



Coln: A Lightweight and Effective Framework for Story Visualization and Continuation

Ming Tao
Nanjing University of Posts and
Telecommunications
Nanjing, China
Peng Cheng Laboratory
Shenzhen, China
mingtao2000@126.com

Bing-Kun Bao*
Nanjing University of Posts and
Telecommunications
Nanjing, China
Peng Cheng Laboratory
Shenzhen, China
bingkunbao@njupt.edu.cn

Hao Tang
Peking University
Beijing, China
haotang@pku.edu.cn

Yaowei Wang
Peng Cheng Laboratory
Shenzhen, China
wangyw@pcl.ac.cn

Changsheng Xu
Peng Cheng Laboratory
Shenzhen, China
Institute of Automation, Chinese
Academy of Sciences
Beijing, China
University of the Chinese Academy of
Sciences
Beijing, China
csxu@nlpr.ia.ac.cn

Abstract

Story visualization aims to generate realistic and coherent images based on multi-sentence stories. However, current methods face challenges in achieving high-quality image generation while maintaining lightweight models and a fast generation speed. The main issue lies in the two existing frameworks. The independent framework prioritizes speed but sacrifices image quality with the non-collaborative image generation process and basic GAN-based learning. The autoregressive framework modifies the large pretrained text-to-image model in an auto-regressive manner with additional history modules, leading to large model size, resource-intensive requirements, and slow generation speed. To address these issues, we propose a lightweight and effective framework, namely CoIn. Specifically, we introduce a Context-aware Story Generator to predict shared context semantics for each image generator. Additionally, we propose an Intra-Story Interchange module that allows each image generator to exchange visual information with other image generators. Furthermore, we incorporate DINOv2 into the story and image discriminators to assess the story image quality more accurately. Extensive experiments show that our CoIn keeps the model size and generation speed of the independent framework, while achieving promising story image quality.

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0686-8/24/10
<https://doi.org/10.1145/3664647.3680873>

CCS Concepts

• Computing methodologies → Computer vision.

Keywords

Generative Models, Story Visualization, Story Continuation

ACM Reference Format:

Ming Tao, Bing-Kun Bao, Hao Tang, Yaowei Wang, and Changsheng Xu. 2024. Coln: A Lightweight and Effective Framework for Story Visualization and Continuation. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*, October 28-November 1, 2024, Melbourne, VIC, Australia. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3664647.3680873>

1 Introduction

Story visualization is a challenging task that aims to generate coherent and visually appealing story images based on a sequence of story descriptions. The task requires the model to capture the essence of the story and the relationships between different story descriptions in the story and then translate them into realistic images that tell the story in a coherent manner. Some recent works extend the task to story continuation which provides an initial frame of the story image and generates subsequent frames based on the remaining story descriptions. Due to the practical application, story visualization and continuation have recently become an active research area [16, 18, 20, 22, 26].

Existing story visualization models [2, 18, 20, 21, 26, 29, 39] have significantly advanced in generating story images. However, they still face difficulties in achieving both high-quality image generation and maintaining lightweight models with a fast generation speed. The existing frameworks for story visualization exhibit inherent limitations that hinder their effectiveness and efficiency. The independent framework [16, 18, 20, 21, 39], which prioritizes generation speed, compromises image quality due to its reliance

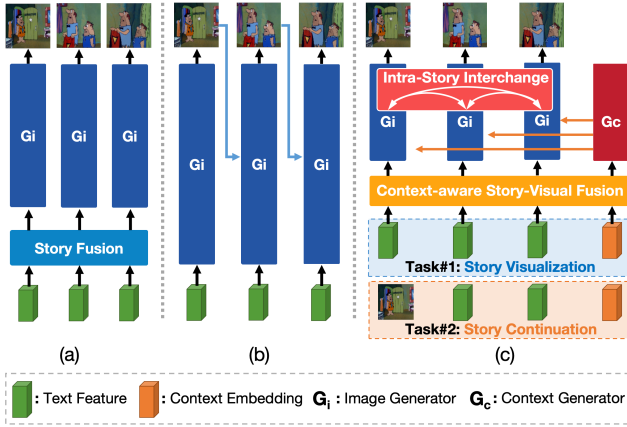


Figure 1: (a) Existing models [16–18] always employ an independent generator design. (b) Recent models [22, 26] employ large pretrained generative models and synthesize story images frame-by-frame. (c) Our CoIn proposes a novel contextualize and interchange story generator.

on independent generators and basic Generative Adversarial Networks (GANs). As depicted in Figure 1(a), in the independent framework, the fusion of textual information for each frame occurs only before it is passed to the image generator. The image generator operates independently during the image generation process without any information exchange, thereby increasing the difficulty of synthesizing coherent story images. Conversely, the autoregressive framework [22, 26, 29] utilizes large pretrained text-to-image models [31, 32, 34], transforming them into an auto-regressive approach for generating story images (see Figure 1(b)). It incorporates an additional history module to retain and encode the generated historical frames. However, as shown in Figure 2, this approach results in significantly larger model sizes, high resource requirements for training and inference, and the frame-by-frame design further slows the generation speed. Moreover, both independent and autoregressive frameworks do not distinguish between local story and global context information. Consequently, the image generator needs to implicitly decode the entangled local story and global context features and synthesize the corresponding visual features simultaneously.

