryGen generate coherent image sequences based on story prompts, the quality of the generated images degrades when story length increases due to the

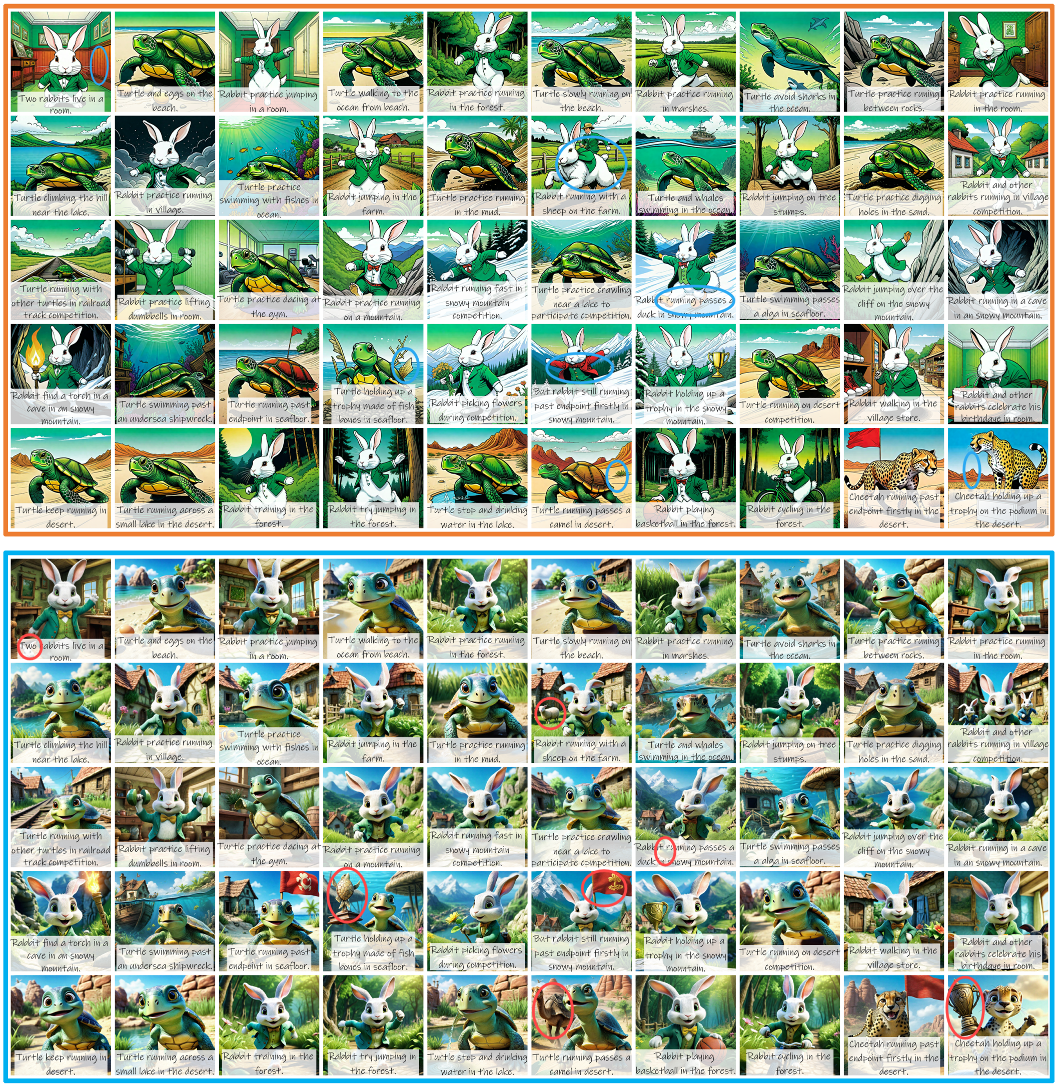

Figure 5: Qualitative comparisons for *long* story visualization. The image sequences in orange and blue boxes are generated by StoryDiffusion and Story-Adapter, respectively. Story-Adapter shows advantages in generating semantic consistency and character interactions. Zoom in for a better view.

error accumulation issue of the AR paradigm. Results of StoryDiffusion (Zhou et al., 2024) and Story-Adapter show satisfactory story visualization performance. However, StoryDiffusion cannot maintain consistency between certain subjects due to lacking global story comprehension (*e.g.*, "*cat*" in Fig. 4). Additionally, since StoryDiffusion requires the first few generated images as references, the visualization results are affected by the reference image flaws (*e.g.*, "*closed-eye issue*" in Fig. 4). In comparison, Story-Adapter performs better in regular-length story visualization benefited from the global features engaged in GRCA.