To achieve high-fidelity and coherent story image generation while maintaining reasonable computational requirements and generation speed, we propose the CoIn (Contextualize and Interchange) for story visualization and Continuation. As shown in Figure 2, compared to WL-SV[16] which is based on the independent framework, our CoIn boosts the image quality significantly while keeping the generation speed and parameters of the story generator. Compared with StoryDALL-E[22] and AR-LDM[26] which adopt pretrained text-to-image model-based autoregressive framework, our CoIn achieves competitive results with $\sim 1000\times$ and $\sim 360\times$ faster synthesis speed and 5.0% and 4.6% generator parameters.

With careful consideration of the characteristics of story visualization and continuation tasks, we develop the novel CoIn framework tailored to meet their specific requirements. Unlike previous

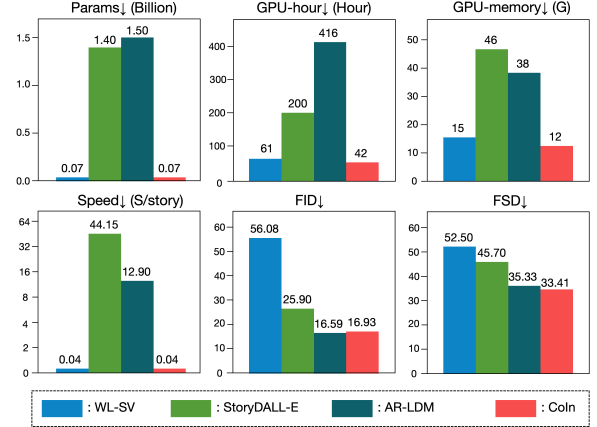


Figure 2: Comparing with the state-of-the-art models on Pororo-SV [14]. Our CoIn achieves comparable image quality metrics while preserving the fast speed, small model parameters, and low resource requirements.

models, our CoIn distinguishes between local story and global context to ensure the overall coherence of generated story images. As depicted in Figure 1(c), we propose the Context-aware Story Generator which consists of Contextual Story-Visual Fusion, local story image generators, and a shared story context generator. The Contextual Story-Visual Fusion fuses the text and source image features for each image generator and extracts the global context information for the context generator. Explicit extraction of global context and specialized context generator alleviate the difficulty for the Context-aware Story Generator in handling both the global context and the local story information simultaneously. Additionally, our CoIn supports information exchange between different frames of the same story during the generation process to ensure a consistent visual appearance. We propose an Intra-story Interchange module, which allows different image generators of the same story to exchange visual features, promoting visual consistency across the generated images. Lastly, we introduce DINO-based discriminators, which incorporate the pretrained DINOv2 [25] visual backbone into the story and image discriminators to assess the quality of story images more accurately.

Overall, our contributions can be summarized as follows:

- We introduce a fast, lightweight, and effective framework for story visualization and continuation to synthesize high-quality and coherent story images.
- We propose a Context-aware Story Generator that explicitly extracts global context information and shares context features with all image generators during the synthesis process.
- We propose the Intra-story Interchange module, which enables the exchange of visual features between intra-story image generators to enhance visual consistency.
- We propose the DINO-based story and image discriminators, which assess the quality of story images more accurately.
- Extensive qualitative and quantitative experiments demonstrate that the proposed CoIn achieves promising results with fast generation speed and a much smaller model size.

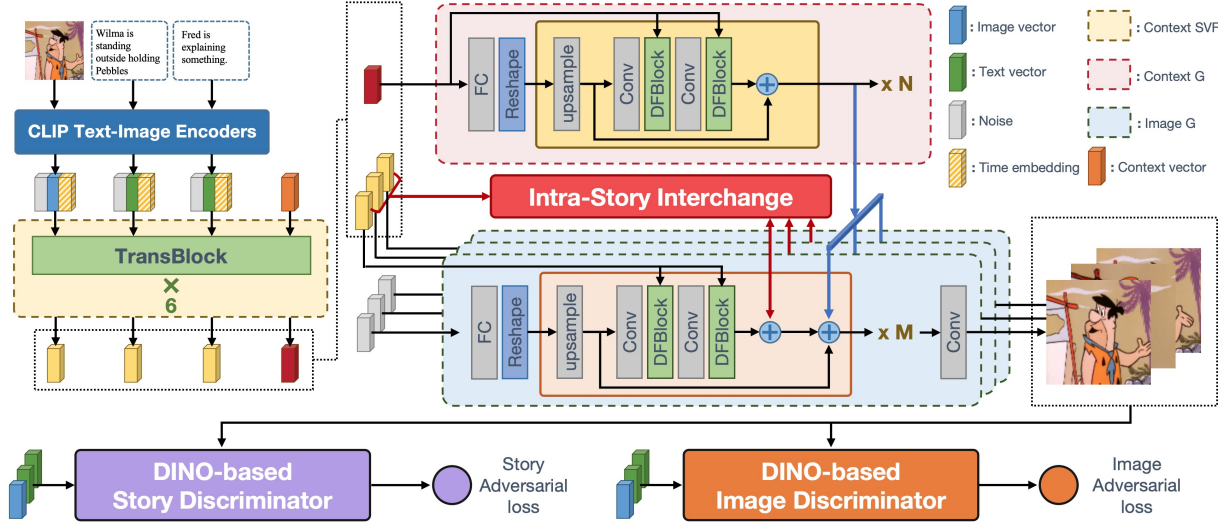


Figure 3: The architecture of the proposed CoIn for Story Visualization and Continuation. Our CoIn proposes a novel contextualize and interchange framework to share the global context visual semantics and interchange local visual features.

2 Related Work

Text-to-Image Synthesis. Recent advancements in text-to-image synthesis have primarily focused on three main frameworks: Generative Adversarial Networks (GANs) [8], autoregressive models [4, 5, 31], and diffusion models [24, 30, 32, 49]. GANs, such as StackGAN [47, 48], AttnGAN [44], DM-GAN [50], DF-GAN [42], ISF-GAN[37], GALIP [41], and StyleGAN-T [36] employ adversarial training strategies between generators and discriminators to generate high-quality images from text descriptions. Large pretrained autoregressive models, including DALL-E [31], Make-A-Scene [6], Parti [45], and VAR[43], have demonstrated scalability and proficiency in synthesizing images. These models generate images by sequentially autoregressively predicting pixel tokens based on previously generated tokens. Diffusion models [3, 12, 13, 23, 38], such as VQ-Diffusion [9], GLIDE [24], DALL-E2 [30], Latent Diffusion Models [32], Imagen [35], eDiff-I [1], and SDXL [27], have gained significant interest.

Story Visualization and Continuation. StoryGAN [18] was the pioneering work that introduced the task of story visualization and proposed a GAN-based sequence generation model. It consisted of a deep RNN-based context encoder and two discriminators for images and stories. Subsequent works have built upon and refined this network. For example, CP-CSV [39] introduced a new foreground segmentation module to optimize the consistency between characters and backgrounds in the story. DUCO [21] and VLC [20] [17] enhanced semantic consistency through dual learning, with DUCO considering inter-image sequence consistency via copy-transform and VLC focusing on textual information by incorporating external common-sense knowledge. WL-SV [16] and Clustering GAN [17] simplified the two-stage GAN network of StoryGAN and further improved story quality through fine-grained word-level features and clustering learning, respectively. These GAN-based models adopt the independent framework. Conversely, recent approaches such as pre-trained DALL-E-based StoryDALL-E [22], and pre-trained

stable diffusion-based AR-LDM [26] have been developed for story visualization. Notably, StoryDALL-E [22] also introduced a new task called story continuation, which involves generating unseen plots and characters based on a given source frame. AR-LDM [26] and Make-A-Story [29] can handle both the story visualization and story continuation tasks. StoryImager [40] enhances the storyboard generative ability inherited from the pre-trained diffusion model. Recently, there are some works focusing on open-ended story visualization. TaleCrafter [7] pioneered an interactive approach with sketch and layout controls for story visualization. CogCartoon [51] introduced a character-plugin generation to minimize data and storage requirements. Intelligent Grimm [19] curated a diverse open-ended story dataset from YouTube and e-books.

3 The Proposed Method

To synthesize high-quality story images while maintaining reasonable computational requirements and generation speed, we propose: (i) a Context-aware Story Generator (Context-aware SG) that explicitly captures and leverages global context information by sharing contextual features with all image generators throughout the synthesis process. (ii) an Intra-story Interchange module (Intra-SI) that facilitates the exchange of visual features among image generators within the story to improve visual consistency. (iii) a pair of DINO-based story and image Discriminators (DINO-based D) that enhance the accuracy in assessing the quality of story images. In the following section, we first present the overall structure. Then, we introduce Context-aware SG, Intra-SI, and DINO-based D in detail.

3.1 Model Overview

As illustrated in Figure 3, CoIn consists of CLIP Text and Image Encoders [28], a Context-aware SG with a Contextual Story-Visual Fusion module (Contextual SVF), an Intra-SI module, a pair of DINO-based story and image discriminators. The CLIP Text-Image Encoder encodes text descriptions and source story images into text

and image vectors. The Contextual SVF takes text and image vectors, Gaussian noise, and a learnable context vector as input and outputs the fused story embedding and extracted context embedding. Afterward, the Context-aware SG utilizes the noise, fused story embedding, and extracted context embedding to synthesize story images. The story generator comprises two classes of sub-generators: image generator and context generator. The image generator employs fused story embedding, while the context generator utilizes extracted context embedding. In the image generator, the noise is passed through a Fully Connected (FC) layer and reshaped to $(4, 4, 256)$. Then, a series of upsampling blocks upsamples the image features and incorporates fused story embedding into the generation process. The Intra-SI collects synthesized image features from different image generators within the same story and performs cross-frame visual feature exchange. The contextual visual features synthesized by the context generator are added to the output features in the upsampling block. Finally, a convolution layer converts the image features into RGB images. DINO-based Story Discriminator and Image Discriminator take the generated and real story images as input and extract their visual features through the frozen DINOv2 visual backbone [25]. Then, they assess the quality of story images based on the extracted visual features along with text and image vectors. By distinguishing synthesized story images from real ones, the discriminators promote the story generator to synthesize higher-quality story images.

3.2 Context-aware Story Generator

In this section, we detail the proposed Context-aware SG. The Context-aware SG shows a novel story generator that explicitly captures and leverages global context information by sharing contextual features with all image generators during the synthesis process. The Context-aware SG comprises three key components: a Contextual Story-Visual Fusion module, a Shared Context Generator, and Story Image Generators.