4.4 Long Story Visualization

To better evaluate generative quality for *long* story visualization (*i.e.*, up to 100 frames), we compare to subject-consistent image generation model IP-Adapter (Ye et al., 2023) in addition to existing story visualization methods. SDM baselines are not included in comparison as they are not suitable to generate long consistent content. We use GPT-4o (OpenAI, 2024) to generate 20 long story cases of ten 50-sentence descriptions and ten 100-sentence descriptions.

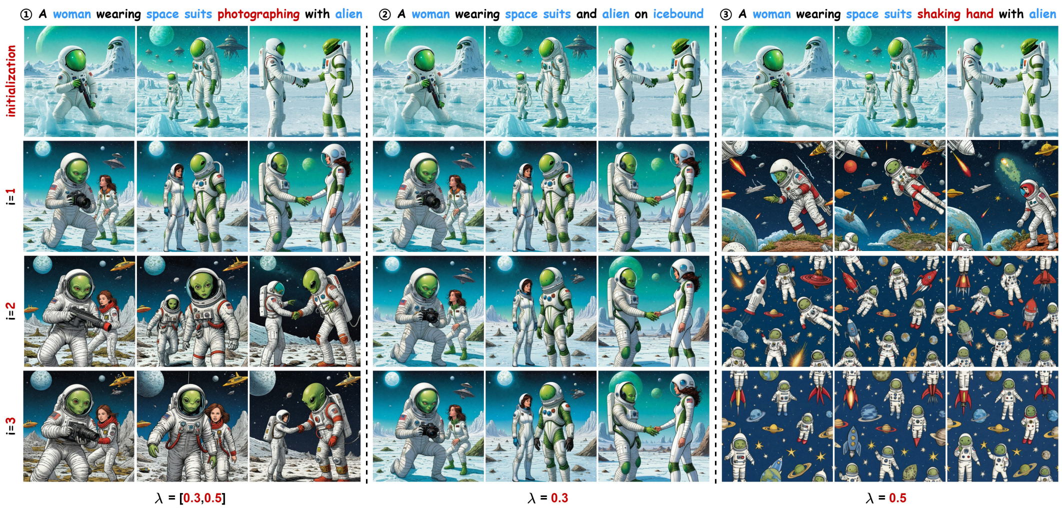

Figure 6: Ablation study of iterative paradigm: the effect of the iterative paradigm and the impact of different fixing λ. Zoom in for a better view. See *Appendix* for results with more iterations.

Quantitative Evaluation. The quantitative results in Tab. 2 show that our Story-Adapter significantly improves the semantic consistency and the generative coherence for fine-grained interactions for long story visualization compared to existing models. Notably, IP-Adapter employs the same guidance image that leads to less aFID. In contrast, our method improves visual consistency without the need to fix the same reference image.

Qualitative Evaluation. Fig. 5 shows the visualization results for long stories, indicating that Story-Adapter can generate high-quality, thematically consistent long image sequences based on the text prompts. In particular, StoryDiffusion cannot convey interactions between multiple characters correctly (*e.g.*, "*turtle lifting the fishbone trophy*" in the *34-th* frame and "*rabbit running past the camel*" in the *46-th* frame), whereas Story-Adapter visualizes the interactions between the characters accurately while maintaining subject consistency.

Computational Cost Comparison. We evaluate the computational cost of single-image generation using CSA in StoryDiffusion (Zhou et al., 2024) and the proposed GRCA with varying numbers of reference images, under the base attention setting for fair comparison. FLOPs are calculated within the diffusion model UNet. As shown in Fig. 8, as the number of reference images increases, StoryDiffusion experiences a significant rise in computation in terms of FLOPs, while Story-Adapter and Story-AdapterXL are slightly affected. This demonstrates the potential of modeling on global embeddings as in GRCA to efficiently sustain global story semantics for long story visualization.

4.5 Ablation Study

Global Reference Cross-Attention. We ablate the effect of global semantics modeling by GRCA for long story visualization. Specifically, for each image visualization in the sequence, we only use the single reference image at the corresponding index during the iteration as guidance. By establishing a global comprehension of the story for the diffusion model, Story-Adapter maintains the semantic consistency in the generated image sequence (Tab. 3 and Fig. 7).