Contextual Story-Visual Fusion. The Contextual SVF aims to fuse the given text and image features and extract the context information from given conditions. However, the inputs of the story visualization and continuation are different. For story visualization, the input is a sequence of encoded text features of story descriptions, where each description describes the visual content of the corresponding frame in the story. For story continuation, the input consists of two parts: the encoded visual features of source story images and the encoded text features of story descriptions corresponding to the last frames. As shown in Figure 3, the Contextual SVF takes the pre-trained CLIP model [28] to encode source story images and story descriptions. Then, the encoded text and image vectors are concatenated with a Gaussian noise vector and combined with the time embedding. This integration allows for image variance in the synthesized story images, while the time embedding incorporates temporal information. Additionally, a learnable context vector is appended to the inputs. This vector is specifically designed to extract context information from both source story images and descriptions.

Unlike previous models that implicitly fuse context information into the story embedding, our Contextual SVF explicitly extracts context information. This explicit extraction enables Contextual

SVF to separate the global context information from the local story features. The Contextual SVF consists of six stacked transformer blocks, each containing eight attention heads. This transformer-based architecture facilitates parallel processing and ensures scalability for story lengths of varying sizes. With contextual story-visual fusion and explicit context extraction in place, the Context-aware Story Generator obtains fused story embeddings and extracted context embeddings.

Story Image Generator. The structure of the Story Image Generator is shown in Figure 3, as highlighted by the blue dashed box. The Gaussian noise vector is fed into the FC layer and reshaped to $(4, 4, 256)$ as an initial image feature. Then we apply a sequence of Story Fusion blocks (SF-BLKs) to upsample the image feature to the target image size and fuse story and context information into synthesized story images. The SF-BLK is composed of one upsample layer, two convolution layers (Conv), and two deep text-image fusion blocks (DFBlock) [42]. As shown in recent text-to-image works [42], DFBlock can fuse text and image features effectively through deep affine transformations. Thus, we employ it to fuse story features and intermediate visual features. After the fusion of the story and visual features, the SF-BLK sends the fused visual features to the Intra-SI and receives the exchanged visual features from it. The exchanged visual features are added to the fused visual features. Lastly, the SF-BLK combines the context visual features generated by the Context Generator and the original features from the shortcut before fusion. Through SF-BLK, the output visual features contain local story information, global context information, and cross-frame visual information. There are M SF-BLKs stacked in the Image Generator, and one convolution layer converts the fused image features into RGB images.

Shared Context Generator. The structure of the Shared Context Generator is also shown in Figure 3, as highlighted by the pink dashed box. To ensure the generation of story images with consistent overall context, the Context Generator and Image Generator are designed with a similar structure, facilitating the effective fusion of context information. However, unlike Image Generator, Context Generator only takes the context embedding as input. There are N Context Fusion blocks (CF-BLKs) stacked in the Context Generator. As SF-BLKs, we adopt the DFBlock[42] to fuse the extracted context embedding at each CF-BLK. Then, the context visual features generated by the Context Generator are shared with all Image Generators. The Image Generator and Context Generator do not share parameters, allowing for a more thorough learning of image and context generation. To strike a balance between sharing story context and preserving different visual details in each story image, the context visual features are only introduced in the first 3 SF-BLKs. Compared to previous methods, the Context Generator enhances the global context information and alleviates the difficulty of the story generator in synthesizing coherent story images.

3.3 Intra-Story Interchange

To enable feature interchange during the intra-story image generation process, we propose the Intra-SI module. As illustrated in Figure 3 and Figure 4, the Intra-SI takes the fused story embeddings from the Contextual SVF and collects the image visual features from

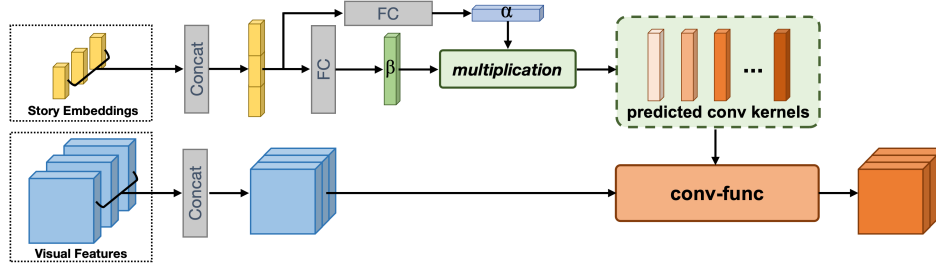


Figure 4: Illustration of the Intra-Story Interchange module which exchanges visual features among image generators.

intra-story image generators. First, Intra-SI concatenates story embeddings and visual features, respectively. The concatenated story embeddings are then fed into two FC layers to predict the vector $\alpha \in \mathbb{R}^C$ and $\beta \in \mathbb{R}^C$, where C is the channel size of the concatenated visual features. Afterward, the vectors α and β are reshaped to $(C, 1)$ and $(1, C)$. We perform matrix multiplication between the reshaped α and β to obtain the predicted 1×1 convolutional kernel weights. Finally, we conduct the 1×1 convolution process on concatenated visual features based on the predicted kernel weights. As part of the Shared Context Generator, the Intra-SI modules are applied only in the first four SF-BLKs to strike a balance between visual appearance consistency and variance. The proposed Intra-SI module in our CoIn framework facilitates interaction among image generators throughout the story generation process. As the Shared Context Generator, the Intra-SI modules are only applied in the first 4 SF-BLKs to balance visual appearance consistency and variance. By incorporating the Intra-SI module, the image generators have the opportunity to exchange visual features among themselves. This exchange of visual features enhances the generated story images, resulting in improved visual appearance consistency.

3.4 DINO-based Discriminators

To improve the efficiency of adversarial learning and the accuracy of story image quality assessment, we propose the DINO-based Discriminator. We first introduce the DINO-based Story Discriminator. As shown in Figure 5, the story image sequence is first encoded by the frozen DINOv2-Small Image Encoder. The visual features from the 2nd, 7th, 9th layers in DINOv2-Small are extracted and further analyzed through two residual blocks. Then we concatenate the output visual features and encode them into a more compact visual feature. We also concatenate the text and image vectors and encode them into a more compact vector. Afterward, we replicate the encoded text-image vector and concatenate it with compact visual features. A story adversarial loss is predicted by two convolution layers to evaluate the story quality. Based on the visual understanding ability of pre-trained DINOv2 [25], the DINO-based Story Discriminator can assess the story quality more accurately. The DINO-based Image Discriminator shares a similar structure but differs in how it processes the input features. Rather than concatenating the visual features of the same story as the whole story visual feature, it concatenates the visual features of each image with their corresponding textual features. This allows the discriminator to assess the quality of each image.

The DINO-based Discriminator is partly inspired by GALIP [41] which employs the pretrained CLIP-ViT [28] to extract visual features in the image discriminator. In this work, we extend it and propose the DINO-based Story and Image Discriminator to assess the story quality. Compared with CLIP-ViT which only utilizes high-level contrastive pretraining, DINOv2 incorporates an additional masked mask to reconstruct masked parts. It enables DINOv2 to extract richer and more detailed visual features. However, we observed that directly using the DINOv2 model resulted in the generation of patch-like patterns in the synthesized images, as shown in Figure 8(a). These patch-like patterns were consistent with the size of the DINOv2 input (16, 16). When we replaced DINOv2 with CLIP-ViT-B/16, the patch-like patterns disappeared. We suspect that this issue may be related to the masked pretraining employed by DINOv2, which might not provide sufficient supervision for smooth transitions between adjacent image patches. Consequently, the generator may have learned to generate adversarial features. To address this issue, we propose the Slight Shift trick. Before feeding the images into DINOv2, we slightly enlarge them and randomly crop an image corresponding to the resolution of DINOv2’s input. This process helps prevent the generator from accurately identifying the patch intervals, thereby avoiding the generation of patch-like patterns.

3.5 Objective Function

We follow previous text-to-image models [41, 42] and employ the hinge loss [46] and MA-GP [41, 42] to stabilize the adversarial training process. Finally, the whole formulation of our CoIn is shown as follows:

$$\begin{aligned}
 L_{D_I} &= -\mathbb{E}[\min(0, -1 + D_I(x, e))] \\
 &\quad - (1/2)\mathbb{E}[\min(0, -1 - D_I(G(z, e), e))] \\
 &\quad - (1/2)\mathbb{E}[\min(0, -1 - D_I(x, \hat{e}))] \\
 &\quad + k\mathbb{E}[(\|\nabla_{c_i} D_I(c_i, e)\| + \|\nabla_e D_I(c_i, e)\|)^p], \\
 L_{D_S} &= -\mathbb{E}[\min(0, -1 + D_S(x, e))] \\
 &\quad - (1/2)\mathbb{E}[\min(0, -1 - D_S(G(z, e), e))] \\
 &\quad - (1/2)\mathbb{E}[\min(0, -1 - D_S(x, \hat{e}))] \\
 &\quad + k\mathbb{E}[(\|\nabla_{c_s} D_S(c_s, e)\| + \|\nabla_e D_S(c_s, e)\|)^p], \\
 L_D &= \lambda L_{D_I} + \gamma L_{D_S}, \\
 L_G &= -\lambda \mathbb{E}[D_I(G(z, e), e)] - \gamma \mathbb{E}[D_S(G(z, e), e)],
 \end{aligned} \tag{1}$$

where z is the Gaussian noise vector; e is the encoded text and image vectors by the CLIP; \hat{e} is the mismatch text and image vectors; x is ground truth story images; G is the Context-aware Story

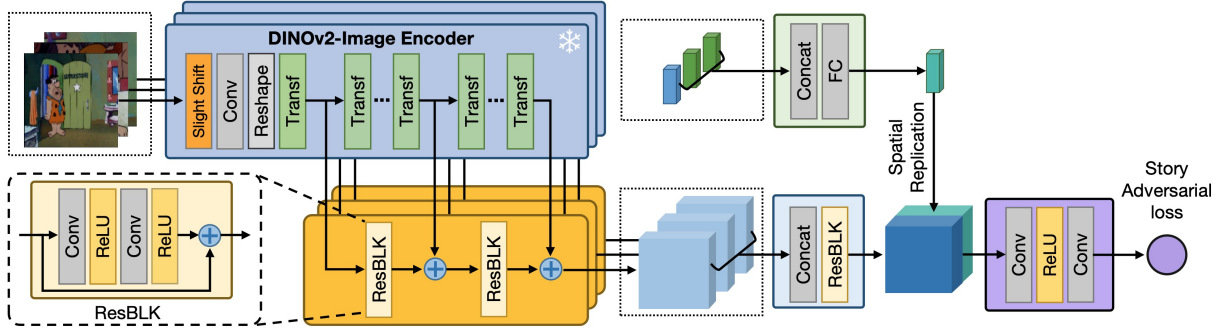


Figure 5: Illustration of the proposed DINO-based Story Discriminator for story visualization and continuation tasks.