Iterative Paradigm. We conduct ablation experiments to evaluate the effect of the proposed iterative paradigm for long story visualization and to validate our linear weighting strategy compared to the fixed weight factors. As shown in Tab. 3 and Fig. 6, the iterative paradigm improves generation quality for fine-grained interactions and semantic consistency. This is mainly because the iterative paradigm offers a global view of the entire story, thus reducing error accumulation and alleviating the propagation of the reference image flaws. A fixed weight factor of 0.3 minimally impacts visualization during iteration, while a fixed factor of 0.5 leads to excessive consistency in the image sequence. This enables flexibility within the iterative paradigm.

Figure 7: Qualitative ablation studies of initialization and GRCA. Zoom in for a better view.

Table 3: Quantitative ablation studies of the design choices of Story-Adapter.

Setting	CLIP-T ↑	aCCS ↑	aFID ↓
w/o Initialization	0.302	0.788	90.30
w/o GRCA	0.319	0.740	97.86
w/o Iteration Paradigm	**0.322**	0.757	105.17
Iteration Paradigm, $\lambda = 0.3$	0.320	0.760	101.55
Iteration Paradigm, $\lambda = 0.5$	0.261	0.753	**81.72**
GRCA	**0.322**	0.757	105.17
CSA	0.315	0.768	102.44
Ours	0.318	**0.802**	94.30

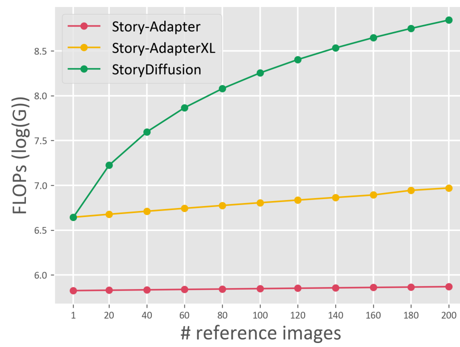

Figure 8: Computational cost of single image generation under different number reference images.

Initialization. To ablate the effect of the proposed initialization, we use a sequence of images consisting of the characters as reference images (*i.e.*, *w/o* initialization). Tab. 3 shows that when removing the proposed initialization, there is a significant decrease in the image-text alignment of Story-Adapter in terms of CLIP-T. Fig. 7 illustrates that without initialization, the diffusion model fails to generate the required objects according to text prompts, *e.g.*, "*nightingale*" and "*robot*".

GRCA vs CSA. We investigate GRCA and CSA in Tab. 3 and Fig. 7, using the outputs of the first iteration from Story-Adapter and StoryDiffusion, respectively. Though GRCA generates less visual consistency during the first iteration than CSA in terms of aCCS and aFID in Tab. 3, GRCA's global comprehension improves the consistency of multiple characters throughout stories shown in Fig. 7. For example, GRCA effectively preserves the consistency of emerging characters (*e.g.*, "*the character 1900*") while CSA fails.

5 Conclusions and Discussions

We introduce Story-Adapter, an *iterative framework* that adapts pre-trained Stable Diffusion models for long story visualization. By using the generated images from previous iterations as references, our method maintains semantic consistency and enhances generative quality for fine-grained interactions throughout the story, effectively reducing error accumulation and avoiding the propagation of flaws. For efficiency, we propose a plug-and-play Global Reference Cross-Attention (GRCA) module, which utilizes global image embeddings to reduce computational costs while preserving essential image information flow. Extensive experiments demonstrate that Story-Adapter outperforms existing methods on the regular-length story visualization dataset, and shows strong results in long story visualization. These findings highlight the potential of our iterative paradigm to advance the quality and coherence of text-to-image story visualization.

Ethical Concerns. All authors of this work have read and commit to adhering to the ICLR Code of Ethics.

Reproducibility. To ensure reproducibility, we provide pseudocode in Algo. 1 and implementation details in Sec. 4.2. The full code can be found in the *Supplementary Material*.

References

Andrew Brock. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Hong Chen, Rujun Han, Te-Lin Wu, Hideki Nakayama, and Nanyun Peng. Character-centric story visualization via visual planning and token alignment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8259–8272, 2022.

Junhao Cheng, Baiqiao Yin, Kaixin Cai, Minbin Huang, Hanhui Li, Yuxin He, Xi Lu, Yue Li, Yifei Li, Yuhao Cheng, et al. Theatergen: Character management with llm for consistent multi-turn image generation. *arXiv preprint arXiv:2404.18919*, 2024.

Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7346–7356, 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.

Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7323–7334, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning*, pp. 13916–13932. PMLR, 2023.

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10124–10134, 2023.

Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2151–2162, 2023.

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*, pp. 1931–1941, 2023.

Bowen Li. Word-level fine-grained story visualization. In *European Conference on Computer Vision*, pp. 347–362. Springer, 2022.

Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6329–6338, 2019.

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*, pp. 8640–8650, 2024.

Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. Intelligent grimm-open-ended visual storytelling via latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6190–6200, 2024.

Haoyu Lu, Guoxing Yang, Nanyi Fei, Yuqi Huo, Zhiwu Lu, Ping Luo, and Mingyu Ding. Vdt: General-purpose video diffusion transformers via mask modeling. *arXiv preprint arXiv:2305.13311*, 2023.

Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2837–2845, 2021.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Adyasha Maharana and Mohit Bansal. Integrating visuospatial, linguistic and commonsense structure into story visualization. *arXiv preprint arXiv:2110.10834*, 2021.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

OpenAI. GPT-4o system card, 2024. URL `https://openai.com/index/gpt-4o-system-card/`.

Xichen Pan, Pengda Qin, Yuhong Li, Hui Xue, and Wenhu Chen. Synthesizing coherent story with auto-regressive latent diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2920–2930, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

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*, pp. 10684–10695, 2022.

Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10219–10228, 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*, pp. 22500–22510, 2023.

Simo Ryu. Low-rank adaptation for fast text-to-image diffusion fine-tuning. *Low-rank adaptation for fast text-to-image diffusion fine-tuning*, 2023.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Fei Shen, Hu Ye, Sibo Liu, Jun Zhang, Cong Wang, Xiao Han, and Wei Yang. Boosting consistency in story visualization with rich-contextual conditional diffusion models. *arXiv preprint arXiv:2407.02482*, 2024.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Ming Tao, Bing-Kun Bao, Hao Tang, Yaowei Wang, and Changsheng Xu. Storyimager: A unified and efficient framework for coherent story visualization and completion. *arXiv preprint arXiv:2404.05979*, 2024.

Wen Wang, Canyu Zhao, Hao Chen, Zhekai Chen, Kecheng Zheng, and Chunhua Shen. Autostory: Generating diverse storytelling images with minimal human effort. *arXiv preprint arXiv:2311.11243*, 2023.

Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. *arXiv preprint arXiv:2112.07804*, 2021.

Haiyang Xu, Yu Lei, Zeyuan Chen, Xiang Zhang, Yue Zhao, Yilin Wang, and Zhuowen Tu. Bayesian diffusion models for 3d shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10628–10638, 2024.

Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Fresco: Spatial-temporal correspondence for zero-shot video translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8703–8712, 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*, 2023.

Ge Yuan, Xiaodong Cun, Yong Zhang, Maomao Li, Chenyang Qi, Xintao Wang, Ying Shan, and Huicheng Zheng. Inserting anybody in diffusion models via celeb basis. *arXiv preprint arXiv:2306.00926*, 2023.

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022.

Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6027–6037, 2023.

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

A PARADIGMS

Existing story visualization methods usually employ the Auto-Regressive (AR) or Reference-Image (RI) paradigms. In this work, we propose a novel iterative paradigm for story visualization. Next, we will discuss different story visualization paradigms in detail.

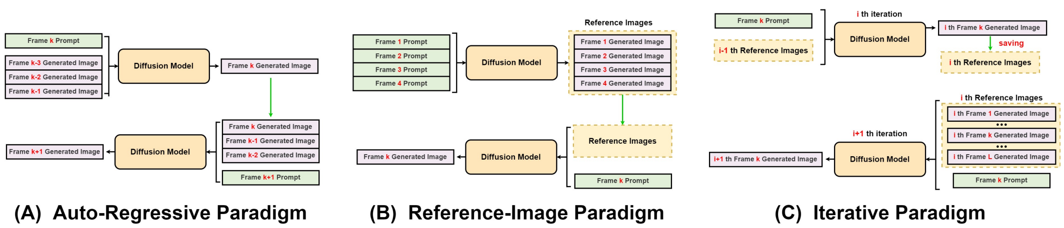

Figure 9: Different paradigms for story visualization. Zoom in for a better view.

A.1 AUTO-REGRESSIVE PARADIGM

Setting. As shown in Fig. 9, AR paradigm-based methods typically use a limited number of previous frames and the corresponding text prompt of the current frame to guide current image generation. This helps the methods maintain semantic consistency between consecutive frames.

Discussion. However, the AR paradigm cannot consider future frames when synthesizing the current image, which makes the AR paradigm only maintain semantic consistency in neighboring frames but not throughout the story. Besides, the AR paradigm easily suffers from error accumulation. Therefore, the image quality of the AR paradigm gets worse as the length of the story increases.