Generator; D_I and D_S are the DINO-based Image Discriminator and Story Discriminator; c_i and c_s are two extracted visual features by the frozen DINOv2-Small in Image and Story discriminators; k and p are two hyper-parameters of Matching-Aware Gradient Penalty [42]; λ and γ are two hyper-parameters of image adversarial loss and story adversarial loss.

4 Experiments

4.1 Datasets

We evaluate our approach on two challenging datasets: Pororo-SV [14] and Flintstones-SV [10]. The Pororo-SV dataset consists of 10191 training samples, 2334 validation samples, and 2208 testing samples. The Flintstones-SV dataset consists of 20132 training samples, 2071 validation samples, and 2309 testing samples. Each story in these datasets consists of five consecutive frames, and each image corresponds to a story description. The partition of datasets into training, validation, and testing subsets follows established practices from previous studies [16, 22, 26, 29].

4.2 Training and Evaluation Details

Our method is developed using PyTorch. The resolution of the output images on Pororo-SV and Flintstones-SV is 256×256 . The network is trained 200 epochs and 150 epochs on Pororo-SV and Flintstones-SV. We use the Adam optimizer [15] with $\beta_1=0.0$ and $\beta_2=0.9$ to train our model. We set the learning rate 0.0001 for the generator and 0.0004 for the discriminator. The hyper-parameters of the objective function k , p , λ , γ , are set to 2, 6, 0.5, and 0.5. All models were trained on $8 \times A6000$ GPUs. Following the previous story visualization works [16, 22, 26], we adopt the Fréchet Inception Distance (FID) [11] to evaluate the image quality of synthesized story images, and adopt the Fréchet Story Distance (FSD) [16, 39] to evaluate the consistency of the story image sequence. To compare the efficiency of different models, we measure the generation speed and the number of parameters in the story generator. The generation speed is evaluated on a single A6000 GPU.

4.3 Quantitative Evaluation

To evaluate the performance of our proposed CoIn, we compare it with several state-of-the-art story visualization and continuation methods [16, 20, 22, 26, 39] in Table 1 and Table 2. From Table 1

Table 1: The comparison results of story visualization on the test set of Pororo-SV and Flintstones-SV.

Method	Params	Speed	Pororo-SV		Flintstones-SV	
			FID ↓	FSD ↓	FID ↓	FSD ↓
StoryGAN[18]	-	-	78.64	94.53	90.55	122.71
CP-CSV [39]	-	-	67.76	71.51	-	-
VLC [20]	-	-	94.30	122.07	-	-
WL-SV [16]	0.07B	0.04s	56.08	52.50	72.37	91.30
Make-A-Story[29]	1.40B	14.00s	27.33	51.20	36.55	53.10
AR-LDM [26]	1.50B	14.50s	16.59	35.33	23.59	39.70
CoIn (Ours)	0.07B	0.04s	16.93	33.41	24.52	36.15

Table 2: The comparison results of story continuation on the test set of Pororo-SV and Flintstones-SV.

Method	Params	Speed	Pororo-SV		Flintstones-SV	
			FID ↓	FSD ↓	FID ↓	FSD ↓
StoryDALL-E [22]	1.30B	44.15s	25.90	45.70	26.49	54.30
MEGASToryDALL-E [22]	2.80B	87.40s	23.48	-	23.58	-
Make-A-Story[29]	1.40B	12.15s	22.66	44.22	23.74	52.08
AR-LDM [26]	1.50B	12.90s	17.40	37.52	19.28	43.32
CoIn (Ours)	0.07B	0.04s	18.63	34.73	19.95	39.17

and Table 2, we can observe that our CoIn achieves promising FID and FSD with fast generation speed and a small number of parameters. Compared with WL-SV [16] which is the best model based on the independent framework, our CoIn significantly improves the FID and FSD of Pororo-SV and Flintstones-SV. Compared with StoryDALL-E [22], which utilizes large pre-trained DALL-E models and synthesizes story images in an autoregressive manner, our CoIn framework achieves superior results despite having significantly smaller generator parameters and faster synthesis speed. This demonstrates the effectiveness and efficiency of our approach in generating high-quality story images. Compared with AR-LDM [26] which is an autoregressive model based on the powerful stable diffusion [33], our CoIn has a significantly smaller number of generator parameters but still achieves a competitive performance.