A.2 REFERENCE-IMAGE PARADIGM

Setting. RI paradigm-based methods employ the beginning visualized frames as reference images to guide the visualization of the rest of the story when performing long story visualization (see Fig. 9). Bootstrapping based on fixed reference images helps the methods to effectively maintain identity consistency in long story visualizations.

Discussion. However, such a setup ignores the consistency of emerging characters in the story, and all visualizations are affected by flaws in the reference images. Both issues affect the quality of long story visualizations with the RI paradigm.

A.3 ITERATIVE PARADIGM

Setting. To address the aforementioned limitations, we propose an iterative paradigm in Story-Adapter (Fig. 9). We constantly consider all generated images in the previous iteration with an iterative mechanism and model on the global embeddings. Specifically, when generating for the k_{th} image, we propose to implement Global Reference Cross-Attention (GRCA) on global embeddings from all generated images in the previous iteration.

Discussion. By using all generated images from the previous iteration as reference images to guide the current generation, we effectively maintain semantic consistency throughout the story. Moreover, all the generated images as references are updated through each iteration. Taken together, the iterative paradigm effectively avoids the influence of defects in some reference images.

B SUBJECT-CONSISTENT GENERATION COMPARISON

In the evaluation phase, we employ GPT-4o (OpenAI, 2024) according to the settings of StoryDiffusion (Zhou et al., 2024) to generate 20 character descriptions and 100 specific activity descriptions, respectively. We combine them as 2000 test descriptions, to compare Story-Adapter and subject-

Figure 10: Qualitative comparison of subject-consistent image generation methods.

consistent image generation baselines, including IP-Adapter (Ye et al., 2023), PhotoMaker (Li et al., 2024), and StoryDiffusion (Zhou et al., 2024).

Quantitative Evaluation. For quantitative comparisons on subject-consistent image generation, we employ CLIP text-to-image similarity (CLIP-T) and image-image similarity (CLIP-I) to measure consistency between the character images and generated images. Tab. 4 shows that Story-Adapter achieves SoTA performance in terms of both quantitative metrics, which demonstrates Story-Adapter's ability to generate subject-consistent image sequences based on text prompts or image prompts.

Table 4: Quantitative comparison with subject-consistent image generation methods.

Method	CLIP-T ↑	CLIP-I ↑
IP-Adapter (Ye et al., 2023)	0.307	0.872
Story-Adapter (Ours)	0.326	0.877
IP-AdapterXL (Ye et al., 2023)	0.312	0.879
PhotoMaker (Li et al., 2024)	0.317	0.880
StoryDiffusion (Zhou et al., 2024)	0.330	0.882
Story-AdapterXL (Ours)	**0.332**	**0.884**

Qualitative Evaluation. Fig. 10 shows the qualitative comparison results. Story-Adapter generates higher-quality images in subject-consistency and detailed interactions. In contrast, IP-Adapter fails to generate correctly, *e.g.*, "*paper*", "*whiteboard*", and "*chainsaw*". PhotoMaker cannot generate images consistently, *e.g.*, maintaining details of the attire. Despite accurately generating content according to text prompts with visual consistency, StoryDiffusion suffers from visualizing complex details due to lacking global story comprehension. By incorporating a global story view in our iterative paradigm, Story-Adapter can maintain visual consistency, especially in details throughout the story.

Table 5: Human evaluation comparison of subject-consistent image generation, regular-length story visualization, and long story visualization. The best is highlighted in red.

Model	Align. ↑	Inter. ↑	Cons. ↑	Qual. ↑	Pref. ↑
Subject-Consistent Image Generation					
IP-Adapter (Ye et al., 2023)	2.51	3.27	4.58	4.33	4.19
IP-AdapterXL (Ye et al., 2023)	2.66	3.36	4.72	4.51	4.26
PhotoMaker (Li et al., 2024)	3.79	4.18	4.25	4.47	4.11
StoryDiffusion (Zhou et al., 2024)	4.15	4.28	4.50	4.54	4.48
Story-Adapter	4.02	4.20	4.41	4.39	4.33
Story-AdapterXL	4.20	4.35	4.58	4.61	4.54
Regular-Length Story Visualization					
SDM (Rombach et al., 2022)	4.11	2.37	2.01	4.17	1.14
Prompt-SDM (Rombach et al., 2022)	4.03	3.49	1.99	4.40	1.26
Finetuned-SDM (Rombach et al., 2022)	3.35	3.82	2.15	3.41	1.60
AR-LDM (Pan et al., 2024)	3.08	3.64	2.90	2.64	2.05
StoryGen (Liu et al., 2024)	3.72	4.17	3.83	3.79	3.39
StoryDiffusion (Zhou et al., 2024)	3.96	4.48	4.52	4.24	4.37
Story-Adapter	3.89	4.21	4.36	4.09	4.10
Story-AdapterXL	4.06	4.60	4.74	4.53	4.62
Long Story Visualization					
AR-LDM (Pan et al., 2024)	3.30	3.68	3.42	2.15	3.27
StoryGen (Liu et al., 2024)	3.51	4.06	3.88	2.72	3.51
IP-Adapter (Ye et al., 2023)	3.79	4.27	4.30	4.19	4.06
IP-AdapterXL (Ye et al., 2023)	3.83	4.23	4.61	4.47	4.11
StoryDiffusion (Zhou et al., 2024)	4.16	4.30	4.53	4.33	4.35
Story-Adapter	3.97	4.15	4.42	4.19	4.29
Story-AdapterXL	4.35	4.47	4.70	4.62	4.65