4.4 Qualitative Evaluation

Figures 6 and 7 show examples of visual comparisons between our CoIn and state-of-the-art story visualization and continuation



Figure 6: Comparison of story visualization results between WL-SV, AR-LDM, and our proposed CoIn.

models, respectively. As shown in Figure 6, the characters synthesized by WL-SV [16] contain broken or incorrect shapes. However, the story images synthesized by our CoIn have the correct shapes and clear backgrounds. Compared with StoryDALL-E [22] in Figure 7, our CoIn can synthesize a more consistent character appearance. Although our CoIn has achieved better image quality than StoryDALL-E, we have observed some limitations in terms of generalization compared to AR-LDM [26]. Due to the powerful Stable Diffusion with large-scale pretraining, AR-LDM can generate uncommon content not frequently seen in the training set. However, it is worth noting that our CoIn still achieves comparable results to AR-LDM in certain scenarios with $\sim 360\times$ faster synthesis speed and 4.6% generator parameters.

4.5 Ablation Study

To verify the effectiveness of different components in the proposed CoIn, we conduct ablation studies on the story visualization and continuation tasks, respectively. The results of Pororo-SV and Flintstones-SV are shown in Table 3. The components being evaluated include DINO-based D, Context G, Contextual SVF, and

Table 3: The performance of different components of our model on the test set of Pororo-SV and Flintstones-SV.

Task	Method	Pororo		Flintstones	
		FID (\downarrow)	FSD (\downarrow)	FID (\downarrow)	FSD (\downarrow)
Visualization	Baseline	52.64	65.08	58.29	64.78
	+ CLIP-based D	31.46	52.19	38.71	55.46
	+ DINO-based D	28.46	49.34	33.71	52.46
	+ Context G	24.39	42.56	29.50	46.55
	+ Contextual SVF	19.38	38.19	27.96	39.32
	+ Intra-SI (CoIn)	16.93	33.41	24.52	36.15
Continuation	Baseline	57.15	66.12	62.11	67.18
	+ CLIP-based D	36.18	52.10	40.66	58.69
	+ DINO-based D	32.74	48.49	35.23	53.69
	+ Context G	26.83	43.41	28.97	48.40
	+ Contextual SVF	23.53	38.14	23.22	45.69
	+ Intra-SI (CoIn)	18.63	34.73	19.95	39.17

Intra-SI. We also compare the results of CLIP-based D and DINO-based D. Our baseline is a modified text-to-image DF-GAN [42] adapted for story visualization tasks. The baseline fuses the story information through FC layers.



Figure 7: Comparison of story continuation results between StoryDALL-E, AR-LDM, and our proposed CoIn.

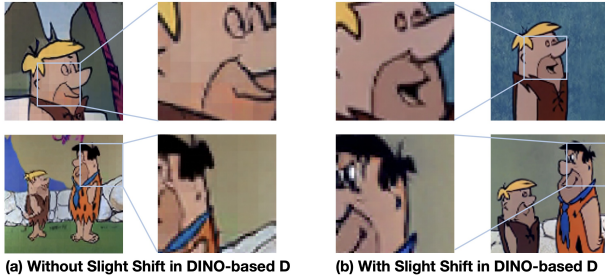


Figure 8: Comparison of the synthesized images between the DINO-based D with and without Slight Shift.

From Table 3, we can observe that our proposed DINO-based D achieves better performance than CLIP-based D on these two tasks. If we further introduce the Context G to decompose the generation task, we can observe a further improvement of FID and FSD. Armed with Contextual SVF, the model also decreases FID and FSD from 24.39 and 42.56 to 19.38 and 38.19 for the story visualization task on Pororo. The proposed Intra-SI further decreases FID and FSD

on two datasets. The ablation studies demonstrate the effectiveness of our proposed modules in both story visualization and continuation tasks. Furthermore, we compare the differences in generated images between the DINO-based D with and without Slight Shift in Figure 8. It is evident that Slight Shift effectively alleviates the issue of generating patch-like patterns.

5 Conclusion

In this paper, we propose a lightweight and effective CoIn (Contextualize and Interchange) framework for story visualization and continuation. Moreover, we propose a Context-aware Story Generator that explicitly extracts the global context information and synthesizes context visual features for each image generator. Furthermore, we propose an Intra-story Interchange module to enable information exchange between intra-story image generators. Lastly, a pair of DINO-based discriminators is introduced to assess the story quality more accurately. Our CoIn achieves competitive synthesis quality with $\sim 360\times$ faster synthesis speed and 4.6% generator parameters. Our CoIn also significantly reduces the required resources and training time. This is particularly crucial for tasks like story visualization that involve generating a large number of images.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants (No.62325206, 61936005), the Key Research and Development Program of Jiangsu Province under Grant BE2023016-4, and the Postgraduate Research & Practice Innovation Program of Jiangsu Province under Grant KYCX22_0947.