C Human Evaluation

Setting. To complement the evaluation metrics to accurately reflect the quality of the generated stories, we involve human evaluation to further compare Story-Adapter and baselines. Referring to the setting in StoryGen (Liu et al., 2024), we invite participants to rate various aspects: text-image alignment (Align.), character interaction (Inter.), content consistency (Cons.), image quality (Qual.), and preference (Pref.) on a scale from 1 to 5. The higher the better.

Results. Tab. 5 shows that our Story-Adapter receives more preference from the participants. It is worth noting that although IP-Adapter receives higher scores for consistency in the subject-consistent image generation task, Story-Adapter is more favored in text-image alignment and generating character interactions. For regular-length and long story visualization, Story-Adapter is more preferred compared to baselines in most evaluation aspects, especially visual consistency and capability to generate character interactions. This is aligned with the quantitative measurement.

D More Iterations

Setting. In this section, we compare results on different iterations in the iterative paradigm and investigate the impact of longer iterations on story visualization. Specifically, we study the visualization results in the initialization, 1_{st}, 5_{th}, 10_{th}, and 15_{th} iterations, respectively.

Results. Tab. 6 shows that as iteration increases, Story-Adapter achieves significant improvement in visual consistency (aCCs and aFID) while text-image alignment (CLIP-T) drops slightly. This further demonstrates the contribution of the iterative paradigm to the semantic consistency of the overall story. However, we also note that a further increase in iterations harms text-image alignment, with limited gain in visual content consistency. This indicates that while Global Reference Cross-

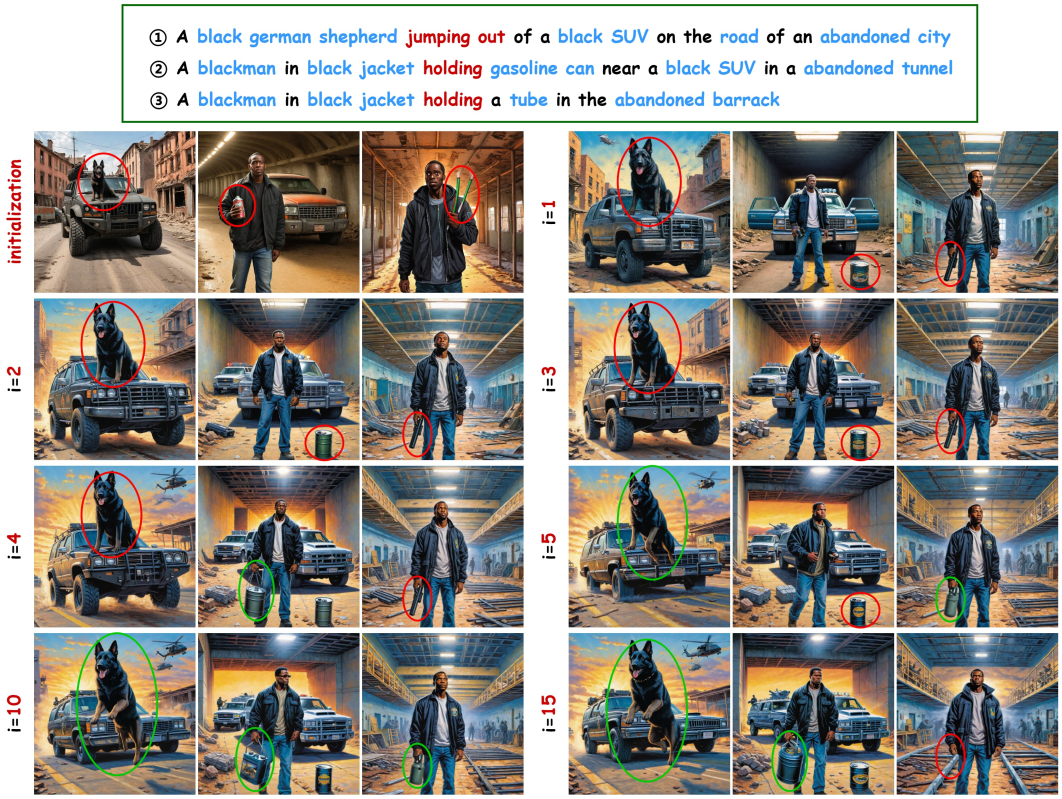

Figure 11: Story visualization results from different iterations by Story-Adapter. Accurate interactions are denoted in green, wrong or missing ones are in red.

Attention (GRCA) effectively improves the content consistency of the long story, the increasing weighting factor of GRCA during the iterations poses a challenge to aligning the text prompts.

Fig. 11 demonstrates a significant improvement in generative quality for fine-grained interactions as the iteration proceeds. The iterative paradigm effectively alleviates the diffusion model's limitations on complex interaction generation by continuously creating input channels for text prompts. But more iterations wouldn't improve the generation quality further. Therefore, 10 iterations in the iterative paradigm is an optimal choice based on the quantitative and qualitative experiments.

Table 6: Quantitative comparison of multiple iterations.

Iteration	CLIP-T ↑	aCCS ↑	aFID ↑
initialization	0.330	0.502	214.94
1_{th} iteration	0.322	0.757	105.17
5_{th} iteration	0.319	0.783	100.81
10_{th} iteration	0.306	0.840	91.35
15_{th} iteration	0.297	0.848	90.62

E More Visualization Results

In this section, we provide more visualization results from Story-Adapter and the baselines.

E.1 Visual Comparison

We compare the long story visualization results of representative work with AR-based, RI-based, and iterative paradigms, respectively. Specifically, Fig. 12, Fig. 13, and Fig. 14 show the generated results of the same "*Pianist*" story from the proposed Story-Adapter (iterative), StoryGen (Liu et al., 2024) (AR-based), and StoryDiffusion (Zhou et al., 2024) (RI-based), respectively.

Results. Fig. 12 shows that the visualization quality from StoryGen constantly gets worse as the length of the story increases. In Fig. 13, StoryDiffusion maintains high visual quality throughout

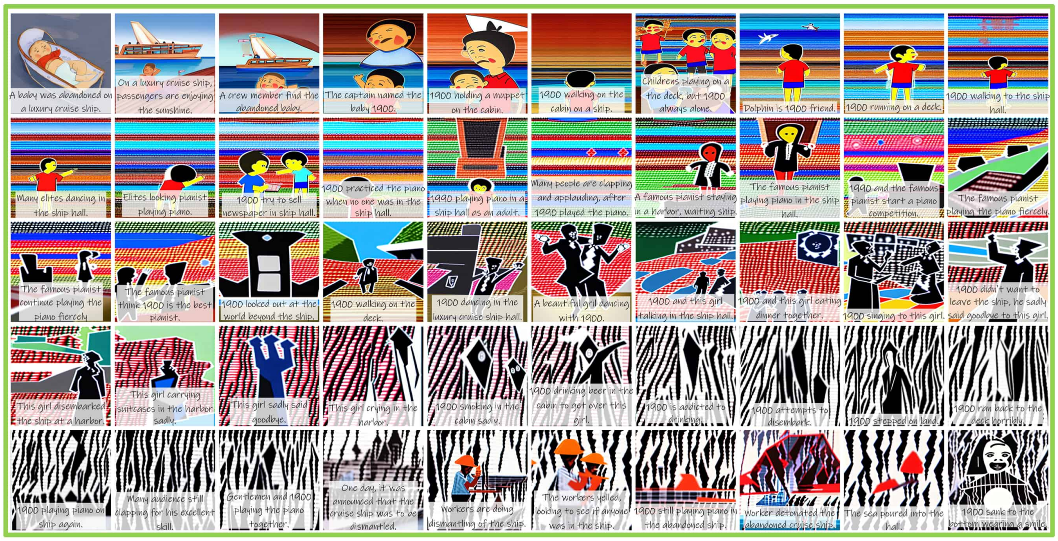

Figure 12: Visualization results of StoryGen for the "*Pianist*" story. Zoom in for a better view.