References

- [1] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. 2022. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324* (2022).
- [2] Hong Chen, Rujun Han, Te-Lin Wu, Hideki Nakayama, and Nanyun Peng. 2022. Character-Centric Story Visualization via Visual Planning and Token Alignment. *arXiv preprint arXiv:2210.08465* (2022).
- [3] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *NeurIPS* (2021).
- [4] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. 2021. Cogview: Mastering text-to-image generation via transformers. In *NeurIPS*.
- [5] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. 2022. CogView2: Faster and Better Text-to-Image Generation via Hierarchical Transformers. *arXiv preprint arXiv:2204.14217* (2022).
- [6] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. 2022. Make-a-scene: Scene-based text-to-image generation with human priors. In *ECCV*.
- [7] Yuan Gong, Youxin Pang, Xiaodong Cun, Menghan Xia, Haoxin Chen, Longyue Wang, Yong Zhang, Xintao Wang, Ying Shan, and Yujiu Yang. 2023. Tale-Crafter: Interactive Story Visualization with Multiple Characters. *arXiv preprint arXiv:2305.18247* (2023).
- [8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NeurIPS*.
- [9] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. 2022. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*.
- [10] Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. 2018. Imagine this! scripts to compositions to videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 598–613.
- [11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*.
- [12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *NeurIPS* (2020).
- [13] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. 2022. Cascaded Diffusion Models for High Fidelity Image Generation. *JMLR* 23 (2022), 47–1.
- [14] Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. 2017. Deepstory: Video story qa by deep embedded memory networks. *arXiv preprint arXiv:1707.00836* (2017).
- [15] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.
- [16] Bowen Li. 2022. Word-Level Fine-Grained Story Visualization. In *European Conference on Computer Vision*. Springer, 347–362.
- [17] Bowen Li, Philip HS Torr, and Thomas Lukasiewicz. 2022. Clustering generative adversarial networks for story visualization. In *Proceedings of the 30th ACM International Conference on Multimedia*. 769–778.
- [18] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. 2019. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6329–6338.
- [19] Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, and Weidi Xie. 2023. Intelligent Grimm—Open-ended Visual Storytelling via Latent Diffusion Models. *arXiv preprint arXiv:2306.00973* (2023).
- [20] Adyasha Maharana and Mohit Bansal. 2021. Integrating Visuospatial, Linguistic and Commonsense Structure into Story Visualization. *arXiv preprint arXiv:2110.10834* (2021).
- [21] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. 2021. Improving generation and evaluation of visual stories via semantic consistency. *arXiv preprint arXiv:2105.10026* (2021).
- [22] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. 2022. StoryDALL-E: Adapting Pretrained Text-to-Image Transformers for Story Continuation. *arXiv preprint arXiv:2209.06192* (2022).
- [23] Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *ICML*.
- [24] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *ICML*.
- [25] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).
- [26] Xichen Pan, Pengda Qin, Yuhong Li, Hui Xue, and Wenhui Chen. 2022. Synthesizing Coherent Story with Auto-Regressive Latent Diffusion Models. *arXiv preprint arXiv:2211.10950* (2022).
- [27] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023).
- [28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.
- [29] Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. 2023. Make-a-story: Visual memory conditioned consistent story generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2493–2502.
- [30] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).
- [31] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *ICML*.
- [32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.
- [33] Robin Rombach and Patrick Esser. [n.d.]. Stable diffusion v1-4. <https://huggingface.co/CompVis/stable-diffusion-v1-4>.
- [34] Chihyeon Kim Doyup Lee Saehoon Kim, Sanghun Cho and Woonhyuk Baek. 2021. minDALL-E on Conceptual Captions. <https://github.com/kakaobrain/minDALL-E>.
- [35] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *arXiv preprint arXiv:2205.11487* (2022).
- [36] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. 2023. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. In *International conference on machine learning*. PMLR, 30105–30118.
- [37] Yefei Sheng, Ming Tao, Jie Wang, and Bing-Kun Bao. 2024. ISF-GAN: Imagine, Select, and Fuse with GPT-Based Text Enrichment for Text-to-Image Synthesis. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 7 (2024), 1–17.
- [38] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*.
- [39] Yun-Zhu Song, Zhi Rui Tam, Hung-Jen Chen, Huiao-Han Lu, and Hong-Han Shuai. 2020. Character-preserving coherent story visualization. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*. Springer, 18–33.
- [40] Ming Tao, Bing-Kun Bao, Hao Tang, Yaowei Wang, and Changsheng Xu. 2024. StoryImager: A Unified and Efficient Framework for Coherent Story Visualization and Completion. *arXiv preprint arXiv:2404.05979* (2024).
- [41] Ming Tao, Bing-Kun Bao, Hao Tang, and Changsheng Xu. 2023. GALIP: Generative Adversarial CLIPs for Text-to-Image Synthesis. *arXiv preprint arXiv:2301.12959* (2023).
- [42] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Xu Changsheng. 2022. Df-gan: A Simple and Effective Baseline for Text-to-Image Synthesis. In *CVPR*.
- [43] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. 2024. Visual Autoregressive Modeling: Scalable Image Generation via Next-Scale Prediction. *arXiv preprint arXiv:2404.02905* (2024).
- [44] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiao-Lei Huang, and Xiaodong He. 2018. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*.
- [45] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789* (2022).
- [46] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. 2019. Self-attention generative adversarial networks. In *ICML*.

- [47] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiao lei Huang, and Dimitris N Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*.
- [48] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiao lei Huang, and Dimitris N Metaxas. 2018. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE TPAMI* 41, 8 (2018), 1947–1962.
- [49] Wendi Zheng, Jiayan Teng, Zhuoyi Yang, Weihang Wang, Jidong Chen, Xiaotao Gu, Yuxiao Dong, Ming Ding, and Jie Tang. 2024. CogView3: Finer and Faster Text-to-Image Generation via Relay Diffusion. *arXiv preprint arXiv:2403.05121* (2024).
- [50] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. 2019. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *CVPR*.
- [51] Zhongyang Zhu and Jie Tang. 2023. CogCartoon: Towards Practical Story Visualization. *arXiv preprint arXiv:2312.10718* (2023).