Figure 13: Visualization results of StoryDiffusion for the "*Pianist*" story. Zoom in for a better view.

the story, but it suffers from the flaw in the beginning frame that serves as the reference image, *e.g.*, "*closed-eye"*. In addition, the subject "*the character 1900*" is not consistently generated as baby, kid, and adult. In contrast, our Story-Adapter effectively achieves high-quality story visualization and addresses the aforementioned limitations (see Fig. 14).

E.2 Longer Story Visualization Results

In Fig. 15, we show the visualization results of the long story (up to 100 frames).

E.3 Different Style

We provide the long story visualization results from Story-Adapter in a realistic style in Fig. 16. The experiment results suggest that Story-Adapter can be applied to different visual styles as well.

Figure 14: Visualization results of our Story-Adapter for “*Pianist*”. Zoom in for a better view.

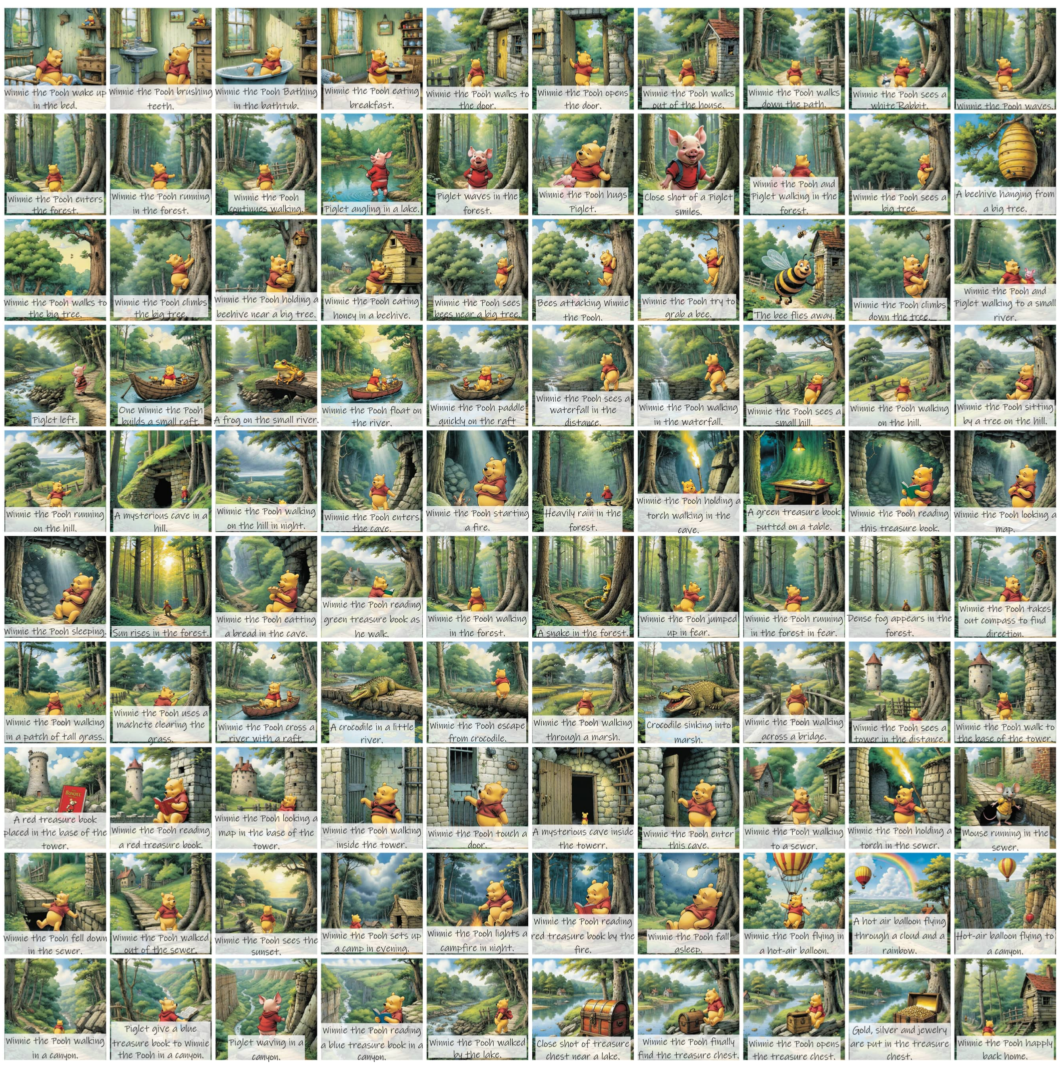

Figure 15: Our long story visualization results for “*Winnie the Pooh*”. Zoom in for a better view.

Figure 16: Our realistic style story visualization results for "*loyal dog*". Zoom in for a better view.